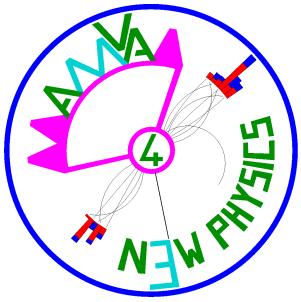
This Report is part of a project that has received funding from the **European Union's Horizon 2020 research and innovation programme under grant agreement Nº675440**



AMVA4NewPhysics ITN

Work Package 4 - Deliverable 4.2

Report on a Statistical Learning method for Model-Independent searches for New Physics

December 28, 2017

Abstract

Model-Independent searches for New Physics at the Large Hadron Collider can be stated as a collective anomaly detection problem. We propose the parametric approach that uses a semi-supervised learning algorithm. It is based on a penalized likelihood and is able to jointly perform appropriate variable selection and detect possible collective anomalous behavior in data with respect to a given background sample. The proposed algorithm is tested both on synthetic and simulated data from the Large Hadron Collider.

1 Introduction

One of the main goals of the Large Hadron Collider (LHC) physics program is to probe the current theory of elementary particle physics, known as the Standard Model, and to further discover phenomena beyond that theory if they occur. The LHC accelerates protons in a 27 km ring that lead to collisions with a center of mass energy of 13 TeV each 25 ns, making it the largest and most powerful collider to date. General-purpose detectors at the LHC, namely ATLAS and CMS, have been primarily designed to measure the properties of the phenomena emerging from proton-proton collisions.

Since its conception over four decades ago the Standard Model has proven to be a successful theory, its predictions go from precise values of parameters to the existence of particles. Nonetheless, many theoretical extensions of the Standard Model aim to solve its well-known problems (e.g. the absence of a quantum description of gravity), establishing guidelines for experimental searches for phenomena beyond the Standard Model, referred to as New Physics (NP). Experimental searches for NP usually assume hypotheses associated to the existence of some theorized NP phenomena, which are to be confirmed or ruled out to a certain extent, for example, by comparing the data with the predictions that include such phenomena in a specific region of a parameter space. However, this model-dependent approach is limited in two senses: firstly, being guided by a NP model, it fails to cover a wider space of parameters where untheorized NP could appear; and secondly, it focuses on an often small set of experimental signatures which, even after combining all existing model-dependent NP searches at the LHC, leaves a broad number of signatures unexplored [1, 2]. In order to address these issues, Model-Independent, multi-signature searches have to be performed.

This work presents a novel anomaly detection method and an application to a model independent search at the LHC. The method is oriented on collective anomaly searches and employs a semi-supervised learning approach to cope with the scenario in which no physics knowledge for an anomaly is provided a priori. Special care is taken to perform proper variable selection and in the use of the two related input datasets: one that contains only Standard Model physics events and another that could additionally contain a set of anomalous events. Results for simulated synthetic data and for a dijet search at the LHC are presented as a proof of concept.

This document is organized as follows. In Section 2 we describe the physics problem and pose it in statistical terms. We then describe the method introduced by this work on section 3, followed by results of the application of the method in artificial data and in a search at the LHC using a simulated datasets that contain a NP signal, in Sections 4 and 5, respectively. Concluding remarks are presented in Section 6.

2 Description of the problem

Detectors such as ATLAS and CMS can identify and measure with precision most final-state particles produced in proton-proton collisions: photons (γ), leptons (e, μ, τ) and jets (j), including those that originate from a b quark decay, known as b-jets. One can classify each collision event according to its particle content in the final state, a so-called experimental signature, thereby allowing to separate the events in different case studies.

The goal of NP search analyses is to find evidence of the existence of new possible particles not predicted by the Standard Model. The main underlying assumption is that such new particles, if they do exist, show an anomalous behavior with respect to the known Standard Model physics. Within the Model Independent approach one relies on the prediction of the Standard Model processes, the so-called background, in general obtained by Monte Carlo (MC) simulations, to be compared with the actual experimental data that could additionally contain events coming from a NP process. For the purpose of this study, both background and experimental data were produced using the MC simulation packages MadGraph5 [3], Pythia8 [4, 5], and Delphes [6], as described in section 5, where NP signal events are injected in the experimental data with different proportions on top of the Standard Model processes in order to assess the power and limitations of the method presented.

There have been several efforts in the CMS and ATLAS collaborations to perform Model Independent searches. In particular, there are signature-specific searches as, e.g., the ones in [7, 8]. More generally, both experimental collaborations have performed Model Independent searches for NP in a systematic way in many signatures [9, 10, 11, 12, 13], an approach denominated General Search. Automatically exploring multiple signatures entails analyses of high volumes of data and in general LHC searches have been based on a simple uni-variate analysis. This approach consists in separating data and simulated events into hundreds of independent experimental signatures; then for corresponding class of events, one-dimensional histograms for data and simulation are produced for a few kinematic variables. A simple algorithm can then be used to identify the region on each histogram where the disagreement between data and simulation is the largest. By construction, this approach is not able to capture potential anomalies that become manifest when probing the datasets in multiple variables at once, which requires a multi-variate approach, such as the one presented in this work.

From a statistical point of view, the aforementioned problem of NP search can be framed in the context of anomaly detection [14] when one searches for events, also referred to as observations, that are not consistent with an assumed model. One of the most common ways to define anomalies is to refer them as unusual events, unlikely to be observed under the assumed model. If such definition were applied to the considered problem, any signal event that has a substantial probability to be also generated by the background process would be falsely classified as non-anomalous. Consequently, the interest is put on finding groups of observations that separately

do not need to have anomalous properties (according to the former definition) but their common occurrence in a particular part of the data domain is unexpected. This is known as "*collective anomaly*" detection.

Two data samples are at hand: the background - MC data, simulated according to the Standard Model, and the experimental sample, possibly also including observations from the unknown signal. Since there is only a partial knowledge of processes generating the datasets, the anomaly detection problem emerges as semi-supervised learning approach. Simulated data are labelled as drawn from the background, while no direct information about the generating process of the experimental data, which we then refer to as unlabelled.

Hence, we shall operate in an unsupervised manner and later strengthen the classification power based on the accessible labels of the background observations. The starting point for our approach is found in Vatanen et al. [15] and Kuusela et al. [16] and is described in the following section.

3 Methods

3.1 Modeling data by mixture models

Let $X = (\boldsymbol{x}_1,...,\boldsymbol{x}_n)'$ be the sample of the MC background data, where for $i = 1,...,n$, $\boldsymbol{x}_i = (x_{i1},...,x_{ip},...,x_{iP})'$ are i.i.d. P-dimensional observations with unknown probability density function f_B. Also, denote by $Y = (\boldsymbol{y}_1,...,\boldsymbol{y}_m)'$ the experimental data, where for $i = 1,...,m$, $\boldsymbol{y}_i = (y_{i1},...,y_{ip},...,y_{iP})'$ are i.i.d. observations from the unknown probability density function f_{SB}. It is natural to assume for f_{SB} a mixture of the two components:

$$f_{SB}(\boldsymbol{y}) = (1-\lambda) f_B(\boldsymbol{y}) + \lambda f_S(\boldsymbol{y}) \tag{1}$$

where f_B and f_S are the background and signal distributions respectively. Parameter $\lambda \in [0,1]$ represents the proportion of the signal events in the whole data Y.

The models for the background and signal should be arbitrarily flexible so that even complex processes could be well represented. For this reason we assume that both f_S and f_B are mixtures. In particular we consider finite mixtures of Gaussian distributions as they are capable of approximating well every distribution. The Gaussian mixtures suit particularly well as they have been proven to serve well for density estimation and for classification purposes [17]. Hence, let us specify the two models for the background and signal respectively as

$$f_B(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{y}|\boldsymbol{\mu}_k, \Sigma_k) \quad \text{and} \quad f_S(\boldsymbol{y}) = \sum_{q=1}^{Q} \pi_q \phi(\boldsymbol{y}|\boldsymbol{\mu}_q, \Sigma_q) \tag{2}$$

where K and Q are the numbers of Gaussian components in the mixtures, π_k and π_q for $k = 1,...,K$ and $q = 1,...,Q$ are the mixing proportions (with the constraints

$\sum_{k=1}^{K} \pi_k = 1$ and $\sum_{q=1}^{Q} \pi_q = 1$) and $\phi(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{P}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$ denotes the P-variate Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Consequently, the model for the experimental data Y is also the Gaussian mixture

$$f_{SB}(\boldsymbol{y}) = (1-\lambda)\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{y}|\boldsymbol{\mu}_k, \Sigma_k) + \lambda \sum_{q=1}^{Q} \pi_q \phi(\boldsymbol{y}|\boldsymbol{\mu}_q, \Sigma_q)$$

$$= \sum_{l=1}^{K+Q} \tilde{\pi}_l \phi(\boldsymbol{y}|\boldsymbol{\mu}_l, \Sigma_l),$$

where $\tilde{\pi}_k$ are mixture proportions accounting for the λ and naturally $\sum_{l=1}^{K+Q} \tilde{\pi}_l = 1$ holds.

Parameters of a mixture model are usually estimated via maximum likelihood. Consider, for the sake of simplicity, the estimation of the parameters involved in f_B (equation 2). Conditionally to the observation of the X data, the log-likelihood is defined as:

$$\log L_p(\theta) = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k)\right],$$

where $\theta = (\pi_1, ..., \pi_K, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \Sigma_1, ..., \Sigma_K)$ are model parameters. In literature the mixture components are often related to data clusters [17] and the background estimation could be referred as model-based clustering. Since the optimization of the likelihood has no explicit solution, numeric methods have to be used. Here we employ the Expectation-Maximization (EM) algorithm [18]. The method relies on an iterative two-steps loop. The first step is to compute the expectation for each observation to belong to a given component and it is computed based on the current parameter values. Subsequently, conditioning on the expectations (posterior probabilities of cluster membership) the parameter estimates are updated as the result of maximum likelihood optimization.

The algorithm is described by the following formulas:

- Expectation step - given current values of the estimates the probability of the observation $\boldsymbol{x}_j$ belonging to the l^{th} cluster (component) is:

$$\tau_{il}^{(r)} = \frac{\pi_l^{(r)} \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_l^{(r)}, \Sigma_l^{(r)})}{\sum_{k=1}^{K} \pi_k^{(r)} \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k^{(r)}, \Sigma_k^{(r)})}. \tag{3}$$

- Maximization step - given posterior probabilities τ_{il} update the current parameter estimates

– Component proportions for $k = 1, ..., K$

$$\hat{\pi}_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(r)}. \tag{4}$$

– Component mean parameters for $k = 1, ..., K$ and $p = 1, ..., P$

$$\hat{\mu}_{kp}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} x_{ip}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{5}$$

– Component covariance matrices

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} * (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})'}{\sum_{i=1}^{n} \tau_{ik}^{(r)}}. \tag{6}$$

The upper index (r) indicates the subsequent iterations of the algorithm. Each iteration of the EM algorithm has been proven to increase the model likelihood. The loop iterations are performed until the estimates converge, i.e. changes of the likelihood between the subsequent iteration are arbitrarily small.

In order to obtain a solution for the semi-supervised problem of anomaly detection we stem from the *fixed-background* model introduced in Vatanen et al. [15] and Kuusela et al. [16]. The main idea underlying those works is to estimate the model parameters in a 2-step procedure: first based on the background data X the model $f_B(\boldsymbol{x})$ is fitted. In the second step for the experimental data Y, the model parameters are estimated by keeping the background parameters fixed (component proportions are adjusted according to λ value).

3.2 Dimensionality reduction methods in mixture models

The fixed-background model is built upon a natural idea and makes a logical use of the two datasets. However, as pointed out by the authors, for high dimensional spaces the model suffers from the curse of dimensionality. This affects both the ability to find possible signal components and, possibly, the algorithm convergence. For data of dimension P, the number of free parameters to be estimated is large - for each covariance matrix $\frac{P(P+1)}{2}$ parameters, KP for mean vectors and $K-1$ for mixture proportions, giving a total amount of $\frac{K(P+1)(P+2)}{2} - 1$ free parameters.

It follows the need of reducing the dimensionality of the parameter space. Vatanen et al. [15] perform principal component analysis of the original space and run the anomaly detection algorithm in the reduced space spread on the first few components. However, there is absolutely no guarantee that the first few selected principal components exhibit any deviation of the signal distribution versus the background, in presence of an anomalous component. Alternatively, in the context of

anomaly detection,[19] perform variable selection based on some heuristics, i.e. criteria related to the divergence between the marginal distributions of the background and experimental data. The proceeding estimation of parameters is performed subsequently as if the two tasks were independent. In the unsupervised context of model-based clustering, several approaches have been introduced to reduce the number of parameters. In Banfield and Raftery [20] parsimonious mixtures of Gaussian distributions are proposed. The specific constraints on the component covariance matrices are set so that the number of free parameters is greatly reduced (for example by using component-specific diagonal covariance matrices). Based on that idea the model-based clustering framework was proposed by Celeux and Govaert [21] . In a similar manner Bouveyron et al. [22] proposed another family of Gaussian mixture models with other type of constraints on component covariance matrices.

An alternative approach to reduce the number of parameters to be estimated is to jointly perform parameter estimation and variable selection. In the field of statistical regression such approach is common [23, 24] and it makes use of an estimates-based penalty function. In this setting, the most relevant example is the LASSO [25], where the number of model free parameters is decreased by introducing a penalization of the optimization function, causing parameters shrinkage to a fixed value (possibly to 0). In general, instead of the log-likelihood optimization, the aim is to maximize the penalized log-likelihood expression:

$$\log L_p(\theta) = \log L(\theta) - \gamma p(\theta).$$

were γ is a regularization parameter and $p(\theta)$ is a penalty function of the model parameters θ. The regularization parameter γ serves as a balance between bias of the estimates and small penalty value. Penalty introduces bias with respect to the standard maximum likelihood solution but on the other hand, it decreases the variance of the estimates that results in better model performance in terms of mean square error.

The origin of the penalized approach for mixture models should be found in Pan and Shen [26]. The authors proposed a penalized log-likelihood that automatically performs variable selection. For the standardized data, it is proposed to shrink component mean parameters to 0 under assumption that the component covariance matrices are equal to identity. The authors considered the following form of the penalized log-likelihood

$$\log L_p(\theta) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right] - \gamma \sum_{k=1}^{K} \sum_{p=1}^{P} |\mu_{kp}|. \tag{7}$$

The Maximum Penalized Log-likelihood Estimates (MPLE) are found using a modified EM algorithm (for details we refer to the source article). Shrinkage of the mean parameters to 0 is not only causing the regularization but serves for feature selection, as described in subsection 3.3.3.

Xie [27] relaxes the assumption of the identity covariance matrices to the component-specific diagonal covariance matrices. A second penalty term is introduced, which encourages diagonal elements of the covariance matrices to be equal to 1 while it also leads to variable selection in similar manner. In parallel Zhou et al. [28] consider another penalty as a function of the covariance matrices. Specifically, it is a function of the elements of the component-specific covariance matrix inverses, that is applied to enable the needed regularization.

In contrast to Pan and Shen [26] where a l_1 penalty of the mean parameters is considered (equation 7) Xie et al. [29] introduces a l_2 penalty on the means that simultaneously shrinks a whole vector of parameters for particular variables. They consider the following penalized log-likelihood formula:

$$\log L_p(\theta) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right] - \gamma \sum_{p=1}^{P} \|\mu_{\cdot p}\|,$$

where $\|\mu_{\cdot p}\| = \sqrt{\sum_{k=1}^{K} \mu_{kp}}$ for $p = 1, ..., P$. The l_1 penalty causes shrinkage of the mean parameters individually and variable selection is performed if and only if for some $p = 1, ..., P$ and all $k = 1, ..., K$ $\mu_{kp} = 0$. In order to obtain the effective shrinkage to 0 for a given variable of all the component-specific mean parameters, the grouped shrinkage is more suitable, i.e. it is the case for the grouped LASSO [30]. For the grouped penalty it is more likely that the true uninformative variables are removed and less bias is introduced to the estimates of the informative variable-dependent parameters.

In the semi-supervised context, the idea for a penalized algorithm for dimension reduction was introduced in Pan et al. [31]. The authors consider the mixture of Gaussian models and propose a shrinkage method for parameters to obtain sparse solutions in a related manner to the LASSO. However in Pan et al. [31], a slightly different anomaly detection problem is considered. In their case a single partially labeled dataset is considered and it is assumed that each Gaussian component of the model corresponds to a single label. However, for the considered problem of this paper, an interplay of the two datasets at hand is studied. The datasets have different origin and in general could vary in size. First, if the two sets were merged, the possible signal fraction in the produced data is decreased resulting in poorer anomaly detection power. The second discrepancy with the considered problem in this paper is that the background observations should not be modeled by a single Gaussian component as the true distribution of the physical processes is much more complex.

3.3 A penalized approach in mixture models

3.3.1 Penalization of the background

In the following we consider a comprehensive approach to jointly estimate the parameters and reduce the data dimensionality P by borrowing and extending the fixed-background model idea [15, 16] and the parameter penalization approach described in [31].

Since the fixed-background model is a two-step procedure, firstly requiring estimation of the background distribution in an unsupervised manner, variable selection based on penalization has to be primarily defined in the model-based clustering setting. In the following, we extend the approach of Xie [27] by relaxing constraints on the component-specific covariance matrices to be arbitrary. The proposed procedure makes use of two penalties as the mean for variable selection: the first is a function of the component mean parameters and the second a function of the component-specific covariance matrix eigenvalues. In the following the method is referred as Mean And Eigenvalue Shrinkage algorithm (MAES).

The first penalty of the MAES algorithm (a function of components mean vectors) borrows the idea of the grouped shrinkage presented in Xie et al. [29]. However, for the problem at hand the algorithm should be sensitive to precise estimation of infrequent component parameters. According to Bühlmann and Van De Geer [32] if the true proportions of the components differ substantially, penalty should be a function of appropriately weighted parameters. It aims to balance an influence of unequal components proportions, so that all the mean parameters are encouraged to be shrunk uniformly, not mostly the one corresponding to the components with the smallest proportions. The penalty is then formulated as follows:

$$p_1(\theta) = \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \pi_k \mu_{kp}^2}$$

were proportions π_k serve as the weights.

The second penalty of the MAES algorithm depends on the covariance matrices of the components. It relies on eigenvalue decomposition of each component covariance matrix and the idea is to perform a shrinkage of the smallest eigenvalues to 0. The aim is to obtain a low rank approximation of the clusters covariance matrices, that should be especially effective if clusters lie in the lower dimensional sub-spaces. In fact, shrinkage of parameters defining the component covariance matrices, may result in non-positive-definite matrices which entails impossibility of the likelihood computation. Alternatively, it is possible to formulate a generalized Gaussian distribution by using a pseudo-determinant and a generalized inverse matrix. We considered this approach but we found that the optimization of such a generalized penalized log-likelihood using the EM algorithm was unstable as it often converged to the meaningless local maxima, the likelihood steps were not

monotone and the algorithm often locked itself in loops (back and forth changes of the parameter values between the subsequent iterations). For this reason no further results of this approach are shown.

In order to circumvent these problems we propose to shrink the eigenvalues to a component-specific small positive value ϵ_k. In this way, the expected regularization is performed and the likelihood can be written explicitly. Additionally, such a shrinkage prevents the EM algorithm from running into singularities of the likelihood if the smallest eigenvalue for some component covariance matrix tends to 0. As pointed out in [15, 16], this is a frequent issue of the EM algorithm employment.

If the L_k smallest eigenvalues for the k^{th} component are shrunk to ϵ_k then the number of free parameters is decreased by $\sum_{k=1}^{K}(L_k-1)+\frac{(L_k-1)*L_k}{2} = \sum_{k=1}^{K}\frac{(L_k-1)*(L_k+2)}{2}$. The first term results from equality of L_k eigenvalues, the second from the arbitrariness of choice of the eigenvectors associated to the smallest and equal eigenvalues (the last eigenvectors are arbitrarily chosen to fill up the orthogonal bases of the other eigenvectors associated to the larger eigenvalues).

In order to obtain a component-specific eigenvalue shrinkage, the eigenvalue decomposition for each component covariance matrix is performed:

$$\Sigma_k = Q_k D_k Q_k'$$

where D_k is a diagonal matrix of eigenvalues and Q_k matrix containing orthonormal eigenvectors q_{kp}. Let us represent by δ_{kp} the p^{th} largest eigenvalue of the matrix Σ_k. Consequently, the penalty is formulated as:

$$p_2(\theta) = \sum_{k=1}^{K}\sum_{p=1}^{P}\max(\delta_{kp}, \epsilon_k).$$

In contrast to the p_1 penalty, for p_2 penalty it is not necessary (although possible) to use weighs in the penalty function. In practice the components covariance matrices estimation is more robust to a presence of the unbalanced clusters than the one of the components mean.

Selection of ϵ_k is performed based on the asymptotic distribution of the eigenvalues [33]. Assuming that the L smallest eigenvalues of the population covariance matrix is equal to δ_{const}, the asymptotic distribution of the L smallest unsorted eigenvalues δ_l of the sample covariance matrix is normal with the mean δ_{const} and variance $\frac{2\delta_{const}^2}{nL}$. Hence the estimate $\hat{\epsilon}_k$ is an unbiased estimator of δ_{const} if $\hat{\epsilon}_k = \frac{1}{n}\sum_{p=P-L+1}^{P}\delta_{kp}$ that is the mean of L smallest eigenvalues.

The parameter L_k is selected based on a sequential test with an adjustment for the significance level α. The sequential test uses partially the same data and a Bonferroni-like correction [34] is applied so that the type 1 error does not increase for the multiple testing paradigm. First, it is tested if all the eigenvalues of the k^{th} component covariance matrix are equal ($L_k = P$) against the general alternative. Let

$\bar{\hat{\delta}}_k$ be the mean of all the eigenvalues of the sample covariance matrix for the k^{th} component. The rejection regions for the tested hypothesis are determined as

$$\frac{\hat{\delta}_{k1}}{\bar{\hat{\delta}}_k} > 1 - \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2}} \quad \vee \quad \frac{\hat{\delta}_{kP}}{\bar{\hat{\delta}}_k} < 1 + \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2}}$$

where $z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ quantile of the normal random variable. In summary, the null hypothesis is rejected if the ratio between the largest eigenvalue to the mean of eigenvalues is too large or the ratio between the smallest eigenvalue to the mean is to low. If there is no reason to reject the null then $L_k = P$, otherwise a sequential test is performed if the $P-1$ smallest eigenvalues are equal. For this aim the largest eigenvalue estimate is discarder from the considered set, mean value $\bar{\hat{\delta}}_k$ recomputed and the new rejection region determined

$$\frac{\hat{\delta}_{kh}}{\bar{\hat{\delta}}_k} > 1 - \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2h}} \quad \vee \quad \frac{\hat{\delta}_{kP}}{\bar{\hat{\delta}}_k} < 1 + \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2h}}$$

where h numerates upon the number of tests being performed. Note that quantiles are taken from different points as previously. If the correction is not applied, the first type error of the sequential test will grow uncontrolled. The step is repeated until there is no reasons to reject the null hypothesis and based on that stage parameter L_k is selected.

In summary, the parameters estimation for the MAES algorithm is performed via the optimization of the following penalized likelihood

$$\begin{aligned} \log L_p(\theta) &= \log L(\theta) - \gamma_1 p_1(\theta) - \gamma_2 p_2(\theta) && (8) \\ &= \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right] - \gamma_1 \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \pi_k \mu_{kp}^2} - \gamma_2 \sum_{k=1}^{K} \sum_{p=1}^{P} \max(\delta_{kp}, \epsilon_k) && (9) \end{aligned}$$

3.3.2 Parameters estimation of the background

Estimation of the parameters involved in 8 is obtained via a suitable modification of the EM algorithm.

The component proportions of the Gaussian mixture components are expressed equally to the unpenalized case (equation 4).

Due to the shrinkage, the MPLE are shifted with respect to the MLE due to the shrinkage. The maximum penalized likelihood estimates are shifted with respect to the standard non-penalized estimates, and expressed as:

$$\hat{\mu}_{kp}^{(r+1)} = \begin{cases} 0 & \text{if } \left(\sum_{k=1}^{K} \left(\sum_{i=1}^{n} \tau_{ik}^{(r)} x_{ip} \right)^2 \right)^{\frac{1}{2}} \leq \gamma_1 M_p^{(r)} \\ \tilde{\mu}_{kp}^{(r+1)} - \gamma_1 \frac{\hat{\mu}_{kp}^{(r)} \Sigma_{k,pp}^{(r)}}{\|\hat{\mu}_{\cdot p}^{(r)}\| \sum_{i=1}^{n} \tau_{ik}^{(r)}} & \text{otherwise} \end{cases} \qquad (10)$$

where $\tilde{\mu}_{kp}^{(r)}$ are the MLEs for the means given by equations 5 and $M_p^{(r)} = \max_{k=1,\ldots,K} \hat{\Sigma}_{k,pp}^{(r)}$. Obviously, based on equation 10, for a sufficiently large γ_1, all the mean components for the p^{th} variable are $\mathbf{0}$ when a given threshold is crossed.

Concerning the covariance matrices, denoted by $\tilde{\Sigma}_k^{(r)}$ the MLE estimates, eigenvalue decomposition of $\tilde{\Sigma}_k^{(r)}$ is performed for each iteration r

$$\tilde{\Sigma}_k^{(r)} = \hat{Q}_k^{(r)} \tilde{D}_k^{(r)} \left(\hat{Q}_k^{(r)}\right)'. \tag{11}$$

As no penalty is applied to component eigenvector matrix $\hat{Q}_k^{(r)}$, their penalized estimates are the same as the MLEs. If the eigenvalues were shrunk to 0 then their MPLE of $\hat{D}_{k,pp}^{(r+1)}$ would be equal to $\bar{D}_{k,pp}^{(r+1)}$ expressed as:

$$\bar{D}_{k,pp}^{(r+1)} = \frac{-n\hat{\pi}_k^{(r+1)} + \sqrt{\left(n\hat{\pi}_k^{(r+1)}\right)^2 + 8\gamma_2 n\hat{\pi}_k^{(r+1)} \tilde{D}_{k,pp}^{(r)}}}{4\gamma_2}. \tag{12}$$

However, as the eigenvalues are shrunk to $\epsilon_k > 0$ then the sequential tests are performed based on which particular eigenvalue estimates $\hat{D}_{k,pp}^{(r+1)}$ are assigned either to be equal to ϵ_k or $\bar{D}_{k,pp}^{(r+1)}$.

3.3.3 Variable selection for the background

Under assumption of component identity covariance matrices as considered by Pan and Shen [26], applied penalization leads to straightforward variable selection. If for all components the estimated mean value corresponding to the p-th variable is 0 then the p-th variable is uninformative for cluster classification, hence should be removed from the considered data. This follows from equation 3 of the posterior probability of observation membership

$$\tau_{il} = \frac{\pi_l \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_l, I_P)}{\sum_{k=1}^K \pi_k \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, I_P)} = \frac{\pi_l \phi(x_{ip}|0,1)\phi(\boldsymbol{x}_{i,-p}|\boldsymbol{\mu}_{l,-p}, I_{P-1})}{\sum_{k=1}^K \pi_k \phi(x_{ip}|0,1)\phi(\boldsymbol{x}_{i,-p}|\boldsymbol{\mu}_{k,-p}, I_{P-1})} \tag{13}$$

where I_P is a diagonal matrix of dimension P and index $i,-p$ denotes the i^{th} vector with the removed p^{th} variable. After simplification of the equation it is clear that the observation for the p-th variable does not contribute to the classification based on the posterior probability.

For a more general case of the component-specific diagonal covariance matrix as in Xie [27], in order to remove the p-th variable the two conditions have to be met. The first one corresponds to mean estimates equal 0 as described above, while the second relies on the marginal component variances for a considered p-th variable

that should all be equal to 1 (due to data standardization). In that case, the posterior probability that an observation $\boldsymbol{x}_i$ belongs to the l-th cluster is equal to

$$\tau_{il} = \frac{\pi_l \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \Sigma_l)}{\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k)} = \frac{\pi_l \phi(x_{ip}|0,1)\phi(\boldsymbol{x}_{i,-p}|\boldsymbol{\mu}_{l,-p}, \Sigma_{l,-p})}{\sum_{k=1}^{K} \pi_k \phi(x_{ip}|0,1)\phi(\boldsymbol{x}_{i,-p}|\boldsymbol{\mu}_{k,-p}, \Sigma_{k,-p})} \tag{14}$$

where notation $\boldsymbol{\mu}_{l,-p}$ denotes removal of p-th variable from the vector and $\Sigma_{l,-p}$ removal of p-th row and column from the matrix. In analogy to the previous case the p-th variable again does not contribute to the classification, hence it is uninformative.

In the general case where the covariance matrix is not diagonal (hence also the correlations between variables are modeled) such a simple factorization cannot be performed. Without loss of the generality let us index all the features and take the two subsets $A = (1, ..., L)$ and $B = (L+1, ..., P)$ for any $L \in [1, P-1]$. The data and the respective parameters could now be partitioned into two sets as follows: $X = (X_A, X_B)$, component mean vectors $\boldsymbol{\mu}_k = (\boldsymbol{\mu}_{k,A}, \boldsymbol{\mu}_{k,B})$ and component covariance matrices are represented by the block matrices $\Sigma_k = \begin{pmatrix} \Sigma_{k,AA} & \Sigma_{k,AB} \\ \Sigma_{k,BA} & \Sigma_{k,BB} \end{pmatrix}$ where $\Sigma_{k,AB}$ is a block matrix constructed from Σ_k by taking rows from the set A and columns from the set B. The goal is to split the joint distribution of $f(X_A, X_B)$ into a marginal probability of X_B and a conditional probability for X_A. For equations 13 and 14 it was done almost in an automatic way because uncorrelated Gaussian distributions are conditionally independent.

From the factorization $f(X_A, X_B) = f(X_A|X_B)f(X_B)$ we could generalize equation 14 and obtain the following formula:

$$\tau_{il} = \frac{\pi_l \phi(x_{iB}|\mu_{l,B}, \Sigma_{l,BB})\phi(x_{i,A}|\mu_{l,A} + \Sigma_{l,AB}\Sigma_{l,BB}^{-1}(x_{i,B} - \mu_{l,B}), \Sigma_{l,AA} - \Sigma_{l,AB}\Sigma_{l,BB}^{-1}\Sigma_{l,BA})}{\sum_{k=1}^{K} \pi_k \phi(x_{iB}|\mu_{k,B}, \Sigma_{k,BB})\phi(x_{i,A}|\mu_{k,A} + \Sigma_{k,AB}\Sigma_{k,BB}^{-1}(x_{i,B} - \mu_{k,B}), \Sigma_{k,AA} - \Sigma_{k,AB}\Sigma_{k,BB}^{-1}\Sigma_{k,BA})}. \tag{15}$$

The first necessary condition for removing set B of variables as uninformative for classification purpose is (as previously) to have 0 mean estimates for all the component means corresponding to the set B (for all $k = 1, \ldots, K$ and $p \in B$ $\mu_{kp} = 0$). However, in contrast to the suggestion in Zhou et al. [28], the condition is not sufficient for considering the variables from the set B as uninformative. The posterior probability of observation membership if the first condition is met is

$$\tau_{il} = \frac{\pi_l \phi(x_{iB}|\mathbf{0}, \Sigma_{l,BB})\phi(x_{i,A}|\mu_{l,A} + \Sigma_{l,AB}\Sigma_{l,BB}^{-1}x_{i,B}, \Sigma_{l,AA} - \Sigma_{l,AB}\Sigma_{l,BB}^{-1}\Sigma_{l,BA})}{\sum_{k=1}^{K} \pi_k \phi(x_{iB}|\mathbf{0}, \Sigma_{k,BB})\phi(x_{i,A}|\mu_{k,A} + \Sigma_{k,AB}\Sigma_{k,BB}^{-1}x_{i,B}, \Sigma_{k,AA} - \Sigma_{k,AB}\Sigma_{k,BB}^{-1}\Sigma_{k,BA})}. \tag{16}$$

which implicitly is a function of the parameters modeling the distribution of the presumably uninformative variables, i.e. $\Sigma_{l,BB}$ and $\Sigma_{l,AB}$. Hence, in order to obtain a proper variable selection, the second condition relying on the covariance matrices has to be met (as shown later by the simulations).

The second necessary condition is to have component-wise equal correlations between variables from the set B, that is for all $k = 1, \ldots, K$ $\Sigma_{k,BB} = \Sigma_{BB}$ and $\Sigma_{k,AB} = \Sigma_{AB}$. If both conditions are met for the set B then the posterior probability used for classification expressed formerly in 15 should be rewritten as

$$\tau_{il} = \frac{\pi_l \phi(x_{iB}|0_B, \Sigma_{BB}) \phi(x_{i,A}|\mu_{l,A} + \Sigma_{AB}\Sigma_{BB}^{-1} x_{i,B}, \Sigma_{l,AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})}{\sum_{k=1}^{K} \pi_k \phi(x_{iB}|0_B, \Sigma_{BB}) \phi(x_{i,A}|\mu_{k,A} + \Sigma_{AB}\Sigma_{BB}^{-1} x_{i,B}, \Sigma_{k,AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})}$$

$$= \frac{\pi_l \phi(x_{i,A}|\mu_{l,A} + \Sigma_{AB}\Sigma_{BB}^{-1} x_{i,B}, \Sigma_{l,AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})}{\sum_{k=1}^{K} \pi_k \phi(x_{i,A}|\mu_{k,A} + \Sigma_{AB}\Sigma_{BB}^{-1} x_{i,B}, \Sigma_{k,AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})}. \quad (17)$$

As a result the variables from set B influence the posterior distribution used for classification uniformly and independently on the true cluster membership. Hence if the two listed conditions are met, the variables from set B should be removed as uninformative for the dimensionality reduction purpose.

While the first condition is obtained automatically by shrinking the components mean vectors to 0, for the second condition a model selection has to be performed. Let the set A consist of all the features that do not meet the first condition and then the set B consists of the all potentially uninformative variables. Subsequently, let C be a set of all the possible subsets of set B ($C = C_1, ..., C_{N_B}$ for an appropriate N_B index) and $D_i = C_i^C$ be the complements of sets C_i for $i = 1, ..., N_B$.. Iteratively, the N_B new models are created in a following way. For each i, the common blocks of component covariance matrices are computed as a weighted average of components blocks

$$\Sigma_{C_iC_i} = \sum_{k=1}^{K} \pi_k \Sigma_{k,C_iC_i},$$

$$\Sigma_{D_iC_i} = \sum_{k=1}^{K} \pi_k \Sigma_{k,D_iC_i}$$

as it would be a sum of independent variables covariances. The idea is to substitute the component-specific blocks by the common one. The new model likelihood should be computed for each model. An information criterion is used to select the optimal model with features partitioned into sets of the uninformative variables C_i and the informative one D_i (i.e. the BIC criterion [35]).

The technique might seem computationally expensive, however, there is no need to fit $P!$ models. The first necessary criterion already filters out most of the true informative variables. There is also no need to check all the possible subsets of B as the minimal BIC values are likely obtained for models that are partitioned with subsets C_i with the highest cardinality.

3.3.4 Penalization of the background + signal model

The MAES algorithm and the associated data dimensionality reduction approach introduced above are the fundamental steps to obtain the background model and further to detect an unknown signal process. The developed anomaly detection algorithm (Penalized Anomaly Detection - PAD) integrates the penalized approach for the background density estimation and the fixed-background algorithm. The PAD algorithm introduces a penalty to the log-likelihood of the fixed-background model.

A form of the penalty should be chosen so that the proper variables are selected. A special care should be taken because the uninformative variables for the background model could be possibly powerful for the signal/background discrimination. Hence, the penalty should be a function of both the background and signal estimates. This in turn results in the dependence in estimation of the background and signal dependent parameters. The external loop iterates over the two-step procedure of the fixed-background model to adopt the fixed background model idea to the penalized approach and the following variable selection.

The experimental data model is a mixture of $K+Q$ Gaussian components, from which the first K are estimated based on the background data X and the rest on the mixed data Y. The penalized log-likelihood to be maximized has the following form:

$$\begin{aligned}\log L_p(\theta|\theta(X)) = \sum_{i=1}^{m} \log \Bigg[(1-\lambda) \sum_{k=1}^{K} \pi_k(X) \phi_k(\boldsymbol{y}_i|\boldsymbol{\mu}_k(X), \Sigma_k(X)) + \lambda \sum_{k=K+1}^{K+Q} \pi_k \phi_k(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \Sigma_k) \Bigg] \\ -\gamma_1(Y) \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \pi_k(X) \mu_{kp}(X)^2 + \sum_{k=K+1}^{K+Q} \pi_k \mu_{kp}^2} \\ -\gamma_2(Y) \sum_{p=1}^{P} \left(\sum_{k=1}^{K} \max(\delta_{kp}(X), \epsilon_k(X)) + \sum_{k=K+1}^{K+Q} \max(\delta_{kp}, \epsilon_k) \right),\end{aligned} \tag{18}$$

where $\theta(X)$ are the background model parameters, which are obtained by the maximization of the penalized log-likelihood. When necessary in brackets it is denoted on which dataset (X or Y) the parameters depend. For the background estimation we consider slightly different penalized log-likelihood than in case of MAES algo-

rithm, i.e. the first penalty is also the function of the signal parameters:

$$\log L_p(\theta|\theta(Y)) = \sum_{i=1}^{m} \log\left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k)\right] - \gamma_1(X) \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \pi_k \mu_{kp}^2 + \sum_{k=K+1}^{K+Q} \pi_k(Y)\mu_{kp}(Y)^2} - \gamma_2(X) \sum_{p=1}^{P} \left(\sum_{k=1}^{K} \max(\delta_{kp}, \epsilon_k)\right), \tag{19}$$

where $\theta(Y)$ are the signal model parameters

Equation 18 differs from the penalized log-likelihood of the MAES algorithm (equation 8) not only by the number of components that take part in the likelihood formulation and penalty formulation. The optimization is based on the two separate datasets and the penalty on the mean parameters is a joint function of all the background and signal components, hence it is not trivial to find the MPLE. Changes in the estimates of the signal components influence the penalty value that in turn impacts the estimates of the background density.

It is proposed to loop over the two-step procedure of the fixed background model, i.e. optimize subsequently formulas 18 and 19 until the changes of parameters are negligible.

3.3.5 Parameters estimation of the background + signal model

The parameters estimation for the PAD algorithm is processed by a properly modified MAES algorithm. For simplicity let us first describe the two-step procedure inside the outer loop. For the initialization, the simple background model should be found and for this aim, the MAES algorithm described in the previous subsection is applied. Subsequently, we need to find an estimate for the signal proportion λ and the estimates for the component-specific parameters for the rest of the Q signal components. In analogy to the MAES algorithm, the EM framework is employed.

First, the expectation step is computed for which posterior probabilities of cluster membership is computed based on the already known and fixed background components as:

$$\tau_{il}^{(r)} = \frac{\pi_l^{(r)} \phi\left(\boldsymbol{y}_i|\hat{\boldsymbol{\mu}}_l^{(r)}, \hat{\Sigma}_l^{(r)}\right)}{\sum_{k=1}^{K+Q} \pi_k^{(r)} \phi\left(\boldsymbol{y}_i|\hat{\boldsymbol{\mu}}_k^{(r)}, \hat{\Sigma}_k^{(r)}\right)}. \tag{20}$$

for l in $K+1, ..., K+Q$ (the signal components). The formulas for signal parameters 4, 5, 6, 11 and 12 remain unchanged but the iterations of the estimates are computed based on the Y dataset. The proportion of signal events λ is equal to $\sum_{k=K+1}^{K+Q} \pi_k$.

Given the proportion estimate $\hat{\lambda}$ the known background components proportions should be rescaled so that $\sum_{k=1}^{K+Q} \pi_k = 1$, that is to apply a transformation:

$$\pi_k^{(r+1)} = \frac{\pi_k^{(r+1)}}{\sum_{k=1}^{K} \pi_k^{(}r)} * (1 - \hat{\lambda})$$

for $k = 1, ..., K$.

The MPLE for the signal means are expressed as

$$\hat{\mu}_{kp}^{(r+1)} = \begin{cases} 0 & \text{if } \left(\sum_{k=1}^{K} \left(\sum_{i=1}^{n} \tau_{ik}^{(r)} y_{ip} \right)^2 \right)^{\frac{1}{2}} \leq \gamma_1 M_p^{(r)} \\ \tilde{\mu}_{kp}^{(r+1)} - \gamma_1 \frac{\hat{\mu}_{kp}^{(r)} \Sigma_{k,pp}^{(r)}}{\|\hat{\mu}_{\cdot p}^{(r)}\| \sum_{i=1}^{n} \tau_{ik}^{(r)}} & \text{otherwise} \end{cases} \tag{21}$$

for $k = K+1, ..., K+Q$ where $M_p^{(r)} = \max_{k=1,...,K+Q} \hat{\Sigma}_{k,pp}^{(r)}$. Note that the norm $\|\hat{\mu}_{\cdot p}^{(r)}\|$ should be computed based on all the $K+Q$ components.

Finally, the outer loop should iterate over the explained above fixed-background estimation procedure. During the subsequent iteration, the MAES model is fitted to the background data with the slight change. Instead of equation 10 the formula 21 for components $k = 1, ..., K$ should be used in order to borrow the knowledge of the already found signal components. After the background is refitted the second step of the anomaly detection should be performed. As a result, the background mean estimates from the initial step (even the one that had been shrunk to 0) could change if a significant signal is present in the mixed data.

4 Experimental analysis on toy simulated data

4.1 Goals of the analysis

Collections of artificial data were generated in order to understand the performance of the proposed methodology in terms of classification in an unsupervised setting (i.e. to estimate the background distribution) and in a semi-supervised setting (to find a signal process). The simulations touch different aspects of the algorithm performance with respect to:

- Different implementations of the proposed approach to handle variable selection. Within the penalized model-based clustering approach we consider and test two different scenarios to perform variable selection. The first (M1) is a natural result for the MAES algorithm described above and the model is obtained by a model fit to the data of the full dimension P. The used penalties serve as regularization and variable selection.The second implementation

(M2) also relies on the MAES algorithm but is constructed in two passes. First the regular MAES is fitted to data of the full dimension P as the mean for variable selection. In the second pass the model is fitted again but based only on informative variables that had been selected in the former pass. The M2 model is expected to have better prediction performance because in the reduced space of informative variables smaller penalty (or even none at all) should be applied to the data. This results in smaller bias of the parameter estimates and as well the estimates are more stable. On the other hand, this approach might suffer from a possible error propagation and a decrease of classification performance if some of the informative variables are incorrectly removed. For the simulation setting, when MAES is mentioned it is assumed that the M1 method is used because it is a natural consequence of the penalized approach. The M2 method has less theoretical bases, could suffer from error propagation and is against the paradigm of joint parameter estimation and variable selection, however it is worth to test and compare its performance.

- Comparison with the competing models, i.e. the parsimonious family of models introduced in Fraley and Raftery [17] and also Bouveyron et al. [22], application of the PCA prior to the parameter estimation according to Kuusela et al. [16] and the penalized log-likelihood approach of Pan and Shen [26].

- Varying configurations of the background and possible signal. For this reason, evaluation has considered different degrees of separation between the mixture components and different mixing proportions of the components.

4.2 Simulation settings

The simulated data are generated from mixtures of a few Gaussian distributions as a proof of concept to enable easier analysis for simple examples. Additionally, the true parameters of the generating model are chosen so that the particular features of the algorithm performance could be explored.

For the background data the following model has been considered:

$$f_B(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k),$$

where

- K is chosen in $\{2,3\}$.

- Mean vectors are set as $\boldsymbol{\mu}_k = (1,...,1,0,...,0)' * mult$ with

$$mult = \begin{cases} (m,-m) & \text{for the } K=2 \text{ setting} \\ (m,0,-m) & \text{for the } K=3 \text{ setting} \end{cases}$$

for m in $\{0.1, ..., 0.8\}$ and $\sum_{p=1}^{P} \boldsymbol{\mu}_{kp} = \frac{P}{2} mult.$

- Covariance matrices are set as

$$\Sigma_k = \begin{pmatrix} \Sigma_{k1} & 0 & 0 \\ 0 & \Sigma_{k1} & 0 \\ 0 & 0 & I_8 \end{pmatrix}$$

where $\Sigma_{k1} = P_k D_k P_k'$ with

$$P_1 = \begin{pmatrix} 1 & 0 & -1 & 1 \\ 1 & \sqrt{2} & 1 & 0 \\ 1 & -\sqrt{2} & 1 & 0 \\ -1 & 0 & 1 & 2 \end{pmatrix}, \quad P_2 = \begin{pmatrix} -1 & 0 & 1 & 1 \\ 1 & -\sqrt{2} & 1 & 0 \\ 1 & \sqrt{2} & 1 & 0 \\ 1 & 0 & -1 & 2 \end{pmatrix}, \quad P_3 = I_{\frac{P}{4}}$$

$$D_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.15 & 0 \\ 0 & 0 & 0 & 0.12 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.1 \end{pmatrix}, \quad D_3 = I_P$$

- Component proportion are set to

$$\pi = \begin{cases} (0.5, 0.5) & \text{for the } K = 2 \text{ setting} \\ (0.5, 0.3, 0.2) & \text{for the } K = 3 \text{ setting} \end{cases}$$

- Data sizes n in $\{250, 500\}$ and dimension $P = 16$.

Given this set of the parameters the variables $9 - 16$ are truly uninformative.

For anomaly detection simulations, the background data are generated using $K = 2$ Gaussian components with parameters specified as above. The signal process is simulated by a single Gaussian component with parameters specified as for the third background component described above. The experimental data is generated for different proportions of signal events λ in $\{0.2, 0.1, 0.05\}$. The generated background and experimental data sizes are $n = m = 500$. As the signal component is placed between the background components, for testing purpose (when it is specified), the signal mean for the 14^{th} variable (which is uninformative in respect to the background set) is specified to be non-zero.

4.3 Details

1. The simulations were performed in R environment for statistical computing [36] version 3.4.2.

2. Competing approaches for the model-based clustering of the background data

- Model-based clustering without any restriction on constraint on the covariance parametrization based on the Fraley and Raftery [17] with the associated implementation [37].
- Model-based clustering approach using the unrestricted component covariance matrices with prior reduction of the data dimension to the first two principal components as proposed in [15, 16]
- A family of parsimonious mixture models as proposed in Bouveyron et al. [22] and implemented in the HDclassif R package [38].
- A penalized approach of Pan and Shen [26] with component covariances restricted to identity matrices.

3. Model selection - an important aspect of a model-based clustering and the penalized approach is proper determination of the number of Gaussian components K that should be used for modeling and as well the values for the regularization parameters γ_1 and γ_2. The commonly used approach for the unpenalized mixtures model is to use the Bayesian Information Criteria (BIC) which was defined by Schwarz [35] as

$$BIC_1 = -2logL(\hat{\theta}) + log(n) * d,$$

where d is the number of free parameters in the model, L is the model likelihood and $\hat{\theta}$ are the MLE. According to the criterion a model that minimizes the BIC should be selected as an optimal trade off between the goodness of fit and the model complexity. However for the penalized model some parameters are shrunk and should not be considered any longer as free. Following Pan and Shen [26] and Bühlmann and Van De Geer [32] motivated by a study in context of the penalized regression of Efron et al. [24] a modified BIC criterion could be used. It is formulated as

$$BIC_2 = -2logL(\hat{\theta}) + log(n) * d_{eff},$$

where d_{eff} is the effective number of parameters (parameters not shrunken to the fixed value) and $\hat{\theta}$ are MPLE. Despite the lack of rigorous theoretical arguments justifying the use of the modified BIC criterion, practically in simulations and application the criterion serves very well.

In practice, the minimum value for the BIC is found based on an extensive grid search over the parameters space. In a first phase the grid is coarse and later the search is performed on a more compacted grid trimmed appropriately based on the previous pass. With special care the grid boundaries are chosen in order to select the global minimum for BIC function not only the local one. BIC values are discontinuous with respect to regularization parameters, however as they increase sufficiently the associated BIC values are so

large, that it is clear that the global minimum should be found for lower regularization parameters. Based on this observation the grid points are set.

For the considered simulation for all the models, we consider that the number of Gaussian components is known a priori, so that the results are not influenced by incorrect specification of K.

4. Performance evaluation of the model-based clustering methods are based on the Adjusted Rand Index [39]. In general the Rand Index is a measure of accuracy of correctly classified observations [40]. The adjusted version corrects value of the former Rand Index so that for a random classification the expected index value is equal to 0.

5. It should be noted that the EM algorithm does not guarantee to find the correct maximum of the objective function (penalized log-likelihood) as it could find one of local maxima. It strongly depends on the starting values and for this reason the algorithm should be run a few times from the different initialization points $\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}$ and $\Sigma_k^{(0)}$ for $k = 1, ..., K$. Additionally, maximization of the log-likelihood function of the Gaussian mixture model is not a well-posed problem, i.e. the likelihood tends to the infinity if one of the components becomes a singularity that models a single observation. These issues become more apparent especially for the datasets with a higher number of dimensions. For these reasons, the algorithm is run from many starting points and a stopping rule is implemented if the algorithm does not converge.

   For the grid search it is clever to use the so-called warm-starts as initialization values, i.e. the estimates of the model with smaller regularization parameters. For new regularization parameters the MPLE are not far from the previous run, hence it is often sufficient to run the algorithm from a single starting point. Given the warm-start the algorithm converges much faster (in about 10 iterations versus about 50 needed for arbitrary starting points) and does not need to be run multiply times.

6. Note that the datasets X and Y could differ in size and prior distribution they have been produced. For this reason, regularization parameters γ_1 and γ_2 should be selected separately for each of them. We propose to select the regularization parameters for the background once (despite the further loop iterations) when the initializing background model is obtained. The regularization parameters for the second stage of the PAD algorithm and the number of signal components Q should be found using a grid search and an optimization criterion.

4.4 Results and comments

4.4.1 Model-based clustering

Let us consider a scenario where the data are generated from the Gaussian mixture with two equally weighted components and parameters specified as described above.

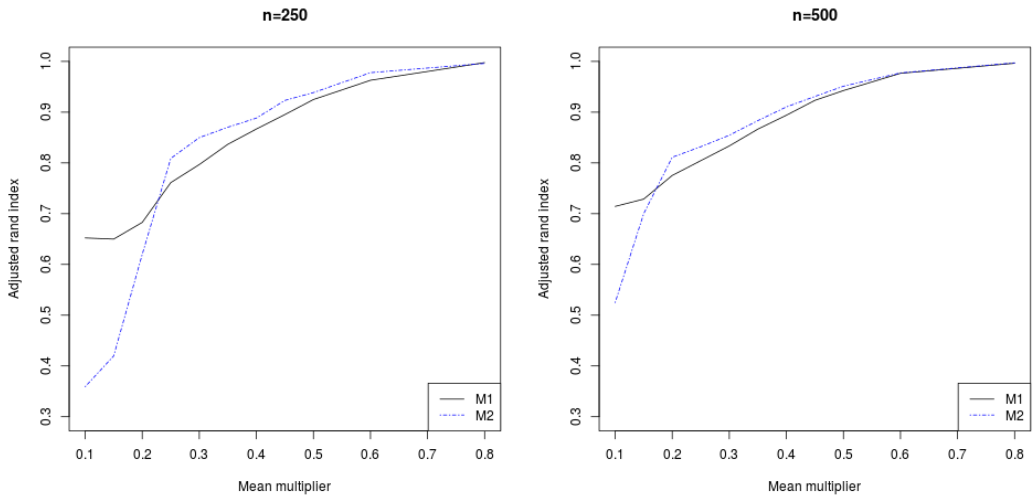


Figure 1: Performance comparison for the two types (M1 and M2) of variable selection methods for the proposed model-based clustering algorithm for the data of size $n = 250$ and $n = 500$ with different values of the mean multiplier. The results are based on the average performance on 30 simulated datasets.

Comparison of variable selection methods (figure 1) for MAES suggests that the M2 model performs best in terms of the adjusted Rand index. However, for the small components separation and $n = 250$, the variable selection methodology incorrectly removes the informative variables. This decreases the performance of the M2.

For the simple scenario of the balanced data the introduced MAES algorithm has as superior performance in comparison to the other methods (figure 2), especially for the data with the lowest size $n = 250$ and small true components separation (the most difficult cases). As the separation increases and classification becomes easier, all but the method of Vatanen et al. [15] have comparable performance. Inferior performance of Vatanen et al. [15] method is presumably due to improper variable selection as for this case the first two principal components explain only a small fraction of the total variability.

Subsequently, let us consider scenario when the data are generated from the Gaussian mixture with 3 components and parameters as specified at the beginning of the section. In this case mixture components are of different proportions and this is an additional difficulty affecting the performance of clustering algorithms.

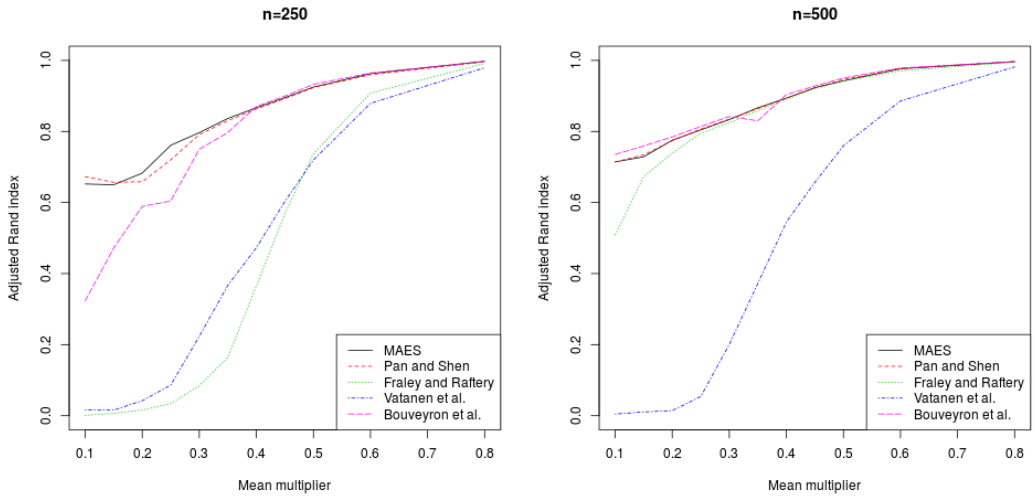


Figure 2: Performance comparison for the 5 model-based clustering methods for the data of size $n = 250$ and $n = 500$ for different values of mean multiplier $mult$ (true separation of classes).

In comparison to the previous simulations where the components had equal proportions, for 3 unbalanced components we naturally observe much lower scores of the adjusted Rand index. Performance for variable selection approaches (models M1 and M2) presented above supports the M2 approach (figure 3). The hierarchy of model performances changes slightly between the balanced and unbalanced scenarios, but MAES, Pan and Shen and Bouveyron et al. are still leading, especially for the larger data size (figure 4). For the smaller data size and small separation MEAS algorithm has far better performance than the competing models.

4.4.2 Anomaly detection

For anomaly detection simulations the respective background and experimental (background + signal) datasets are respectively with distributions specified in subsection 4.2. Each pair of datasets is properly transformed. The background data are standardized according to their sample mean and variance. The experimental data are standardized as well but according to the background sample mean and variance, hence in general mixed data mean is different from 0 especially if a strong signal is present.

The simulations are performed as follow. First, for signal fraction $\lambda = 5\%$ the anomaly detection algorithm is run for varying $mult$ parameter (separation of the background components). The results are presented in the first three rows of table 1. For the high components separation, there is no problem in signal/background classification however for $mult = 1$ the signal misclassification error raises to 76%. In row 4 and 5 is presented the performance for higher signal proportion in the ex-

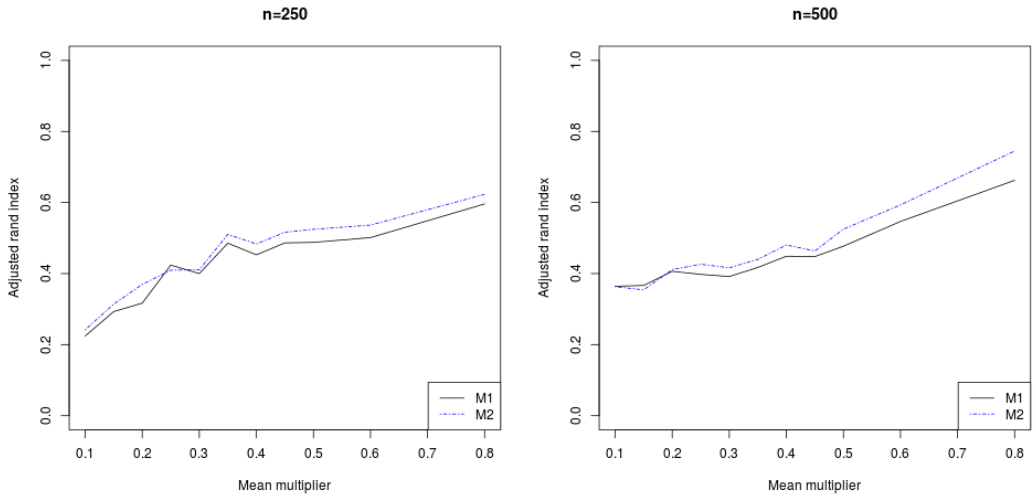


Figure 3: Performance comparison for the two types (M1 and M2) of variable selection methods for the proposed model-based clustering algorithm for the data of size $n = 250$ and $n = 500$ with different values of the mean multiplier. The results are based on the average performance on 30 simulated datasets.

perimental data by having fixed separation. Only the high signal proportion allows for a decrease of the misclassification error. For the next three following simulations, the true signal means for the 14^{th} variable (which is uninformative according to background data) is changed from 0 to 3, so that the signal strongly exhibits as a deviation from the background distribution. This results in much better signal classification. It is due to the fact that true signal observations lie between the high background distribution probability domain. Figure 5 presents the fitted density to the mixed data. Hence the signal classification error is not as low as one could wish, but the signal component is estimated close to the true one.

Finally, the feature space was reduced according to both criteria described above and a new model was refitted in the reduced space (M2 model approach). The results are presented in the last 6 rows of table 1. In each case, the Adjusted Rand Index for classification is increased with respect to the approach without dimensional reduction. The improvement is a result of smaller bias caused by penalty function applied to all the informative and uninformative variables. After dimensionality reduction the penalty is much softer (or even not present) as the uninformative variables could not be present in the data any more.

4.5 Simulation conclusions

In unsupervised setting, the MAES algorithm was tested against the competitors to access classification performance. Half of the variables of the simulated data were taken as uninformative, which favors the proposed algorithm which employs vari-

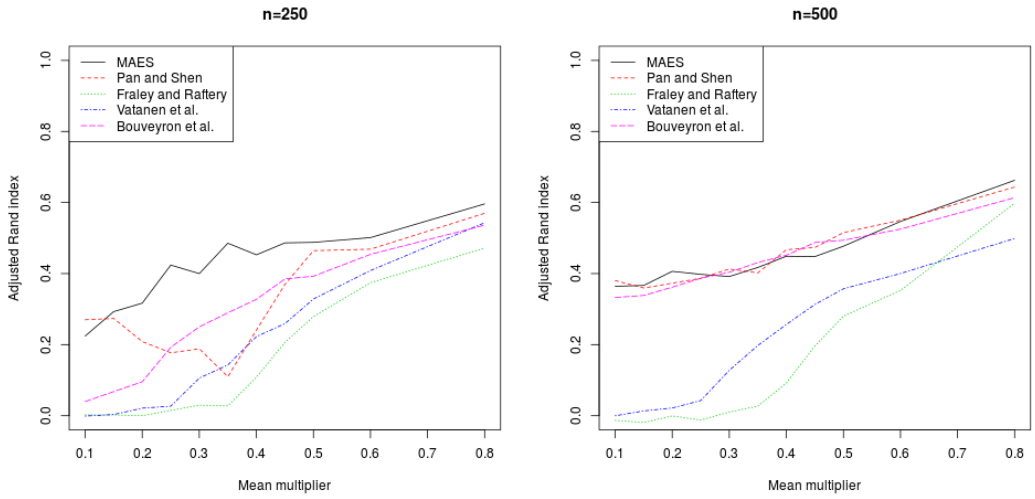


Figure 4: Performance comparison for the 5 model-based clustering methods for the data of size $n = 250$ and $n = 500$ for different values of mean multiplier $mult$ (true separation of classes). By MAES is always understood the M1 variable selection model for clarity (despite superior performance of M2 for the given simulation schema).

able selection jointly with parameter estimation. For this reason Fraley and Raftery [17] and Vatanen et al. [15] approach turns out to perform worst. Difference between algorithm performance is most visible for the most difficult cases that support the MAES. For easier cases or for bigger data the algorithms tend to perform similarly.

We considered two methods for variable selection (M1 and M2). The simulations favour M2 model, it has superior performance versus M1, unless the background components are hardly separated from each other, which results in the removal of informative variables and a consequent worsening of the performance.

For semi-supervised anomaly detection, the PAD algorithm is robust against the separation of the background and signal components. It was presented that if for some uninformative variable according to the background data a strong signal is exhibited ($\mu_{s,14} = 3$), then the variable is not removed and classification is improved. Finally, the M2 method was shown to perform better than the M1 for all the tested cases. We presume that the high penalty used for removing many of uninformative variables causes bias on other parameters leading to poorer classification.

The PAD algorithm was compared with the Kuusela et al. [16] approach for which the simulated data is reduced to the first two principal components. For the given case, the first two components represent only 21% of total data variability. In order to be able to explain at least 80% of the total variability 11 principal components (out of 16) should be selected. Hence, taking only the first two components puts under the risk that any possible signal deviation is omitted by the PCA

Table 1: Anomaly detection results compared for varying background distribution, signal fraction λ and signal true mean μ_s. Also the dimension reduction technique (model M2) is applied and its results are shown in the last 6 rows. There are 8 uninformative variables, but true signal mean vector is equal to **0** (unless the signal mean for the 14^{th} variable ($\mu_{s,14}$) is increased to 3).

$\mu_{s,14}$	$mult$	λ	model	misclassified events	misclassified signal events	Rand Index	misclass. error
0	2	0.05	M1	0	0	1.000	0.000
0	1.5	0.05	M1	15	0	0.937	0.030
0	1	0.05	M1	19	19	0.926	0.038
0	1	0.10	M1	42	42	0.839	0.084
0	1	0.20	M1	37	20	0.825	0.074
3	1	0.05	M1	18	18	0.929	0.036
3	1	0.10	M1	22	21	0.908	0.044
3	1	0.20	M1	14	8	0.928	0.028
0	1	0.05	M2	13	13	0.948	0.026
0	1	0.10	M2	15	14	0.936	0.030
0	1	0.20	M2	17	15	0.897	0.034
3	1	0.05	M2	8	7	0.967	0.016
3	1	0.10	M2	8	7	0.964	0.016
3	1	0.20	M2	11	9	0.945	0.022

procedure. Indeed, in table 2 we see that the algorithm proposed by Kuusela et al. has better performance in terms of Rand index for $\mu_{s,14} = 0$ than for $\mu_{s,14} = 3$ - that is the case when signal distribution diverge more from the background (rather the opposite behavior one would expect). The PAD algorithm of M1 type has slightly worse performance for $\mu_{s,14} = 0$ than Kuusela's but better for $\mu_{s,14} = 3$. However, the PAD algorithm of M2 type is superior to others for all the tested cases.

5 Application to a model-independent search for new physics

For a proof of concept of the method in the context of a physics analysis, we have produced simluated datasets. Specifically, we are interested in the signature containing two jets in the final state (known also as "dijet" final states), denoted as "jj" coming from proton-proton collisions at the LHC; for such signatures, background and hypothetical signal collision events have been generated as described below. A simple analysis selection is then performed on the samples and finally the anomaly detection method is tested.

Table 2: Anomaly detection results based on the algorithm introduced by Kuusela et al. [16] for the data reduced to the first two principal components. The simulation is performed against varying signal fraction λ and signal true mean μ_s. For comparison reasons, adjusted Rand indexes for PAD algorithm for both M1 and M2 models are also included.

$\mu_{s,14}$	$mult$	λ	misclass. events	misclass. sig. events	Rand Index	PAD M1 Rand Index	PAD M2 Rand Index	classif. error
0	2	0.05	3	3	0.987	1.000	-	0.006
0	1.5	0.05	8	8	0.967	0.937	-	0.016
0	1	0.05	17	17	0.932	0.926	0.948	0.034
0	1	0.10	24	22	0.899	0.839	0.936	0.048
0	1	0.20	34	22	0.838	0.825	0.897	0.068
3	1	0.05	25	25	0.902	0.929	0.967	0.050
3	1	0.10	30	28	0.877	0.908	0.964	0.060
3	1	0.20	41	33	0.811	0.928	0.945	0.082

5.1 MC samples

We generated a set of MC samples using standard simulation software for high-energy collisions. The simulated phenomena correspond to proton-proton collisions at a center-of-mass energy $\sqrt{s}$ = 13 TeV and measured by a simplified ATLAS detector simulation implemented in the Delphes package. The main source of Standard Model (SM) background in our signature is the production of two jets via QCD processes, namely from the production of a pair of gluons, a pair of quarks, or a gluon and a quark. The simulated signal corresponds to the production of a stop quark decaying into two jets, in the R-parity violating Minimal Supersymmetric Model (RPV-MSSM) [41, 42], using the package in [43]. A more in-depth description of the simulations performed for signal and background follows.

5.1.1 Signal

As a benchmark for testing the anomaly detection algorithm, we produced a sample of 5 * 10^5 stop quark signal events. The hard process was simulated using MadGraph 5, where we used the four-flavor scheme and nn23lo1 [44] to model the proton PDFs. The resulting events were then ported to Pythia 8.230 for showering, decay and hadronization. All default parameters from MadGraph 5.2.6.0 were used except the value of the pseudorapidity, restricted to $|\eta| < 2.5$.

Within the RPV-MSSM model used, the production and decay (to two jets) of a resonant stop happens via a single Feynman diagram in a four-flavor scheme, as depicted in figure 6.

The production cross section reported by MadGraph 5 is 18.011 ± 0.003 pb.

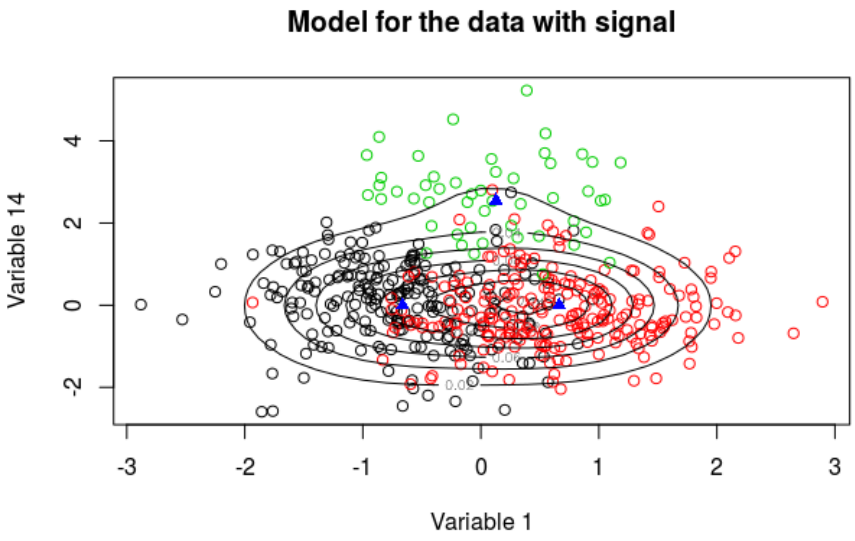


Figure 5: Estimated distribution for the the mixed data with signal fraction $\lambda = 5\%$ and the true signal mean for the 14^{th} variable increased to 3. The blue triangles are estimated Gaussian mean components. The plot shows the mixed data in space spanned by the 1^{st} and 14^{th} transformed variable. Green point correspond to signal, while black and red to the background components.

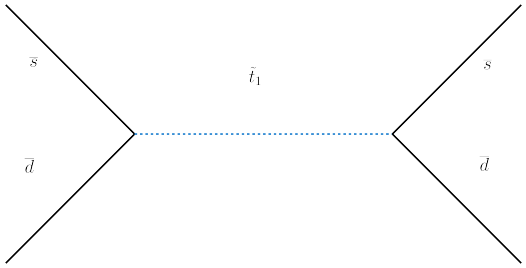


Figure 6: Feynman diagram for the production of a stop quark decaying into two light quarks, in the RPV-MSSM [41, 42].

5.1.2 Background

The production of $5 * 10^5$ dijet events from QCD processes was performed using the same MadGraph 5 and Pythia 8 versions as the ones for the signal. There are many ways in which QCD interactions can lead to a pair of jets in the hard process, namely all possible diagrams that contain two gluons, a quark and a gluon or two quarks in the final state. In a four-flavor scheme at tree level, there are 65 such processes. The production cross section corresponding to the QCD background reported by MadGraph 5 is: 3.382 ± 0.001 mb.

5.1.3 Detector simulation

Both signal and background event samples have been passed through a fast detector simulation using the DELPHES 3.4.1 software. We have kept all default parameters except the jet cone parameter $\Delta R = 0.4$, and used the ATLAS detector card that comes with the software distribution.

5.2 Event selection

The signal and background simulated samples are then analyzed and an event selection is applied. A set of requirements are imposed on the object properties and event variables, inspired by realistic experimental analyses, in order to e.g. mitigate detector effects, simulate trigger selection. Furthermore, several features (variables) are extracted and calculated. The features of the events that pass the selection comprise the input of the anomaly detection algorithm. Typical values of object selection are already included in the DELPHES ATLAS detector card; at this level, the only additional requirement we impose is that the event contains only two jets and each of them has a transverse momentum of 20 GeV or more.

Given that we are performing a model-independent search, the selection requirements are not optimized for any particular signal, even if we could, in principle, devise such a procedure for the stop production described above.

The variables extracted and calculated for each event are the following:

- Transverse momentum for each of the jets: $p_T(j_1)$, $p_T(j_2)$.
- Reconstructed invariant mass of the dijet system: $M_{\text{inv}}(j_1, j_2)$.
- Missing transverse energy: E_T^{miss}.
- Angular distance of the two jets in the $\eta - \phi$ plane: $\Delta R(j_1, j_2)$.
- Sphericity and centrality, as defined in [45].

Normalized distributions for the background and signal described above can be found in figure 7.

5.3 Method performance

Marginal distributions for the produced background and signal samples are unimodal and most of them stand out to be heavily skewed. Certainly, skewed distributions could be approximated by the Gaussian mixtures but this requires to use many components that could lead to over-parametrization. Hence to obtain a more accurate model, there is a need for proper data transformation, so that the marginal variables distributions are as close to the normal distribution as possible. For this reason, Tukey's Ladder of Powers transformation is applied for each variable of the

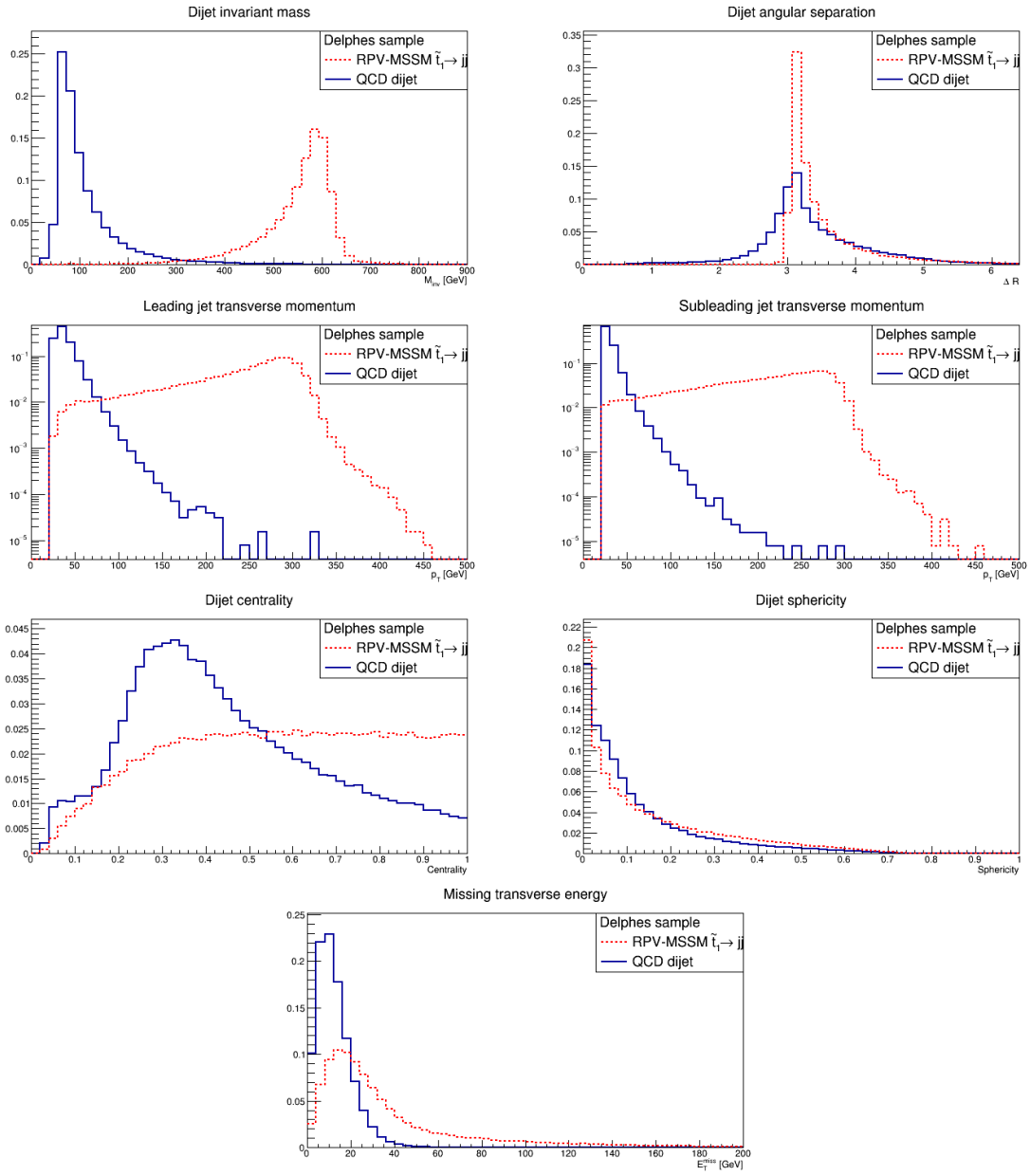


Figure 7: Normalized distributions of signal and dijet background for seven kinematic and angular variables.

background data subsequently [46, 47]. Formally, separate transformations are applied to each variable based on a properly chosen parameter ρ_k, that is:

$$x_{ik}* = \begin{cases} x_{ik}^{rho_k} & \text{if } \rho_k > 0 \\ \log(x_{ik}) & \text{if } \rho_k = 0 \\ -x_{ik}^{rho_k} & \text{if } \rho_k < 0 \end{cases}$$

for $i = 1,...,n$ and ρ_k selected separately for each k in $1,...,K$. Parameter ρ_k is selected based on the background and the same transformation is applied to simulated experimental data which are a mixture of the background and signal events. Afterwards, the background data are scaled according to their sample mean and variance. Scaling is also applied to the experimental sample but according to the sample mean and variance of the background.

Although the univariate background data distributions after the transformations are unimodal and resemble Gaussian, it would be an oversimplified model if the background density would be estimated by a single multivariate Gaussian component. If the 2-dimensional data distributions are analyzed, they do not exhibit elliptical Gaussian shape. One reason for this is preprocessing of the produced collision events from which only a group of potentially interesting are selected. Also ordering of event variables (i.e. $p_T(j_1) \geq p_T(j_2)$) introduces strict limits onto feature space. Finally the background is produced by many underlying processes that results in complex distribution. For this reasons the background distribution needs to be modelled by a mixture of many components.

The data subset of size $n = 4000$ was taken from the background sample. The optimal number of Gaussian components to model the background was selected to be equal to 10 based on the described modified BIC criteria. In total there are 359 parameters to be estimated. Due to the background complexity, the regularization parameter γ_1 was chosen to be 0, hence all the variables play an important role in the model-based clustering. The second regularization parameter γ_2 is positive and plays an important role for component covariance matrices regularization. As the first necessary condition for variable selection cannot be met, the approaches for variable selection (M1 and M2) are equivalent.

The background model serves as an initialization point for the PAD algorithm. However for $\gamma_1 = 0$, subsequently found signal components do not influence the background estimation through the penalty, hence the PAD algorithm is reduced to the fixed-background model with regularization on the component covariance matrices.

For the varying proportion of signal observations λ in the experimental data, the PAD algorithm is tested. For each λ, 50 datasets of size $n = 4000$ are sampled from the respective background and signal datasets so that the simulated experimental data are constructed. In following, we search only for a single signal component. This is an already challenging task, and its results are sufficient to provide insight in the possible presence of an anomalous process in the data.

The algorithm is able to find and estimate a signal component corresponding to an unknown process not present in the background data for the signal proportion $\lambda = 0.05$ without any problem. However, for smaller signal proportion λ problems start to arise. For many algorithm runs estimates for the signal proportion $\hat{\lambda}$ is about 0.80. This is a frequent example of EM algorithm finding local (instead of global) maxima of the penalized log-likelihood. For such the case the likelihood of the background model is higher, which implies that the found signal component is

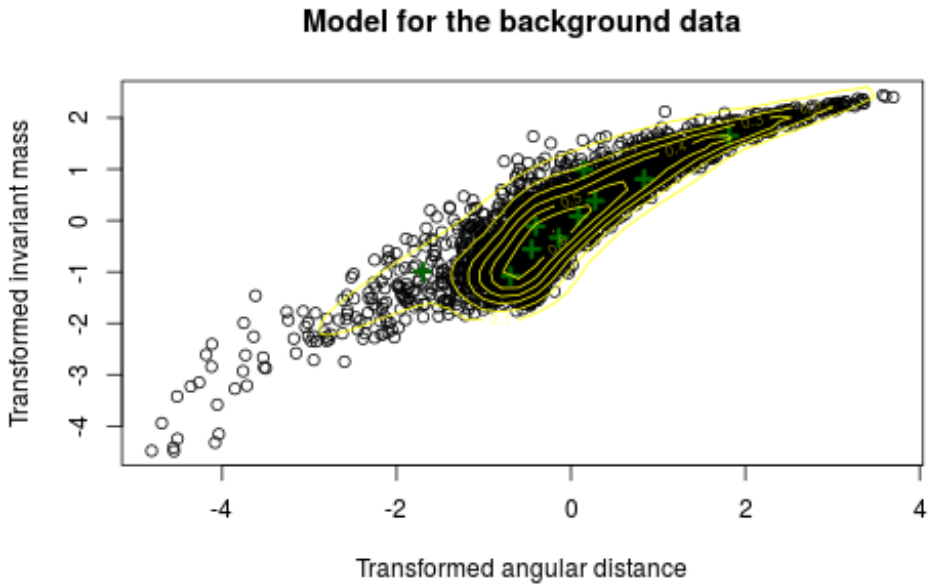


Figure 8: The plot of the two variables for the transformed background data to which the 10 component model was fitted. The green crosses are the component means and the contour reflects the estimated distribution of the sample.

spurious.

In order to find a non-spurious signal component, the algorithm should start with different initialization values as the EM algorithm is known to be sensitive to the starting point. The best practice is to use signal events for a starting value of signal component mean. In this way it is more likely that a signal component is found and the global maximum is reached. However in general, it is not known which observations are generated by the signal process and hence, the algorithm should be run multiple times starting from different initial values. The runs are performed as long as some fitted signal component increases the model likelihood with respect to the pure background model likelihood on the mixed data. In the following, we consider not only an increase of the likelihood but rather more conservative decrease of the modified BIC criteria.

The simulation results are presented in tables 3 and 4. The first one shows that the estimated fraction of the signal events is accurate, although a little biased toward the higher values. With decreasing true signal proportion λ the number of misclassified signal events rises 4. This affect the mean adjusted Rand index.

For signal proportions λ lower than 0.013, the algorithm has severe problems to fit a proper signal component. Note that already for $n = 4000$ and $\lambda = 0.013$ there are only 52 signal observations that the algorithm is sensitive to signal proportion $\lambda = 0.015$ is already satisfactory. However, it should be noted that distribution of

Table 3: Summary of the anomaly detection results performed by the PAD algorithm for datasets with different signal proportions λ. For each λ 50 datasets were generated in order to obtain also the standard deviations of the average estimates (in brackets).

True signal proportion λ	Average estimate $\hat{\lambda}$	Average adj. Rand index
0.050	0.05099 (0.00021)	0.9526 (0.0015)
0.030	0.03093 (0.00025)	0.9364 (0.0022)
0.020	0.02129 (0.00043)	0.9203 (0.0044)
0.015	0.01658 (0.00033)	0.8997 (0.0062)

Table 4: Average percent of misclassification errors for the anomaly detection results performed by the PAD algorithm for the datasets with different signal proportions λ for the mixed data. For varying parameter λ, 50 datasets were generated in order to obtain the average estimates and their standard deviations (in brackets).

Signal prop. λ	Average percent of misclassified signal observations	Average percent of misclassified background observations
0.050	2.39 (0.31)	0.38 (0.11)
0.030	3.87 (0.41)	0.24 (0.07)
0.020	6.66 (0.94)	0.17 (0.05)
0.015	10.70 (0.99)	0.14 (0.05)

the simulated signal is quite different from the background distribution, that causes the anomaly detection problem to be trivial.

6 Conclusions

We have presented a novel method for collective anomaly detection that makes use of a semi-supervised approach and is able to jointly perform variable selection and parameter estimation (referred as PAD). Motivations for the algorithm development derive from the needs of New Physics model-independent searches at the LHC. PAD algorithm makes use of Gaussian mixture models as the mean for data density estimation and further classification of anomalous observations. It extends the idea of fixed-background model that has been previously applied in the same context. The main contribution with respect to the mentioned model is to perform proper feature selection so that variables that exhibit signal deviation are not removed. Dimensionality reduction is performed based on the penalized approach stemmed from the unsupervised penalized model-based clustering methods.

As a proof of concept, we have tested the PAD algorithm in two scenarios: one

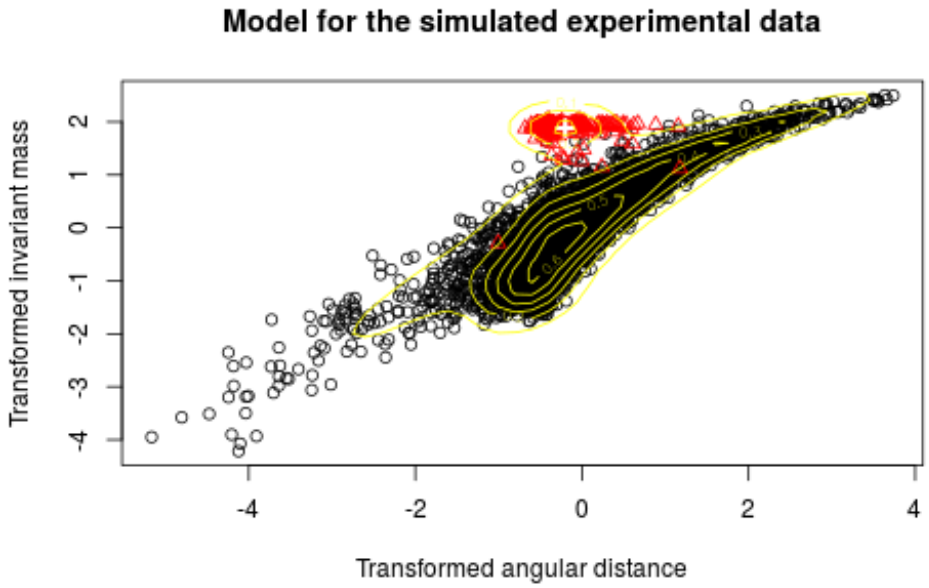


Figure 9: The plot of angular distance and invariant mass for the transformed simulated experimental data. Red dots are red triangulates are respectively background and signal observations. The white crosses is the signal component mean and the contour reflects the estimated distribution of the whole sample.

using idealized synthetic data and another one using data produced from MC simulators of high-energy collisions at the LHC, reconstructed as real data by the ATLAS detector. In the former case, the developed unsupervised model-based clustering method (called MAES) was tested. As well respective anomaly detection algorithm (PAD) was tested against different gaols. The performance of both PAD and MAES was tested against other competing models with good results. The introduced dimensionality technique turned out to be successful in removing uninformative variables. The removal leads to better predictions. In the latter case, the method was used to detect an anomalous set of signal events coming from a simulated new physics process that yields two jets in the final state. The PAD algorithm enables to detect presence of an anomalous process, precisely estimate its proportion and estimate signal distribution.

7 Acknowledgement

The authors are grateful to the European Union's Horizon 2020 research and innovation program under grant agreement number 675440 for financial support.

8 Appendix

8.1 Pseudo-code for MAES algorithm

Function input:

- Background data - x
- Number of fitted components - K
- Maximal number of iteration allowed - n_iter
- Regularization parameters - γ_1 and γ_2
- Warm start initialization values for μ_k, π_k, Q_k and D_k for $k = 1;...,K$ (the default is NA)
- Stopping parameter ν

Algorithm

1. $P =$ dimension of x
2. $n =$ size of x
3. IF (any of $\pi_k^{(0)}$, $\mu_{k\cdot}^{(0)}$, $Q_k^{(0)}$ or $D_k^{(0)}$ for $k = 1,...,K$ is NA)
   Initialize $\pi_k^{(0)} = \frac{1}{K}$ $Q_k^{(0)} = I_P$, $D_k^{(0)} = I_P$ and let $\mu_{k\cdot}^{(0)}$ be the centers of k-mean algorithm
   ELSE Check if initialization parameters are of right dimension, $\sum_{k=1}^{K} \pi_k = 1$
4. LikeLast = $-\infty$
5. LikeNew = the likelihood for the 0^{th} step
6. Iterator $r = 0$
7. While(|LikeLast-LikeNew| > ν AND r <n_iter)
   (a) Covariance matrices
   For (k in 1:K)
   $$\hat{\Sigma}_k^{(r)} = \hat{Q}_k^{(r)} \hat{D}_k^{(r)} \left(\hat{Q}_k^{(r)}\right)'. \tag{22}$$
   (b) Posterior probability
   For (i in 1:n, l in 1:K)
   $$\tau_{il}^{(r)} = \frac{\pi_l^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_l^{(r)}, \hat{\Sigma}_l^{(r)}\right)}{\sum_{k=1}^{K} \pi_k^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k^{(r)}, \hat{\Sigma}_k^{(r)}\right)} \tag{23}$$

(c) Components proportions
For (k in 1:K)

$$\hat{\pi}_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ik}^{(r)}. \tag{24}$$

(d) MLE estimates
For (k in 1:K)

$$\tilde{\mu}_{kp}^{(r+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(r)}x_{ip}}{\sum_{i=1}^{n}\tau_{ik}^{(r)}} \tag{25}$$

$$\tilde{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(r)} * (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})'}{\sum_{i=1}^{n}\tau_{ik}^{(r)}} \tag{26}$$

(e) For (p in 1:P)

$$M_p^{(r)} = \max_{k=1,\ldots,K}\hat{\Sigma}_{k,pp}^{(r)}$$

(f) For (k in 1:K) perform eigenvalue decomposition

$$\tilde{\Sigma}_k^{(r+1)} = \hat{Q}_k^{(r+1)}\tilde{D}_k^{(r+1)}\left(\hat{Q}_k^{(r+1)}\right)'. \tag{27}$$

(g) FOR (k in 1:K,p in 1:P)

- IF $\left(\sum_{k=1}^{K}\left(\sum_{i=1}^{n}\tau_{ik}^{(r)}x_{ip}\right)^2\right)^{\frac{1}{2}} \leq \gamma_1 M_p^{(r)}$

$$\hat{\mu}_{kp}^{(r+1)} = 0$$

- ELSE

$$\hat{\mu}_{kp}^{(r+1)} = \tilde{\mu}_{kp}^{(r+1)} - \gamma_1\frac{\hat{\mu}_{kp}^{(r)}\Sigma_{k,pp}^{(r)}}{\|\hat{\mu}_{\cdot p}^{(r)}\|\sum_{i=1}^{n}\tau_{ik}^{(r)}}$$

(h) FOR (k in 1:K)

- $$\bar{D}_{k,pp}^{(r+1)} = \frac{-n\hat{\pi}_k^{(r+1)} + \sqrt{\left(n\hat{\pi}_k^{(r+1)}\right)^2 + 8\gamma_2 n\hat{\pi}_k^{(r+1)}\tilde{D}_{k,pp}^{(r)}}}{4\gamma_2}. \tag{28}$$

- To surpass numerical instability it is used
For(p in 1:P)
IF $\bar{D}_{k,pp}^{(r+1)} < 0.0005$ then $\bar{D}_{k,pp}^{(r+1)} = 0.0005$.

(i) FOR (k in 1:K,p in 1:P)

i. $\epsilon_k =$ mean ($P-p+1$ smallest eigenvalues of $\bar{D}_k^{(r+1)}$).

ii. For arbitrarily $\alpha = 0.05$

$$Logical = \frac{\bar{D}_{k,pp}^{(r+1)}}{\epsilon_k} < 1 - \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2p}} \vee \frac{\bar{D}_{k,PP}^{(r+1)}}{\epsilon_k} > 1 + \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2p}}$$

where $z_{\frac{\alpha}{2p}}$ is a normal distribution quantile.

iii. If $(Logical = true)$ then the smallest $P-p+1$ eigenvalues $\hat{D}_{k,pp}^{(r+1)} = ... = \hat{D}_{k,PP}^{(r+1)} = \epsilon_k$
BREAK the inner loop over 1:P.

iv. else $\hat{D}_{k,pp}^{(r+1)} = \bar{D}_{k,pp}^{(r+1)}$.

(j) LikeLast = LikeNew

(k) LikeNew = likelihood for the $(r+1)^{th}$ step.

8. Return all the parameters value for the last step and an error code if n_iter was reached (issues with convergence in n_iter steps)

The algorithm should be run for different values of γ_1, γ_2 and K to perform the optimal selection, that is the one that minimizes the modified BIC criteria

$$-2logL(\hat{\Theta}) + \log(n)d_{eff}$$

where d_{eff} is effective number of degrees of freedom.

8.2 Pseudo-code for anomaly detection PAD algorithm

Function of:

- Mixed data - y
- Number of signal components - L
- Maximal number of iteration allowed - n_iter
- Regularization parameters - γ_1 and γ_2
- Warm start initialization values for λ, μ_k, π_k, Q_k and D_k for $k = 1;...,K+L$ (the default is NA)
- Stopping parameter ν

Algorithm

1. $P =$ dimension of y

2. $n =$ size of y

3. IF (any of $\pi_k^{(0)}$, $\mu_{k\cdot}^{(0)}$, $Q_k^{(0)}$ or $D_k^{(0)}$ for $k = K+1,...,K+L$ is NA)
   Initialize $\pi_k^{(0)} = \frac{1}{L} * \lambda$ $Q_k^{(0)} = I_P$, $D_k^{(0)} = I_P$ and let $\mu_{k\cdot}^{(0)}$ be random
   ELSE Check if initialization parameters are of right dimension, $\sum_{k=1}^{K} \pi_k = 1$

4. For (k in 1:K) $\pi_k^{(0)} = (1-\lambda)\pi_k^{(0)}$

5. LikeLast = $-\infty$

6. LikeNew = the likelihood for the 0^{th} step computed based on y

7. Iterator $r = 0$

8. While(|LikeLast-LikeNew| > ν AND r <n_iter)

   (a) Covariance matrices
   For (k in K+1:K+L)

$$\hat{\Sigma}_k^{(r)} = \hat{Q}_k^{(r)} \hat{D}_k^{(r)} \left(\hat{Q}_k^{(r)}\right)'. \tag{29}$$

   (b) Posterior probability
   For (i in 1:n, l in 1:K+L)

$$\tau_{il}^{(r)} = \frac{\pi_l^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_l^{(r)}, \hat{\Sigma}_l^{(r)}\right)}{\sum_{k=1}^{K} \pi_k^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k^{(r)}, \hat{\Sigma}_k^{(r)}\right)} \tag{30}$$

   (c) Components proportions
   For (k in K+1:K+L)

$$\hat{\pi}_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(r)}. \tag{31}$$

   (d) $\hat{\lambda}^{(r+1)} = \sum_{k=K+1}^{K+L} \hat{\pi}_k^{(r+1)}$

   (e) Reweighing of the background proportions
   For (k in 1:K)

$$\hat{\pi}_k^{(r+1)} = \hat{\pi}_k^{(r)} * (1 - \lambda^{(r+1)}) / \left(\sum_{k=1}^{K} \hat{\pi}_k^{(r)}\right)$$

   (f) MLE esitmates
   For (k in K+1:K+L)

$$\tilde{\mu}_{kp}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} x_{ip}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{32}$$

$$\tilde{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(r)} * (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})'}{\sum_{i=1}^n \tau_{ik}^{(r)}} \tag{33}$$

(g) For (p in 1:P)

$$M_p^{(r)} = \max_{k=1,\ldots,K+L} \hat{\Sigma}_{k,pp}^{(r)}$$

(h) For (k in K+1:K+L) perform eigenvalue decomposition

$$\tilde{\Sigma}_k^{(r+1)} = \hat{Q}_k^{(r+1)} \tilde{D}_k^{(r+1)} \left(\hat{Q}_k^{(r+1)}\right)'. \tag{34}$$

(i) FOR (k in K+1:K+L,p in 1:P)

- IF $\left(\sum_{k=1}^K \left(\sum_{i=1}^n \tau_{ik}^{(r)} x_{ip}\right)^2\right)^{\frac{1}{2}} \leq \gamma_1 M_p^{(r)}$

$$\hat{\mu}_{kp}^{(r+1)} = 0$$

- ELSE

$$\hat{\mu}_{kp}^{(r+1)} = \tilde{\mu}_{kp}^{(r+1)} - \gamma_1 \frac{\hat{\mu}_{kp}^{(r)} \Sigma_{k,pp}^{(r)}}{\|\hat{\mu}_{\cdot p}^{(r)}\| \sum_{i=1}^n \tau_{ik}^{(r)}}$$

where the L_2 norm is computed based on background and signal mean parameters.

(j) FOR (k in K+1:K+L)

- $$\bar{D}_{k,pp}^{(r+1)} = \frac{-n\hat{\pi}_k^{(r+1)} + \sqrt{\left(n\hat{\pi}_k^{(r+1)}\right)^2 + 8\gamma_2 n \hat{\pi}_k^{(r+1)} \tilde{D}_{k,pp}^{(r)}}}{4\gamma_2}. \tag{35}$$

- To achieve numerical stability
  For(p in 1:P)
  IF $\bar{D}_{k,pp}^{(r+1)} < 0.0005$ then $\bar{D}_{k,pp}^{(r+1)} = 0.0005$.

(k) FOR (k in K+1:K+L,p in 1:P)

i. $\epsilon_k =$ mean ($P - p + 1$ smallest eigenvalues of $\bar{D}_k^{(r+1)}$).

ii. For $\alpha = 0.05$

$$Logical = \frac{\bar{D}_{k,pp}^{(r+1)}}{\epsilon_k} < 1 - \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2p}} \vee \frac{\bar{D}_{k,PP}^{(r+1)}}{\epsilon_k} > 1 + \sqrt{\frac{2}{n}} * z_{\frac{\alpha}{2p}}$$

where $z_{\frac{\alpha}{2p}}$ is the normal distribution quantile.

iii. If $(Logical = true)$ then the smallest $P - p + 1$ eigenvalues $\hat{D}_{k,pp}^{(r+1)} = ... = \hat{D}_{k,PP}^{(r+1)} = \epsilon_k$
BREAK the inner loop over 1:P.

iv. else $\hat{D}_{k,pp}^{(r+1)} = \bar{D}_{k,pp}^{(r+1)}$.

(l) Perform the "Background fit" on x with slightly changed formulas

- in 7e
$$M_p^{(r)} = \max_{k=1,...,K+L} \hat{\Sigma}_{k,pp}^{(r)}$$
- in 7g mind the signal components
$$\|\hat{\mu}_{\cdot p}^{(r)}\| = \sqrt{\sum_{k=1}^{K+L} \hat{\mu}_{kp}^{(r)}}$$
insted of used previously
$$\|\hat{\mu}_{\cdot p}^{(r)}\| = \sqrt{\sum_{k=1}^{K} \hat{\mu}_{kp}^{(r)}}$$

(m) LikeLast = LikeNew

(n) LikeNew = likelihood for the $(r+1)^{th}$ step.

9. Return all the parameters value for the last step and an error code if n_iter was reached (issues with convergence in n_iter steps)

References

[1] S. Caron, A General Search for New Phenomena (2004). Talk given at HERA-LHC Workshop (2004), available at http://www.nikhef.nl/~scaron/talks/heralhctalk.pdf.

[2] N. Craig, P. Draper, K. Kong, Y. Ng, D. Whiteson, The unexplored landscape of two-body resonances, arXiv 1610.09392 (2016).

[3] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, JHEP 07 (2014) 079.

[4] T. Sjostrand, S. Mrenna, P. Z. Skands, PYTHIA 6.4 Physics and Manual, JHEP 05 (2006) 026.

[5] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, P. Z. Skands, An Introduction to PYTHIA 8.2, Computer Physics Communications 191 (2015) 159–177.

[6] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, M. Selvaggi, DELPHES 3, A modular framework for fast simulation of a generic collider experiment, JHEP 02 (2014) 057.

[7] ATLAS Collaboration, Search for high-mass dilepton resonances in pp collisions at $\sqrt{s} = 8$ tev with the atlas detector, Physical Review D-Particles, Fields, Gravitation and Cosmology 90 (2014).

[8] CMS Collaboration, Search for lepton flavour violating decays of heavy resonances and quantum black holes to an eμ pair in proton-proton collisions at $\sqrt{s} = 8$ TeV, The European Physical Journal C76 (2016) 317.

[9] CMS Collaboration, Model Unspecific Search for New Physics in pp Collisions at $\sqrt{s} = 7$ TeV (2011).

[10] CMS Collaboration, MUSiC, a Model Unspecific Search for New Physics, in pp Collisions at $\sqrt{s} = 8$ TeV (2017).

[11] ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV (2014).

[12] ATLAS Collaboration, A general search for new phenomena with the ATLAS detector in pp collisions at sort(s)=7 TeV. (2012).

[13] ATLAS Collaboration, A model independent general search for new phenomena with the ATLAS detector at $\sqrt{s} = 13$ TeV (2017).

[14] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, Association for Computing Machinery computing surveys (CSUR) 41 (2009) 15.

[15] T. Vatanen, M. Kuusela, E. Malmi, T. Raiko, T. Aaltonen, Y. Nagai, Semi-supervised detection of collective anomalies with an application in high energy particle physics, The 2012 International Joint Conference on Neural Networks (2012) 1–8.

[16] M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen, Y. Nagai, Semi-supervised anomaly detection – towards model-independent searches of new physics, Journal of Physics: Conference Series 368 (2012) 012032.

[17] C. Fraley, A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal of the American statistical Association 97 (2002) 611–631.

[18] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the royal statistical society. Series B (methodological) (1977) 1–38.

[19] R. Alexandridis, S. Lin, M. Irwin, Class discovery and classification of tumor samples using mixture modeling of gene expression data - a unified approach, Bioinformatics 20 (2004) 2545–2552.

[20] J. D. Banfield, A. E. Raftery, Model-based gaussian and non-gaussian clustering, Biometrics (1993) 803–821.

[21] G. Celeux, G. Govaert, Gaussian parsimonious clustering models, Pattern recognition 28 (1995) 781–793.

[22] C. Bouveyron, S. Girard, C. Schmid, High-dimensional data clustering, Computational Statistics & Data Analysis 52 (2007) 502–519.

[23] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.

[24] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al., Least angle regression, The Annals of statistics 32 (2004) 407–499.

[25] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[26] W. Pan, X. Shen, Penalized model-based clustering with application to variable selection, Journal of Machine Learning Research 8 (2007) 1145–1164.

[27] B. Xie, Variable selection in penalized model-based clustering, University of Minnesota (2008).

[28] H. Zhou, W. Pan, X. Shen, Penalized model-based clustering with unconstrained covariance matrices, Electronic journal of statistics 3 (2009) 1473.

[29] B. Xie, W. Pan, X. Shen, Variable selection in penalized model-based clustering via regularization on grouped parameters, Biometrics 64 (2008) 921–930.

[30] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 49–67.

[31] W. Pan, X. Shen, A. Jiang, R. P. Hebbel, Semi-supervised learning via penalized mixture model with application to microarray sample classification, Bioinformatics 22 (2006) 2388–2395.

[32] P. Bühlmann, S. Van De Geer, Statistics for high-dimensional data: methods, theory and applications, Springer Science & Business Media (2011) 298–300.

[33] M. L. Eaton, Multivariate statistics: A vector space approach, Beachwood, Ohio, USA: Institute of Mathematical Statistics (2007).

[34] C. E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Libreria internazionale Seeber (1936).

[35] G. Schwarz, Estimating the dimensions of a model, Annals of Statistics (1978) 461–464.

[36] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, https://www.R-project.org/ (2017).

[37] L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, mclust 5: clustering, classification and density estimation using Gaussian finite mixture models, The R Journal 8 (2016) 205–233.

[38] L. Bergé, C. Bouveyron, S. Girard, HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data, Journal of Statistical Software 46 (2012) 1–29.

[39] J. M. Santos, M. Embrechts, On the use of the adjusted rand index as a metric for evaluating supervised classification (2009) 175–184.

[40] W. M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical association 66 (1971) 846–850.

[41] B. Fuks, Beyond the Minimal Supersymmetric Standard Model: from theory to phenomenology, International Journal of Modern Physics A27 (2012) 1230007.

[42] R. Barbier, et al., R-parity violating supersymmetry, Physical Review 420 (2005) 1–202.

[43] The minimal supersymmetric standard model with r-parity violation (2012).

[44] R. D. Ball, et al., Parton distributions with LHC data, Nuclear Physics B867 (2013) 244–289.

[45] C. Chen, New approach to identifying boosted hadronically-decaying particle using jet substructure in its center-of-mass frame, Physical Review D85 (2012) 034007.

[46] J. W. Tukey, Exploratory data analysis, Reading: Addison-Wesley (1977).

[47] Z. S. Abdallah, L. Du, G. I. Webb, Data preparation, Encyclopedia of Machine Learning and Data Mining (2016) 1–11.