UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXI

Advanced statistical methods for data analysis in particle physics

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Giovanna Menardi

Co-supervisores: Prof. Bruno Scarpa, Prof. Livio Finos

Dottorando: Grzegorz Kotkowski

Abstract

The thesis has been developed focusing on the use of multivariate statistical methods in the High Energy Physics framework. Stemming from the framework described by the current dominant physical theory, known as the Standard Model, the thesis has been developed by following two directions, associated with two different physical research questions.

The first route takes the steps from the need of improving the knowledge within the Standard Model. From a statistical point of view, such improvement refers to the aim of obtaining more accurate estimates of the parameters describing the Standard Model in order to gain a better knowledge of the probability distribution of the underlying physical process, known as the background. In practice, estimation of such probability distribution builds on the use of Monte Carlo simulated data, which, in turn, can be costly and imprecise. To prevent these problems, the physical community has developed a novel procedure to generate artificial background data from the experimental ones. Within the thesis, a formal validation of the physical procedure is performed by means of introducing a statistical permutation-based two-sample test for density equality. The test relies on kernel density estimation and is suitably adjusted to be applied to high dimensional data.

The second direction of research derives from the incompleteness of the Standard Model, known to be unable to fully describe the Universe and the interactions among its characterising forces. The goal of going beyond the Standard Model is reached through model-independent searches of new physics which aim at looking for new possible particles not predicted by the Standard Model. Such particles, referred to as a signal, are expected to behave as a deviation from the known background. From a statistical perspective, the problem is recasted to a peculiar classification one where only

partial information is available. Therefore a semi-supervised approach shall be adopted, either by strengthening or by relaxing assumptions underlying clustering or classification methods respectively. Within this context, the thesis follows two distinct approaches. The first approach consists of developing a parametric semi-supervised method which originates from the framework of model-based clustering. A dimensionality reduction technique is proposed by resorting to penalised methods to circumvent issues related to parameters estimation and the curse of dimensionality. The proposed variable selection approach is extended from the unsupervised to the semi-supervised context with attention to features exhibiting anomalous properties. The second approach followed with the aim of new physics searches consists of suitably adjusting and statistically validating an existing procedure, developed within the physical community. Some improvements to the algorithm are also proposed regarding, among others, cases of high dimensional and correlated data.

Sommario

Questa tesi si concentra sull'uso di metodi statistici multivariati in un contesto della fisica per le alte energie. Partendo dall'ipotesi dominante nella teoria fisica, conosciuto come Modello Standard, questa tesi si muove in due direzioni, associate a due diverse domande di ricerca provenienti dalla fisica.

Il primo contributo parte dalla necessità di comprendere meglio i dettagli del Modello Standard. Da un punto di vista statistico, il miglioramento della conoscenza del Modello Standard può essere tradotto nell'obiettivo di ottenere stime più accurate dei parametri che lo descrivono, al fine di avere una migliore conoscenza della distribuzione di probabilità dei processi fisici sottostanti, noti come *background*. Nella pratica tali stime partono da simulazioni Monte Carlo che a loro volta possono essere computazionalmente onerose e imprecise. Per ovviare a questo problema la comunità scientifica ha elaborato nuove procedure per generare il *background* dai dati sperimentali. All'interno della tesi si propone un metodo per validare in maniera formale queste procedure fisiche, basato su un test di permutazione a due campioni per l'uguaglianza in distribuzione. Il test proposto si basa sull'uso stime *kernel* della densità, ed è stato opportunamente aggiustato in modo da poter essere applicato a dati elevata dimensionalità.

Il secondo contributo parte dalla considerazione che il Modello Standard è incompleto, essendo incapace di descrivere l'universo che ci circonda e l'interazione tra le forze che lo caratterizzano. L'obiettivo di superare il Modello Standard è attuato ricercando nuove possibili particelle non predette dalla teoria. Queste particelle definite *segnale*, si assume si manifestino come deviazione rispetto al comportamento del *background*. Da un punto di vista statistico questa ricerca può essere interpretata come un problema di classificazione dove solo una parte dell'informazione è disponibile. L'approccio, che assume dunque caratteristiche semi-supervisionate, può essere affrontato o rilassando le ipotesi proprie dei metodi di classificazione, o rafforzando quelle dei metodi

di raggruppamento. In questo contesto, la tesi segue due approcci. Il primo consiste nello sviluppare un metodo parametrico basato su modelli di raggruppamento, in cui si propone una tecnica per la riduzione della dimensionalità basata su metodi penalizzati, in modo da prevenire problemi relativi alla stima dei parametri e alla maledizione della dimensionalità. Il metodo proposto per selezione delle variabili è esteso dal caso non supervisionato a quello semi supervisionato, con particolare attenzione per le variabili con caratteristiche anomale. Il secondo approccio, consiste nel tarare e validare da un punto di vista statistico, procedure già esistenti, e sviluppate in contesti fisici. Alcune migliorie sono state proposte, riguardando, tra le altre, casi ad alta dimensionalità e dati correlati.

To my lovely wife and children

Acknowledgements

Over the past three years, I have received invaluable help and support from many people regarding the realisation and completion of my PhD. Without their assistance, the finalisation of the thesis would not have been possible, or at most, it would have been miserable. In these short words, I would like to acknowledge those who impacted my work the most and to whom I have the greatest gratitude.

Firstly, I would like to express gratefulness to the thesis supervisor Giovanna Menardi for her continuous support and guidance during the research project. I am grateful for her accurate feedback, advice, patience and confidence in my abilities. I respect her hard-working attitude and devotion to all the actions she takes. By observing this mindset, my approach to the research work has been strongly impacted and furthermore, extends to several areas of my life. Subsequently, I would also like to express my gratitude to Livio Finos and Bruno Scarpa. It was an excellent opportunity to learn from them and to cooperate. Through this time they have been helpful to me and have contributed with plenty of ideas and devoted a lot of their time for my research studies. A great acknowledgement is also addressed to Tommaso Dorigo for his support, encouragement and joint work on research. As well I would like to thank the University of Padova, in particular, to the Department of Statistical Sciences and the associated staff, for the reach education offer, encouraging environment and plenty development opportunities.

I would also like to formulate priceless thanks to the referees of my thesis - professors Lara Lusa and Andrea Giammanco - for agreeing for being my thesis referees; for their insightful comments and questions that encouraged me to widen the research from various perspectives, hopefully resulting in an improvement of the thesis final version.

Finally, I wish to express my sincerest thanks to my family for their continuous support. Primarily, I gratitude to my fantastic wife Maja for her most profound love and her enthusiastic support. To her and our kids, for their accompany and understanding throughout the unusual student-life conditions. As well, I want to give my gratitude to my parents and grandparents who have seeded in me a curiosity about science and have supported me in all my pursuits.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement

AMVA4NewPhysics No 0675440. Without the project, my research would have been impossible, for which I am very grateful.

Contents

List of Figures

List of Tables

Introduction

Overview

Since the early Seventies, the Standard Model has represented the state of the art in Particle Physics. It describes the structure of the Universe - the elementary particles, and the inherent interactions - the fundamental forces. Despite its apparent empirical confirmations, it is evident that the Standard Model is not a complete theory, as it fails to explain several phenomena like, for instance, the gravity, the nature of dark matter, and the dark energy. For this reason, there are persistent attempts to either extend the Standard Model or to build an entirely new theory (Grossman and Rakshit, 2004). To this aim, physical experiments are conducted within large accelerators, as e.g., the LHC at CERN. The experiments involve propelling charged particles, making them collide and detecting the created products. Collected data are then used to validate and delve into physical theories or to find evidence of possible new physics, not predicted by the Standard Model.

In broad terms, two physical processes are of interest within the considered framework: the *background* process refers to the known physics described by the Standard Model. Although well rooted by definition, and with the due specifications which will be clarified in the rest of the thesis, the knowledge of such process is required to be deepened for specific purposes. The latter process, referred to as *signal*, represents, in turn, the unknown physics, which is required to complete our knowledge of the Universe. While the physical theory allows for conjecturing possible expressions for such process, its existence is itself brought into question and, whenever acknowledged, it is required to expect its extreme rarity.

In this framework, statistics plays a pivotal role, aiming at providing tools to analyse the data and thus answer the physical research questions. In this perspective, and consistently with the aforementioned problems, this thesis addresses the following goals:

1. Improving the accuracy in estimating the probability density function underlying the background process.

2. Providing the tools to identify a possible, unknown, signal and to discriminate it from the background process.

Main contributions of the thesis

Stemming from a characterization of the aforementioned research goals, in the following, the investigation path which has been pursued to attain them, as well as the main results of the thesis are summarised.

- Despite that the background process represents, by definition, the known physics, its probabilistic generating mechanism is not explicitly defined or numerically computable, and requires to be estimated. However, the background data needed to estimate the underlying probability density function are not always accurate or available. To circumvent the problem, Dall'Osso *et al.* (2017) have designed an algorithm aimed at generating background-like data. A statistical validation is then necessary to test whether the method performs according to its goals.

  The problem can been framed in a hypothesis testing framework, with the null hypothesis establishing the equality of two distributions. While this problem admits, in principle, a number of standard solutions, the available approaches are based on specific assumptions that do not hold in the considered application, where data at hand are multivariate and exhibit non-Gaussian properties, such as skewness and multimodality.

  Duong and Schauer (2012) have proposed a kernel density-based global two-sample test, which appears suitable for the application. However, nonparametric methods are particularly affected by the curse of dimensionality which prohibits their use for high dimensional data.

  As a first contribution of the thesis, it is proposed to perform the mentioned test multiple times in low-dimensional subspaces, and apply a proper combination function to the obtained results for verifying the previously stated hypothesis. Due to correlations between the multiple test results, statistical inference from their combination is not straightforward. For this reason, it is proposed to embed the test in a permutation framework, which allows for obtaining the empirical distribution of the combination function values under the null hypothesis. The obtained permutation-based two-sample test is validated concerning its first type error rate and its power. Finally, it is applied to the physical data to answer the primary question of interest.

- The main assumption underlying empirical searches of new physics is that any possible signal would behave as a deviation from the background process. From a statistical perspective, the problem can be then expressed in the framework of anomaly detection, where observations not consistent with the assumed background model are searched in the experimental data (Pimentel *et al.*, 2014). Unlike above, in this setting the background process is entirely known and a sample of virtually infinite size can be drawn from it. Two different sources of data are then available: a first sample generated from the background process - in the following referred to as labelled since the generating process of the observations is known, and a second, unlabelled experimental sample, whose generating mechanism is unknown, as it surely include observations from the background but might also include observations from the signal. The anomaly detection problem can be then faced according to a semi-supervised approach, due to the partial labelling of the available data.

  Among several alternatives, in the thesis two different approaches to the semi-supervised anomaly detection problem are followed:

  - The first approach stems from the idea of semi-supervising the signal detection by strengthening unsupervised (clustering) methods via the inclusion of the additional information available on the background. The proposed method originates from the family of model-based clustering approaches, and assume Gaussian mixtures densities to model the background and signal distributions. Due to the curse of dimensionality, the approach can be sub-optimal or even not feasible to be performed on high-dimensional data. Pan and Shen (2007)) and Xie *et al.* (2008) introduce penalised methods for variable selection in the context of model-based clustering, but rely on restrictive assumptions on the covariance matrices of the clusters. In the thesis, the penalised approach is extended to allow for a more flexible modelling without constraining the mixture component covariance matrices. Additionally, a variant of the Expectation-Maximization algorithm (Dempster *et al.*, 1977) is derived and implemented, to estimate the parameters of the mixture model via the numerical maximization of the penalized likelihood function. Subsequently, the idea of variable selection within model-based clustering is extended for anomaly detection purposes in the semi-supervised setting, starting from the works of Vatanen *et al.* (2012) and Kuusela *et al.* (2012).
  - The second approach for semi-supervised anomaly detection phrases the problem in terms of hypothesis testing. In summary, a signal is identified in the

experimental data if there is evidence that such data are not compatible with the background probability distribution. This natural idea has been developed by Vischia and Dorigo (2017) which make use of sampling and multiple hypothesis testing to study anomalous properties of data at hand. The proposed procedure depends on many parameters, but their influence on the method performance has been yet unclear. In the thesis, an optimal selection of such parameters is studied both theoretically and based on simulations. The performance of the procedure is validated and compared with competing methods given artificial data and within applications to real ones. Some improvements of the method are proposed concerning its performance for high dimensional data.

The rest of the thesis is organized as follows. In Chapter 1 an overview of the physical framework is provided. Chapter 2 introduces the problem of background density estimation, illustrates the physical procedure for generating background data, discusses and validates the proposed permutation test. Chapter 3 and 4 focus on the signal detection problem, and illustrate the clustering-based and, respectively, the hypothesis testing-based procedures.

Chapter 1

The physical framework

1.1 The Standard Model

Since the dawn of time, men have been trying to make sense to the surrounding word. Over the years, incredible progresses have been done in understanding the physics of the Universe, ranging from the microscopic scale of atom building quarks, to giant stars and quasars. One question of interest concerns understanding the structure of the Universe - *the elementary particles* - and the inherent interactions - *the forces*. Physical theories provide an effective mathematical formulation describing physical systems, subsequently validated and possibly confirmed based on physical experiments specifically designed. Once a theory is confirmed it becomes a starting point for further extensions which, in turn, allow for a more accurate description of the world.

Over time, particle discoveries, as for example, the ones of electron (1896), proton (1919) or neutron (1932), have allowed for a rough understanding of matter building blocs. With the increasing knowledge of the universe structure, an advance in understanding the physical forces has been made. Four fundamental particle interactions have been classified - namely the electromagnetic, weak interaction, strong interaction and the gravitational forces. Around 1970 a complex theoretical framework was established to comprehensively describe the elementary particles known at that time, and three of the four fundamental forces (gravity not included). The theory, referred to as the *Standard Model*, postulates the existence of matter constituents - the *fermions* (see Figure 1.1) - and the mediators of interactions - the *bosons* (see Figure 1.2). Fermions and bosons are characterised by the electrical charge, spin and a life-span, which establish their properties, possible interactions or bounding configurations.

At the time of the first Standard Model formulation, many of the model-assumed elementary particles were not empirically discovered yet. Only recently the last missing

	Leptons				Quarks			
	Particle		Q	mass/GeV	Particle		Q	mass/GeV
First	electron	(e^-)	−1	0.0005	down	(d)	−1/3	0.003
generation	neutrino	(ν_e)	0	$< 10^{-9}$	up	(u)	+2/3	0.005
Second	muon	(μ^-)	−1	0.106	strange	(s)	−1/3	0.1
generation	neutrino	(ν_μ)	0	$< 10^{-9}$	charm	(c)	+2/3	1.3
Third	tau	(τ^-)	−1	1.78	bottom	(b)	−1/3	4.5
generation	neutrino	(ν_τ)	0	$< 10^{-9}$	top	(t)	+2/3	174

FIGURE 1.1: The twelve postulated fundamental constituents of matter in the Standard Model - fermions (Thomson, 2013).

Force	Strength	Boson		Spin	Mass/GeV
Strong	1	Gluon	g	1	0
Electromagnetism	10^{-3}	Photon	γ	1	0
Weak	10^{-8}	W boson	$W^\pm$	1	80.4
		Z boson	Z	1	91.2
Gravity	10^{-37}	Graviton?	G	2	0

FIGURE 1.2: The four known forces of nature. The relative strengths are approximate indicative values for two fundamental particles at a distance of $1 fm = 10^{-15} m$ (Thomson, 2013).

elements of the Standard Model have been proved - the top quark (CDF Collaboration, 1995), the tau neutrino (DONUT Collaboration, 2001) and the Higgs boson (ATLAS Collaboration, 2012; CMS Collaboration, 2012). The modern discoveries deeply rooted the Standard Model (with some later adaptations) to be the current widely-acceptable state of the art of the physical theory.

Although the Standard Model is sometimes claimed to be a *uniform theory of everything*, the physical community is far to consent with such statement. The Standard Model is not a self-contained theory, as for example, it does not account for the gravitational force; it is not able to explain the mechanism of the dark matter creation, the neutrinos non-zero mass and many other phenomena (Fuks, 2012). For these reasons, there are persistent attempts to either complete the Standard Model or to build an entirely new theory (Grossman and Rakshit, 2004). In general, any further development of the current state of art needs to address the two following questions of interest:

1. Is it possible to improve our knowledge within the Standard Model framework?
2. Is it possible to go beyond the Standard Model, by completing it or defining an alternative theory?

With some due specifications, these are the general problems on which the thesis focuses. The aim is pursued via an investigation of suitable statistical methods aimed

at providing an answer to these questions.

1.2 The experimental settings

A natural way to answer the central research questions is to design and perform suitable experiments to find a required justification. Nowadays, such physical experiments and the experimental infrastructures to make them possible are incredibly complex, costly and require years of careful planning before any answer could be found. The complicated process of an experiment design is only sketched underneath, as our focus is instead put on statistical analyses of collected experimental data to find possible evidence for backing up any theoretical claims.

In general, experiments with elementary particle physics often require the use of very high energy to allow for an insight into processes that are rare or do not occur at low energy. For this aim, specific infrastructures are built - namely particle accelerators, that use an electromagnetic field to propel charged particles to nearly the light speed. There are many accelerators types with different topological structure (linear, circular) or mechanical design (electrostatic, electrodynamic) but have the common aim of increasing particle kinematic energy to a given high value. The currently most famous accelerators from their impact on discoveries are the *LHC* (CERN, Switzerland), *Tevatron* (Fermilab, the USA) and *KEKB* (KEK, Japan).

For the high energy physic purposes, recent accelerators are often built in a circular shape - the so called synchrotrons. Such design allows for a continuous acceleration of groups of charged particle - *beams* - to reach an exceptional energy of the propelled particles. The LHC is a record holder for reaching the energy equal to $6.5\,TeV$, in which particles are accelerated in an underground ring of circumference equal to 27 km. Usually, in synchrotrons, two beams are accelerated in opposite directions and cross in specific locations where the propelled particles collide.

Each particle collision, referred to as an event, produces complex individual interactions, i.e. state transitions, hadronisation, bosons and quarks production, particles decay, and others. Large particle detector systems are placed around the collision points, to enable reconstruction of the primary particles produced for each event. Such detectors are highly-advanced sensors which use a wide range of technologies to identify and measure properties of the produced particles. In general, detectors are cylindrically shaped barrels stretched along the beam path. They consist of specific zones for particle tracking (an inner region) and calorimeters for their energy measurements. See Figure

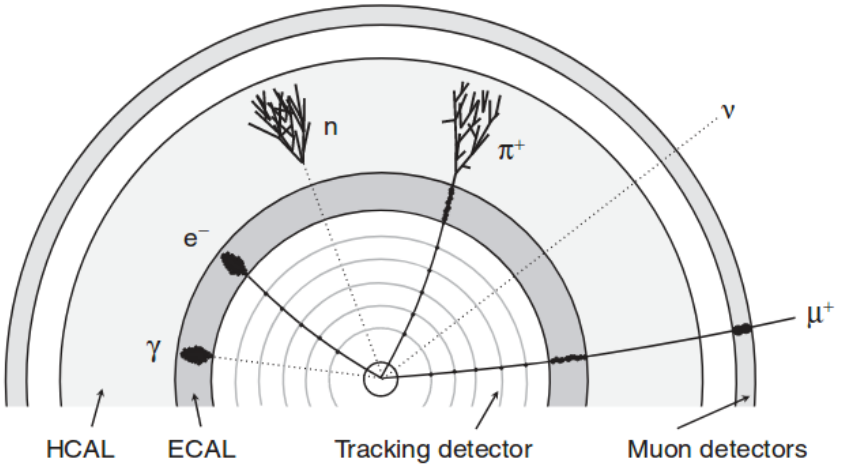


FIGURE 1.3: Typical layout of a particle detector equipped with a tracking system (here shown with cylindrical layers of a silicon detector), an electromagnetic calorimeter (ECAL), a hadron calorimeter (HCAL) and muon detectors. Usually, around the detector a solenoid is wrapped (not shown in Figure) to produce the magnetic field which bends the charged particles trajectories (Thomson, 2013).

1.3 for a schematic detector cross-section. Outside of the colourimeters, additional layers of muon detectors are often placed. Ideally, all the produced particles (apart from neutrinos) should be absorbed by the respective sensors and their existence evidenced. In the LHC at CERN, there are two main particle detectors: ATLAS, which has around 25 meters of diameter, and the more compacted CMS detector.

For each high-energy collision event, many particles can be produced. For each particle, its energy and a 3-dimensional momentum are measured, respectively by the calorimeters and the tracking elements of the particle detector. The four variables - the so called 4-vector - are necessary to identify each collision product distinctly. In the LHC a Cartesian coordinate system is defined as follows: the origin is located at the collision point, the x-axis points to the centre of the ring, the y-axis vertically upwards and the z-axis is tangent to the beam. Alternatively, an equivalent system of polar coordinates is used which is invariant under certain transformations. The azimuthal angle ϕ is measured in the xy plane from the x-axis and the radial coordinate in the plane is denoted by r. The polar angle θ is defined in the rz plane but preferably it is expressed in terms of the pseudorapidity $\eta = \ln\left(\tan\left(\frac{\theta}{2}\right)\right)$. The transverse momentum - p_T - is computed as the momentum component perpendicular to the beam direction. Similarly, the transverse energy is defined as $E_T = E\sin\theta$ where E is the total energy.

Due to a very high frequency of collisions, the amount of the possibly produced data is so large that only a small percentage can be filed. Specifically designed hardware and software solutions, the so called triggers, allow for filtering out only observations

of potential interest, i.e. the part of the data to be discarded is redundant as it is associated with already well-known processes (CMS Collaboration, 2016b). Later, for events admitted by the triggers, a particle identification needs to be accomplished by specific software solutions based on the recorded tracks, energy deposits and the known physical theory (CMS Collaboration, 2009). Unfortunately, the identification is not straightforward as some particles quickly decay before they reach calorimeters. An example is an energetic quark which decays by radiating gluons along its way so that its initial energy cannot be directly measured. However, the quark generates bunch of gluons which travel in approximately the same direction so that a cone is formed - a phenomenon referred to as the *jet* formation. Jet measurements allow for later reconstruction of energy and momenta for the decayed quark. The summary of all the identified particles of a given event (for example one electron and two jets identified for an event) are referred as the event final state. After all these steps, the preprocessed experimental data are finally prepared for the analysis purpose. In conclusion, the data consist of automatically selected events where each of them is composed of a list of identified particles with their respective measurements - the 4-vector.

1.3 Motivation

1.3.1 Improvement of the knowledge within the Standard Model framework

The Standard Model framework describes possible outputs and interactions of physical phenomena in probabilistic terms but, in general, a probability distribution of collision kinematic or angular variables is unknown. In principle, such distribution could be computed by solving multivariate integrals of partition functions and nonlinear Lagrangians. However, for computational and numerical reasons this is not feasible to be performed. Hence, the probability distribution of the physical process described by the Standard Model, in the following referred to as the background, is required to be estimated. In fact, the problem is not trivial, due to the lack of reliable background data. Experimental data are not directly usable, as they are realizations of some, more comprehensive, possibly unknown process, whose the background represents only a (dominant) fraction.

Conversely, background estimation is often based on Monte Carlo simulations, performed under the assumption that the Standard Model is correct. Unlike statistical simulations, aimed at generating data given their distribution, physical Monte Carlo

simulations are based on a different concept, i.e. a probability density function of variables is not necessary to produce a related sample. Such simulated data consist of realisations of possible collisions which are produced by a complex system of adequate generation steps. Without going into the physical details, based on theoretical parton distribution functions, possible collision effects are simulated in an appropriate proportion and with respect to the physical rules, i.e. via the creation of elementary particles, coupling, their interactions and others. The simulations also cover other processes that the produced particles undergo immediately after the collision, as for example their deceleration, scatterings or jet creation. For such simulated events, a detector response is computed based on its measurement efficiency. Later, a specific trigger is applied, to reflect circumstances in which the experimental data are collected. Finally, a particle identification is performed leading to the collision reconstruction. To facilitate the complex computation, many specific software packages have been implemented for this aim, as for example "*MadGraph*" (Alwall *et al.*, 2014), "*Pythia*" (Sjőstrand *et al.*, 2006, 2015), or "*Delphes*" (de Favereau *et al.*, 2014).

Hence, simulating data from the background process is possible, and frequently performed, so that for many final states of particle collisions the generated data are tolerably accurate. Nevertheless, given the degree of complication, some simplifications need to be applied for facilitating the simulation procedure. For example, to avoid computations of multivariate integrals describing complex particle states, approximation methods are used (Frixione and Webber, 2002). Also, not every physical phenomenon is predictable, for instance, whenever the strong force plays a relevant role. In all these circumstances, simulations get imprecise or biased. Additionally, for some rare processes, it can be impossible to produce large data quantity as they are computationally too expensive to be generated.

To respond to the scarcity or the inadequacy deficiencies of Monte Carlo data for complex final states, Dall'Osso *et al.* (2017) have designed a novel approach that addresses the issue of generating background-like data, referred to as the *Hemisphere Mixing*. They propose to use the experimental data themselves which might include non-background observations, and apply a specific permutation scheme, referred to as a mixing procedure. The mixing is driven jointly by the knowledge of the physical process and information from the experimental data. It aims at transforming all the input data into observations distributed according to the dominant background process - to be used for estimating the background density.

In experimental particle physics, the mixing approach is not novel, but uncommon. The method has turned out to be successful for some specific applications, as for example

electron-positron collisions (CMS Collaboration, 2010, 2011b). However, for complex final states of the proton-proton collisions occurring at the LHC, the idea has never been used. In specific, Dall'Osso *et al.* (2017) have adjusted the mixing definition to make it suitable for the multijet collisions.

Although there is a physical and logical reasoning behind the Hemisphere Mixing approach, it needs to be formally verified if the method performs appropriately to its goals. These are:

1. From input data generated dominantly by the background, the Hemisphere Mixing produces data entirely distributed according to the background.

2. Distributions of input data and the related hemisphere mixed output data are equal, if and only if the input data are the background.

The aforementioned questions are addressed in Chapter 2, via an introduction of a suitable statistical test.

1.3.2 Going beyond the Standard Model

Another fundamental research direction within the physical community follows the need of developing the current theory. This can be achieved either by extending the Standard Model or by constructing an entirely new physical framework (Fuks, 2012). Research is often performed based on a data-driven evidence of new physics which can appear, for instance, as a sign of a new particle unpredicted by the Standard Model.

There have been several efforts in new physics searches within the community. Such searches are often driven in a model-dependent guise where supervised classifiers are trained to find particular phenomena expected to be seen under hypothetical theory extensions (ATLAS Collaboration, 2014b; CMS Collaboration, 2016a). As a result, only a narrow subspace of possible alternative extensions is tested. Without being constrained to any physical hypothesis, a more general approach - called model-independent - has also been applied in this context (CMS Collaboration, 2011a, 2017; ATLAS Collaboration, 2014a, 2017; Popov, 2011). Such approach allows the data to speak for themselves and searches within a broader range of alternative signal processes. It aims at new signal detection rather than confirmation of any physical theory.

The main assumption underlying model-independent searches is that new possible physics - referred to as a *signal* process - would show an anomalous behaviour with respect to the background. In the considered setting, Monte Carlo simulations are usually trusted to provide accurate data from the background and feasibly produced

with an arbitrary size. Hence, the problem of detecting a possible signal can be addressed according to a semi-supervised approach, by comparing the the Monte Carlo data generated from the background, with the experimental ones, which might include observations from the signal.

This problem is addressed in Chapters 3 and 4, by considering two alternative logics, based on the semi-supervision of clustering methods and the employment of hypothesis testing respectively.

Chapter 2

Validation of a physical algorithm to improve background estimation

2.1 Motivation and goals

The background process is defined by the Standard Model established on the known physical theory, however, its underlying probability density function is generally unknown and requires a precise estimation. Such goal has been usually attained relying on Monte Carlo simulated data. Nevertheless, simulations are suitable only for some scenarios for which the artificially generated data can be trusted and are available in large quantity. Otherwise, the density estimate is unobtainable or inaccurate. The issue of the Monte Carlo data imperfectness is present for instance in multi-jet final states analysis, which is especially useful for a more in-depth understanding of the Standard Model. In specific, non-resonant pair production of Higgs bosons decaying into a $b\bar{b}b\bar{b}$-quark final state provides an excellent measure to determine the self-coupling λ - a crucial Standard Model parameter. However, as the multi-jet final states have a high potential for a more exhaustive understanding of the Standard Model, it calls for investigating of novel methods that can replace the insufficient Monte Carlo simulations and improve the required background estimation.

An ideal solution to overpass the aforementioned issue would be to generate background data completely (or at worst approximately) independent on the underlying physical theory. Within this logic, the *Hemisphere Mixing* algorithm has been recently proposed by Dall'Osso *et al.* (2017). It takes as an input the experimental collision data $\mathcal{Y}$, and applies on them a specific permutation transformation driven by the knowledge of the physical context so that new synthetic data $\mathcal{Z}$ are produced. Under the assumption that the input data are generated entirely (or at least dominantly) by the background

process, it is aimed that the generated data $\mathcal{Z}$ have the background distribution. Provided that the method works according to its goal, it could be successively used for the production of background samples from the experimental data $\mathcal{Y}$ without resourcing to either Monte Carlo simulations or being heavily based on the physical theory.

From the statistical point of view, an extended analysis needs to be performed to verify if the Hemisphere Mixing algorithm produces data with adequate properties, i.e. distributed according to the background density. Ideally, a two-sample statistical comparison test should be applied to the background data and the hemisphere mixed output data to verify if the algorithm application can retain the data background properties and smear out the signal evidence if present. A permutation-based hypothesis test is introduced for this aim, its type-1 error and power are evaluated. The test is applied to physical data in several scenarios so that the algorithm performances for specific conditions are verified.

2.2 Description of the Hemisphere Mixing algorithm

The Hemisphere Mixing algorithm is based on the so-called observation mixing approach. In a shorthand, each input event is adequately replaced with components originating from other events. For each event, the algorithm groups collision jets into two disjoint sets and reconstitutes them using different sets with similar properties. The selected sets are jointed so that a new event is composed, and if applied to all the events, the algorithm output is produced. The mixing idea *per se* is not utterly new in physical applications (CMS Collaboration, 2010, 2011b) but it had required in-depth adjustments for the peculiar application setting, and to our knowledge, it has never been applied to the multi-jet studies.

Let us consider proton-proton collisions resulting in the production of multiple jets originating from collision points. As it has been explained in Section 1.2, each particle and jet can be distinctly defined by specifying their measured 4-vectors. However, to compare events, it is required to use more general event summary statistics as collisions can differ in the number of produced jets, and, in principle, the jets are not ordered. In practice, various event-based statistics are computed from the corresponding 4-vectors which give rise to data variables. The variables are chosen so that they have a meaningful physical interpretation, they are invariant under certain transformations or powerful for the background and a possible signal discrimination.

The first step of the algorithm is to bi-partition the event jets so that the two resulting sets have possibly null or negligibly small between-group physical interactions. This is

performed within the data framework defined in Section 1.2 using a "*thrust*" axis $\boldsymbol{T}$. The axis $\boldsymbol{T}$ is determined in the Euclidean coordinate system separately for each event based on its jet momenta. It passes through the collision point and its direction is specified in the way that event jet momenta projections along it are maximised, i.e.

$$argmax \left(\sum_{h=1}^{Nj} ||\boldsymbol{p}_h|| \; |\cos \Delta \left(\boldsymbol{p}_h, \boldsymbol{T} \right)| \right), \tag{2.1}$$

where Nj is a number of the event jets, $\Delta \left(\boldsymbol{p}_h, \boldsymbol{T} \right)$ is the angle between the h^{th} jet momentum vector $\boldsymbol{p}$ and the thrust axis $\boldsymbol{T}$.

A perpendicular plane to the thrust axis $\boldsymbol{T}$, which passes through the collision point, divides the event topological space into two sub-spaces. Each subspace and jets enclosed in it are referred to as the *hemisphere*. A graphical visualisation of a collision, its resulting jets, the respective thrust vector and the hemisphere division is presented on the left-hand side of Figure 2.2.

A fundamental assumption of the Hemisphere Mixing algorithm is that interactions between event hemispheres are negligible. The independence assumption should be explained based on the known physical principles at least to the so-called "*first order*" effects. In fact, for the considered proton-proton collisions at the LHC, the produced hemispheres are dependent as, in general, the event centre of mass is not at rest. However, if the phase space is reduced to the transverse plane in which the thrust axis is bounded to be determined, the independence assumption holds, and the hemispheres are only related by the QCD radiations, pile-up effects or multiple parton scatterings (the detailed explanation is given in AMVA4NewPhysics ITN, 2017). Thus, without loss of generality, we consider only a 2-dimensional space of the transverse plane in which the thrust axis is determined.

The experimental data $\mathcal{Y}$ of size m are supplied to be the Hemisphere Mixing algorithm input. For $l = 1, ..., m$, the algorithm takes subsequent observations for which the respective thrust axis are determined and two corresponding hemispheres are obtained. Note that the hemispheres are ordering invariant, and, without loss of generality, let us denote them as h_{2l-1} and h_{2l}. After all the iterations, a total of $2m$ hemispheres are collected to compose the so-called *hemisphere library*. In general, using the independence assumption of the hemispheres, one could pick at random two hemispheres from the library, merge them up and obtain a possible valid event of the multi-jet proton-proton collision. The following variables describe the hemispheres: the number of jets they include Nj, the number of b-tagged jets Nt (a specific type of jets), the sum of projected jets transverse momenta along the thrust axis T, the combined mass of jets M, the sum

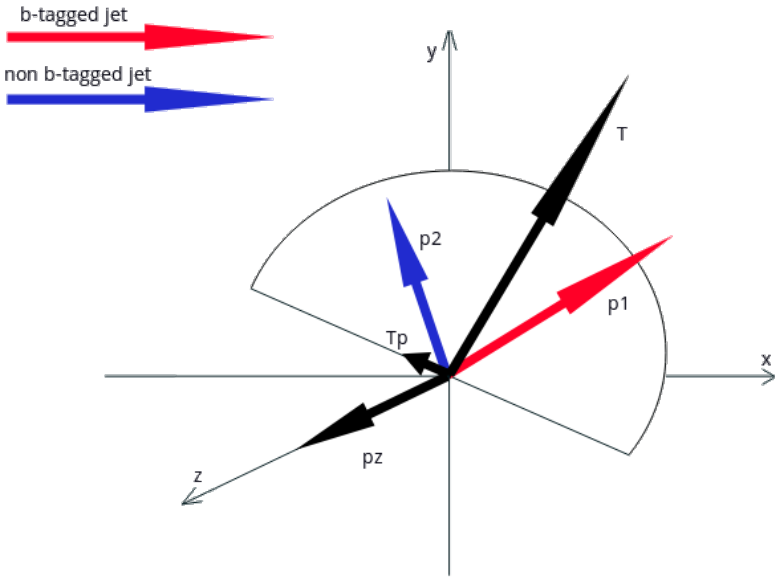


FIGURE 2.1: Graphical representation of an example hemisphere containing two jest ($Nj = 2$) with respective masses m_1 and m_2 and transverse momenta $p1$ and $p2$. One jet is b-tagged $Nt = 1$; the combined mass $M = m_1 + m_2$; T and Tp are chosen according to Equation 2.1.

of momenta projections perpendicular to the thrust Tp, and the sum of jets momenta components along the z-axis Pz. An example graphical representation of a hemisphere is presented in Figure 2.1. From one perspective each hemisphere can be seen as a point in a 6-dimensional space with the associated variables Nj, Nt, T, M, Tp and Pz.

After construction of the hemisphere library, the mixing procedure can be performed. For $j = 1, ..., 2m$, an iterative search within the Hemisphere library is performed to find the most similar hemisphere to the current one h_j. The similarity between the j^{th} and k^{th} hemispheres, for $k \in \{1, ..., 2m\} \setminus \{j\}$, is defined by a distance measure $D(l, k)$ expressed in the hemisphere feature space as

$$\begin{aligned} D(j,k)^2 \;=\; & \frac{(T(h_j)-T(h_k))^2}{Var(T)} \;+\; \frac{(M(h_j)-M(h_k))^2}{Var(M)} \\ & + \frac{(Tp(h_j)-Tp(h_k))^2}{Var(Tp)} \;+\; \frac{(|Pz(h_j)|-|Pz(h_k)|)^2}{Var(Pz)}, \end{aligned}$$

namely the Euclidean distance scaled by the variable variances. Additionally, if the hemispheres h_j and h_k differ for any values Nj or Nt the distance is set to $+\infty$. Finally, a hemisphere h_k with the smallest distance $D(j, k)$ to h_j is selected as the most similar. Let us denoted such hemisphere h_k by h_j^{lib} as the closest to h_j from all the ones in the library. Such search can be performed using a kind of a multi-dimensional nearest-neighbor approach (Bentley, 1975). New events – the algorithm output – are constructed

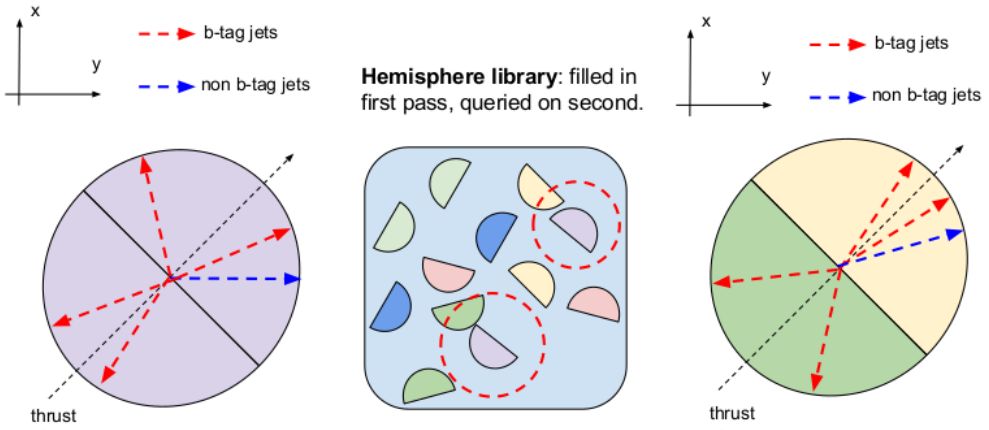


FIGURE 2.2: Graphical visualization of an original collision event (the left-hand side) and a corresponding created artificial event (on the right-hand side) from two closest hemispheres selected from the hemisphere library (central diagram). Figure originates form AMVA4NewPhysics ITN (2017).

Algorithm 1 Pseudo-code of the Hemisphere Mixing

Input: experimental data $\mathcal{Y}$
Parameters: distance measure function $D(j,k)$
1: $m \leftarrow$ size of the data $\mathcal{Y}$
2: allocate HemLibrary - an empty set
3: **for** $l = 1, ..., m$ **do**
4: determine a thrust axis for the observation $\boldsymbol{y}_l$
5: bi-partition $\boldsymbol{y}_l$ and produce hemispheres h_{2l-1} and h_{2l}
6: store h_{2l-1} and h_{2l} in HemLibrary
7: **end for**
8: compute a distance matrix based on the function $D(j,k)$ between all the hemispheres from HemLibrary
9: **for** $j = 1, ..., 2m$ **do**
10: $h_j^{lib} \leftarrow$ the closest hemisphere from h_j within the HemLibrary (excluding h_j)
11: **end for**
12: **for** $l = 1, ..., m$ **do**
13: $\boldsymbol{z}_l \leftarrow$ merged hemispheres h_{2l-1}^{lib} and h_{2l}^{lib}
14: rotate z_l according the the l^{th} thrust axis to closely correspond to $\boldsymbol{y}_l$
15: **end for**
16: **return** Data $\mathcal{Z}$ consisting of events $\boldsymbol{z}_l$ for $l = 1, ..., m$.

from the selected hemispheres, specifically, for $l = 1, .., m$, the two found hemispheres h_{2l-1}^{lib} and h_{2l}^{lib} corresponding to the l^{th} observation are merged up. Each constituted new artificial event is also appropriately rotated so that it matches the thrust axis and closely corresponds to the original one. Figure 2.2 gives a graphical overview of the presented idea and for a exhaustive explanation of the Hemisphere Mixing see Algorithm 1.

The outlined Hemisphere Mixing algorithm produces entirely new data $\mathcal{Z}$, referred to as the *hemisphere mixed data*. The mixing is designed to possibly preserve the dominant properties of the input data, for example, the marginal distribution of their kinematic variables. Hence, it is expected that the hemisphere mixed data produced from the background sample keep their inherent background distribution.

On the other hand, if the experimental input data include observations from the signal process, then the mixing procedure is expected to smear out the signal features, i.e. purify the mixture data to produce fully background-like distributed observations. It is presumed that due to the small number of signal observations, the hemisphere library is going to be poorly represented by the signal-originated hemispheres; consequently, it is likely to model the signal observations using the background originating hemispheres. Such modelling is expected to yield the dominant process properties. However, if the hemisphere similarities are determined by variables greatly discriminating the background and signal processes, then the signal events are likely to be reproduced using the signal originating hemispheres, and the expected smearing does not occur. In any case, whether the signal is present, some background observations can be wrongly modelled using at least partially the signal originating hemispheres, which can severely degrade the algorithm performance.

2.3 Statistical question of interest

2.3.1 Description of the problem

Let us introduce a notation for the datasets at hand. Denote the experimental data $\mathcal{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_m)'$, where $\boldsymbol{y}_l = (y_{l1}, ..., y_{lp}, ..., y_{lP})'$, $l = 1, ..., m$, are supposed to be i.i.d. realizations from an unknown probability density function $f_{BS} : \mathbb{R}^P \to \mathbb{R}$. Consider also a background density $f_B : \mathbb{R}^P \to \mathbb{R}$ which refers to the known processes predicted by the Standard Model. If the process generating the experimental data $\mathcal{Y}$ does not contain any signal component then naturally the distributions f_B and f_{BS} are equal. The hemisphere mixed data $\mathcal{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_m)'$, where $\boldsymbol{z}_l = (z_{l1}, ..., z_{lp}, ..., z_{lP})'$, $l = 1, ..., m$, are realizations from an unknown probability density function $f_{Out} : \mathbb{R}^P \to \mathbb{R}$. Note that all the mentioned densities (f_{BS}, f_B and f_{Out}) are in practice unknown. For the general case for which the algorithm is designed, the background data are not available due to the described issue of the Monte Carlo simulations. However, for the algorithm verification purpose, such collision scenario is chosen that the background and signal data are feasible to be produced in a large quantity. In specific, two datasets are generated: the background data $\mathcal{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$, where $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip}, ..., x_{iP})'$,

$i = 1, ..., n$ i.i.d. realizations from the background density f_B and the respective signal data from the density f_S which are used to generate the experimental data $\mathcal{Y}$.

In order to verify the performance of the Hemisphere Mixing algorithm, a formal statistical test has to be applied, to provide evidence that the density f_{Out} of the hemisphere mixed data $\mathcal{Z}$ is equivalent to the background one f_B, whether the input data $\mathcal{Y}$ include signal or not. The null hypothesis is

$$H_0 : \quad f_B(\cdot) = f_{Out}(\cdot)$$

against the alternative

$$H_1 : \quad f_B(\cdot) \neq f_{Out}(\cdot).$$

The issue of testing two samples for a common distribution is quite common for statistical application. The literature well describes many potential solutions (Tinsley and Brown, 2000). The most well-known is the Kolmogorov-Smirnov test (Sheskin, 2003) whose test statistic is computed based on a distance between empirical cumulative distribution functions of the two compared samples. The test can be used only for unidimensional data, but it has several other attractive features; among them it is the robustness to outliers, as the statistic is just sensitive to the bulk of density function. On the other hand, the test usually has small power in comparison to others (Razali *et al.*, 2011). Multivariate extensions of the Kolmogorov-Smirnov test have been proposed. However, they are computationally complex and do not scale well with the data dimensionality (Friedman and Rafsky, 1979; Justel *et al.*, 1997).

A more powerful substitute is the Wilcoxon rank sum test (Sheskin, 2003). This is a common nonparametric univariate two-sample test, for which the alternative hypothesis is that the two distributions differ by some location shift $\mu \neq 0$ (for the two-sided case). For the considered data this test is not suitable as it is not multivariate and tests a different hypothesis (the same location, in general, does not mean the equality of distributions).

Next, the Multivariate Analysis of Variance (MANOVA) (Sheskin, 2003) seems to be a better alternative as it is oriented at multidimensional cases. However, the test is designed to spot the difference in means, and therefore it also does not satisfy the meant hypothesis. Additionally, the assumption for the test is that the variables have Gaussian marginal distributions which is not the case for the data at hand (Figure 2.5). However, for a large number of observations the distribution of the sample mean is approximately normal (as it follows from the Central Limit Theorem) and for this reason, its use to some extent can be judged (Khan and Rayner, 2003).

As described above, these standard statistical tests are not proper for our purpose. For this reason, we have to identify a more sophisticated method, that is multivariate and designed for the described hypothesis. Duong and Schauer (2012) have proposed a kernel density-based global two-sample comparison test – for short the KDE test. The test makes no assumptions on the data distributions, it is multivariate and tests the required hypothesis. The used test statistic is the integrated square error

$$Z = \int \left[f_B(\mathbf{x}) - f_{Out}(\mathbf{x}) \right]^2 \mathrm{d}\mathbf{x},$$

where for f_B and f_{Out} kernel density estimates are plugged-in

$$\hat{f}_B(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{x}_i) \tag{2.2}$$

and

$$\hat{f}_{Out}(\boldsymbol{x}) = \frac{1}{m} \sum_{l=1}^{m} K_H(\boldsymbol{x} - \boldsymbol{z}_l),$$

K_H is a multivariate kernel with a bandwidth matrix H and the integration is taken over an appropriate Euclidean space. Duong and Schauer (2012) prove that the considered Z statistic has asymptotically Gaussian distribution. Such property has a great computational advantage in comparison to other multivariate tests which often resort to bootstrap procedures to compute its critical values (Aslan and Zech, 2005). However, the relevant drawback of the KDE test is that the kernel density estimation is highly affected by the curse of dimensionality (Azzalini and Scarpa, 2012; Scott, 2015) and by the need of an optimal selection of the bandwidth matrix H (Wand and Jones, 1995). Hence, in principle, the test is applicable, but not recommended for samples with higher dimensionality than 6 (Chacón and Duong, 2010).

2.3.2 Permutation-based statistical test

Within the problem under consideration, the initial number of variables one may observe is much higher than any dimension which could guarantee accurate nonparametric density estimation (typical collision data have about 20 variables). Therefore, the idea is to perform multiple tests on small subsets of data variables and infer from a combination of the test results. Let $\mathbb{T}$ be a set of all the data variables. We take at random S sets of variables from $\mathbb{T}$ so that each set $\mathbb{T}_s$, for $s = 1, ..., S$, contains precisely $U < P$ distinct variables Subsequently, the statistical tests are performed on data with variables given by $\mathbb{T}_s$. In this manner, a vector of S p-values is obtained. Consequently,

a solution for combining multiple test results is required to infer the initially stated hypothesis.

Inference methods for multiple test results have been well described in the statistical literature (Bibby *et al.*, 1979). The so-called *combination functions* – for short combinants – are proposed to reasonably put together the test p-values. A combinant is designed so that its value distribution is known, provided that particular assumptions are met (frequently assumed independence of the corresponding test statistics). Based on the theoretical distribution under the null hypothesis and the obtained combination function value, a single combined p-value is computed, which allows us to decide whether the null hypothesis should be rejected. However, selection of the combination function is relevant for the further inference (Pesarin and Salmaso, 2010, p. 128-134). The most frequently used is the Fisher combinant based on the test statistic

$$p^F = -\sum_{s=1}^{S} \log(p_s)$$

which has the χ^2_{2S} distribution if partial test statistics are independent. The other well-studied combinant is the Liptak combination function based on the statistic

$$p^L = \sum_{s=1}^{S} G^{-1}(p_s)$$

where G is the cumulative distribution function of the partial test statistic. The third popular combination function is the one of Tippett given by

$$p^T = \max_{s=1,\ldots,S}(1 - p_s)$$

or equivalently formulated as $p^M = -\min_{s=1,2,\ldots,S} p_s$ and for this reason it is often referred to as the min-p. For statistical tests with statistics increasing with observed evidence against the null (case of the KDE test) the Fisher and min-p combination functions are recommended (Heard and Rubin-Delanchy, 2018). The Liptak combinant requires to know a test statistic distribution, hence for our case, the approach is not suitable as the distribution G is only known asymptotically. From the aforementioned reasons, in this report, we consider the Fisher and min-p combination functions.

One important point to make is that the distributions for the combinants are only known if the p-values obtained in the multiple tests are independent. Unfortunately, for the studied case the tests are not independent as the subsets $\mathbb{T}_s$ have non-null intersections; besides, the mutual dependence among the variables might cause the sets $\mathbb{T}_s$

TABLE 2.1: Test statistics for performed tests on the $B+1$ permuted datasets regarding the S subsets of variables.

	Subsets of variables				
	$\mathbb{T}_1$	$\dots$	$\mathbb{T}_s$	$\dots$	$\mathbb{T}_S$
1^{st} permuted datasets	$Z_{1,1}$	$\dots$	$Z_{1,s}$	$\dots$	$Z_{1,S}$
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
b^{th} permuted datasets	$Z_{b,1}$	$\dots$	$Z_{b,s}$	$\dots$	$Z_{b,S}$
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
B^{th} permuted datasets	$Z_{B,1}$	$\dots$	$Z_{B,s}$	$\dots$	$Z_{B,S}$
Original datasets	$Z_{(B+1),1}$	$\dots$	$Z_{(B+1),s}$	$\dots$	$Z_{(B+1),S}$

TABLE 2.2: Overview of p-values computation given the corresponding test statistic values from Table 2.1.

	Subsets of variables						
	$\mathbb{T}_1$	$\dots$	$\mathbb{T}_s$	$\dots$	$\mathbb{T}_S$		Combinant
1^{st} permuted datasets	$p_{1,1}$	$\dots$	$p_{1,s}$	$\dots$	$p_{1,S}$	$\rightarrow$	p_1^C
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
b^{th} permuted datasets	$p_{b,1}$	$\dots$	$p_{b,s}$	$\dots$	$p_{b,S}$	$\rightarrow$	p_b^C
	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
B^{th} permuted datasets	$p_{B,1}$	$\dots$	$p_{B,s}$	$\dots$	$p_{B,S}$	$\rightarrow$	p_B^C
Original datasets	$p_{(B+1),1}$	$\dots$	$p_{(B+1),s}$	$\dots$	$p_{(B+1),S}$	$\rightarrow$	p_{B+1}^C

to be dependent. Hence, distributions of the two combinants under the null hypothesis are unknown. One way to overcome this problem and turn out with the distributions is to resort to a permutation framework (Pesarin and Salmaso, 2010). Given the two datasets $\mathcal{X}$ and $\mathcal{Z}$, new data are obtained by randomly swapping observations between the original sets. This guarantees that the distribution of the permuted samples are identical, i.e. that we are under the null hypothesis H_0. Based on the permuted data, the S tests are applied with respect to the sampled variable sets. The procedure is performed B times, and the obtained results can be collected in a table such as Table 2.1.

The results of multiple tests are specifically combined. Firstly for each test statistic value Z_{bs} the p-value is computed by columns of Table 2.1 as

$$p_{b,s} = \frac{\sum_{k=1}^{B+1} \mathbb{1}\{Z_{b,s} \leq Z_{k,s}\}}{B+1},$$

where $\mathbb{1}\{\cdot\}$ is the identity function. In this way the analogous table of p-values is constructed (Table 2.2). Afterwards, the chosen combinants are computed by rows. Note that we obtain $B+1$ combined p-values (denoted as p_b^C).

Given $B+1$ combined p-values p_b^C, we can derive the empirical distribution of the combinant values under the null hypothesis. Subsequently, the distribution is used to obtain the final p-value for the considered permutation framework for the original datasets. It is given as a rank percentile of the original value p_{B+1}^C across all the obtained. In other words, the final permutation-based p-value of a combinant is expressed as

$$p^C = \frac{\sum_{k=1}^{B+1} \mathbb{1}\{p_{B+1}^C \geq p_k^C\}}{B+1}$$

where p_b^C are computed based either on the Fisher p^F or min-p p^M combination function.

2.4 Performance of the statistical test

2.4.1 Simulation settings

In order to evaluate the proposed KDE permutation test, its I-type error rate and power have been evaluated. For this specific purpose, physical Monte Carlo data have been generated, according to the scheme illustrated in Section 1.3.1. The data correspond to proton-proton collisions in the LHC with multijet final states. The signal conforms with the Higgs bosons pair decay into 4 b-quarks ($hh \to b\bar{b}\,b\bar{b}$ channel) and the background represent Standard Model QCD events resulting in a production of at least two jets. A detailed description of the simulation process, driven by physical arguments, is beyond the scope of the thesis, and interested readers may refer to AMVA4NewPhysics ITN (2017). The generated background data have the size equal to 172150 and the signal ones to 1538. The small size of the signal set is due to Monte Carlo simulation issues of producing events from the rare process. Given the data, generated from signal and background processes, a surrogate of the experimental data can be obtained by suitable sampling from the two datasets with adequate proportions. Since a 4-vector distinctly describes a jet, each event is fully expressed using $4Nj$ measurements. The number of jets Nj might vary between the events and can be even higher than 7. In the following application, we consider 20 variables which contain adequate knowledge for the process of merit because jets with the smallest pT are of little relevance. In this way, each event is described by the same features, independetely from the possibly varying number of jets. The chosen data variables are expressed in Table 2.3.

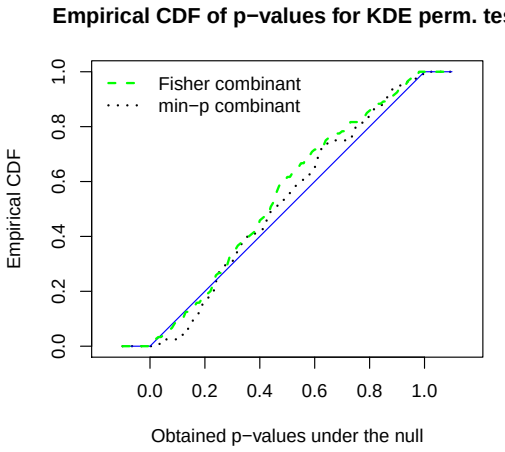
Verification of the test accuracy and its power are estimated through simulations by performing the test multiple times on subsampled datasets $\mathcal{X}^1$ and $\mathcal{X}^2$. The sets, each of size n, are sampled without replacement from the initially generated signal and background data. The consider test itself has been adjusted to strike a compromise between

Variable name	Short description	Remarks
HT	Sum of all the jets transverse momenta pT	
M	All jets invariant mass	
Mjj^{lead}	Leading dijet invariant mass	Dijet is a pair of jets selected based on a specific pairing method (ATLAS Collaboration, 2016).
Mjj^{trail}	Trailing dijet invariant mass	See above.
pTi	Transverse momentum of the i^{th} largest jet pT	We consider momenta for 4 jets, i.e. $pT1$, $pT2$ $pT3$ and $pT4$.
$\Delta\phi ij$	Azimuthal angle in the transverse plane between the i^{th} and j^{th} jets ordered by pT	In specific we consider $\Delta\phi 12$, $\Delta\phi 13$, $\Delta\phi 14$, $\Delta\phi 23$, $\Delta\phi 24$ and $\Delta\phi 34$.
$\Delta\eta ij$	Pseudorapidity difference between the i^{th} and j^{th} jets, ordered in pT	We consider $\Delta\eta 12$, $\Delta\eta 13$, $\Delta\eta 14$, $\Delta\eta 23$, $\Delta\eta 24$ and $\Delta\eta 34$.

TABLE 2.3: Description of the 20 considered data variables for the multijet final state analysis.

the power and computation time, given that for the simulations purpose it is required to perform it many times. It has been chosen to set the samples size n to 2000. As well, we settled on taking $S = 40$ sets of data variables $\mathbb{T}_s$, each consisting of three distinct features. We observed that for more extensive sets $\mathbb{T}_s$, the density estimation could be not accurate enough regarding the moderately small sampled datasets; additionally, the samples dimension is limited with the computation time, which is related quadratically to their dimension. The selection of feature sets $\mathbb{T}_s$ is taken at random from all the possible choices of picking 3 out of 20 considered variables. The number of performed data permutations B within the framework is 400. For such parameters, the single test takes about ten hours on a 20-core computing machine.

While datasets $\mathcal{X}^1$ and $\mathcal{X}^2$ are both, at least partially, repeatedly subsampled from the original background data, the performed statistical test p-values might be correlated. However, as the original background data size is much larger than the subsample size (almost 100 times larger), the possibly appearing correlation is negligible. If one aims at removing this unwanted effect entirely, the initial datasets should be divided into two sets before sampling within the simulations. However, in that case, the resulting p-values might depend on the former division, leading to an incorrect inference.



Selected percentiles	Fisher combinant	Min-p combinant
0.01	0.01	0.00
0.02	0.02	0.01
0.05	0.04	0.01
0.10	0.09	0.02
0.20	0.18	0.17

FIGURE 2.3: On the left-hand side, the empirical cumulative distribution function of p-values for the considered KDE permutation tests under H_0 hypothesis. The number of sampling R is equal to 120. Two combination functions are used: the Fisher (green dashed) and the min-p (black dotted). The blue line is the uniform CDF. On the right-hand side, the table of some selected percentiles of the p-values presented graphically in the adjacent figure.

2.4.2 Type-I error

The accuracy of a statistical test can be defined as the test ability to incorrectly reject the null hypothesis at the nominal level α – the significance level – more precisely the type-I error rate. For a continuous test statistic, when the null hypothesis is true the p-value is uniformly distributed, by definition of α as the probability of the type-I error – the only way to commit the error with an arbitrarily selected probability α when the associated p-value is smaller than α, occurs only when the p-value is uniformly distributed.

We are interested in verifying the type-I error of the KDE permutation test, in particular, if it does not reject the null too often regarding the significance level. For this purpose, we sample without replacement sets $\mathcal{X}^1$ and $\mathcal{X}^2$ from the background data $\mathcal{X}$. For this simulation scenario, both sampled sets have the same distribution, i.e. the datasets conform to the null hypothesis. Given the samples $\mathcal{X}^1$ and $\mathcal{X}^2$, we perform the test and obtain its respective p-value for the tested hypothesis. The procedure is repeated $R = 120$ times to compute the empirical cumulative distribution function (CDF) of the p-values. While to guarantee reliability of the results, the number of simulations should be as large as possible, a single simulation step takes about ten hours on a multi-core computing machine. Hence, due to the computation time burden, only 120 simulations have been performed.

Results are presented in Figure 2.3. The empirical CDF for an accurate test should

be close to the uniform CDF. As the significance level α is usually selected to be a small value, both CDFs should overlap particularly well for the lowest argument values. In the studied case, the Fisher combinant can correctly keep the type-I error while the min-p combinant for the KDE permutation test is much more conservative.

2.4.3 Test power

To analyse the performance of the Hemisphere Mixing, we need to employ a statistical test that not only controls the first-type error but also offers a small second-type error probability, i.e. one that with high probability correctly rejects the null hypothesis when it is indeed false. The rate of type-II error β is equivalently determined by the *power* of the statistical test – denoted by $1 - \beta$. We need to evaluate how often the test is capable of rejecting the null hypothesis when the tested samples are drawn from different distributions. This is, in general, an ill-posed question, as we are not specifying the alternative hypothesis; however, a simplified procedure can be undertaken.

In the considered framework, the more "separated" are the tested distributions, the easier is to reject the null. Hence the test power can be measured as a function of the signal contamination in the samples. We compute it for a sequence of signal proportions contaminating the datasets.

To be more specific, in analogy to the previous section, from the original data we subsample two sets $\mathcal{X}^1$ and $\mathcal{X}^2$. The set $\mathcal{X}^1$ is sampled entirely from the background data $\mathcal{X}$, while the set $\mathcal{X}^1$ consists in $\lambda\%$ of signal and $(100-\lambda)\%$ of background observations sampled from the respective original datasets (to closely resemble the experimental data $\mathcal{Y}$). In contrast to the previously described type-I error analysis performed under the null, the two samples $\mathcal{X}^1$ and $\mathcal{X}^2$ are indeed taken from different distributions. The difference between them increases with the signal fraction λ. The proposed KDE permutation test is performed $R = 80$ times for the different generated datasets, and the summarised results are presented in Table 2.4. In specific, the test power results for the Fisher combination function are presented because the ones of the min-p are much lower due to its conservativeness. In similarity to the previous simulation scenario, the number of simulations R is low due to the long time of performing each simulation step.

In order to be satisfactory, results should ehibit a power at least greater than the II-type error as, in fact, does not occur. It should be noted, however, that to keep simulations realistic, since a possible signal is expected to be poorly represented in the data, the considered alternative hypothesis are, in fact, almost indistinguishable from the null. Additionally, with respect to a standard hypothesis testing framework, in the current case we would rather aim at not rejecting H_0 - as this would mean that

	Signal fraction		
Significance level α	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.10$
0.01	0.013	0.018	0.038
0.05	0.050	0.118	0.175
0.10	0.138	0.200	0.275

TABLE 2.4: Fraction of cases for which the null hypothesis was correctly rejected by the KDE permutation test with the Fisher combinant for significance levels α equal to 0.01, 0.05 and 0.10. 80 pairs of samples were generated under the alternative hypothesis for each background contaminated data with values of signal fraction s equal, in turn, to 1%, 5% and 10%.

the Hemisphere Mixing algorithm could be applied for a large-scale experiment. In this perspective, the first-type error is more important despite the seemingly low test power. Also, to keep acceptable the computation burden and get simulation results in reasonable times, the sample size has been set to be rather small with respect to a standard physical framework where the Hemisphere algorithm would be applied. While such analysis is object of future research, the expectation is that a larger sample size would influence the power growth. Finally, the proposed method is the only one working for the given criteria and answers the question of interest by accounting for the data specificities.

2.5 Physical application

2.5.1 Exploratory analysis

As the first step, we perform an exploratory analysis of the simulated data at hand to inspect the algorithm performance visually. Later the introduced statistical test is applied to evaluate a potential significant difference between the considered distributions.

For the exploratory step, we compare graphically the empirical distribution of the background data with the hemisphere mixed one. We consider univariate representations of four kinematic variables which are frequently used as background/signal discriminators in experimental physics; these are HT, $pT1$, Mjj^{lead} and Mjj^{trail} (see Table 2.3).

In the physical community, it is particularly common to visualise data by drawing their univariate histograms constructed from variables binning (Kramer and Spiesberger, 2018). However, the histogram comparison of two samples with similar distributions is problematic due to the large per-bin variance. In the context, kernel densities are

more suitable for visualisation of univariate overlapping densities (Sheskin, 2003). It mimics histograms while allowing for greater flexibility and smoothness. With respect to Expression (2.2), kernel density estimation is used here in a simplified univariate form. Given a set of one-dimensional observations $x_1, \ldots, x_i, \ldots, x_n$, a kernel density estimate at x is defined as

$$\hat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{2.3}$$

where K is the kernel and h is a smoothing parameter selected manually or according to an optimality criterion. Herein, the Gaussian kernel is used and the bandwidth parameter h is chosen according to Silverman's "rule of thumb" so that the mean integrated squared error of the estimated density is minimised under specific conditions (Silverman, 1986, p.48–55).

For the sake of comparison, it is illustrated the difference between background and signal distributions of the data at hand. The plots of marginal distributions of the two different samples are presented in the left column of Figure 2.4. In the right-hand side column of Figure 2.4, the 10% contamination effect by the signal on the background is presented from which it is evident that a 10% contamination is very well distinguishable, especially for the invariant mass.

To check the algorithm performance regarding its ability to retain background properties of the input data, the univariate kernel density estimates for the background and hemisphere mixed background samples are displayed (Figure 2.5). While not exactly identical, the distributions exhibit very similar behaviour, thus indicating first descriptive evidence of the Hemisphere Mixing algorithm effectiveness. Figure 2.5 also includes normalised stacked plots of the respective densities to ease the comparison (Bolker, 2008). If the two distributions were equal, their percentage in the composition should oscillate about 0.5 without exhibiting any significant peaks. This is what we roughly observe apart from relevant disproportions for the variable extremes, which is likely to be caused by the well-known inconsistency of kernel estimates at support boundaries (Karunamuni and Alberts, 2005).

The second important point to verify is to check the signal smearing property (Section 2.2), the Hemisphere Mixing is used to produce the output data $\mathcal{Z}$ from the mixed data $\mathcal{Y}$ of the background with 10% of signal observations. The hemisphere mixed data are compared with the pure background sample $\mathcal{X}$ (devoid of any signal contamination). The impact of the mixing is presented against the background data in Figure 2.6 which visually indicate that the algorithm works according to its expectations, although some

of the distributions do show slight differences. Especially noteworthy is accurate modelling of the dijet mass distributions as it is often the relevant variable for the signal discrimination. Additionally, in practice, we deal with possible signal contaminations well below one per cent, here the signal fraction is voluntarily increased in order to be able to see by eye the effect of the mixing procedure. A 10% signal contamination would be clearly detectable as a peak for the dijet mass distribution (Figure 2.4), while Figure 2.6 displays that this is not the case for the hemisphere mixed data.

The qualitative comparisons discussed above indicate preliminarily that the Hemisphere Mixing algorithm can enjoy the expected properties. However, before making any early claims, a corresponding multivariate statistical analysis and sensitive hypothesis tests are applied for that aim.

2.5.2 Application of the framework

Since the accuracy and, to some extent, the power of the KDE permutation test with the Fisher combination function have been verified, the test can be applied to the hemisphere mixed data. In contrast to the previous approaches, we do not have to perform many tests on different samples to analyse the distribution of its p-values. Instead, preferably a single test applied to the whole data should be performed. However such computation is infeasible. Hence again, due to the long computation time, the tests are applied given subsamples from the respective data. Although in contrast to the former simulations, the size of subsamples is increased from $n = 2000$ to 15000, which would as well allow drawing inference on the stated hypothesis, and the test itself would have a larger power.

The Hemisphere Mixing is expected to produce output data distributed according to its input data dominant process. For first, the hemisphere mixed data $\mathcal{Z}$ are produced from the background data. Such output is compared against the other background sample using the KDE permutation test. The resulting p-value for the test is presented in the first row of Table 2.5 showing that there is no evidence against the null hypothesis at any reasonable significance level α.

A further desirable feature of the Hemisphere Mixing algorithm is that its output data are smeared out from the signal evidence contained in the input data. A hemisphere mixed sample from a mixture of background with 5% of signal observations is produced, to validate such expectation. A 5% contamination is absolutely off-scale in the case of the search for the tiny $hh \to b\bar{b}b\bar{b}$ signal predicted by the Standard Model in the LHC data. This test is meant to try and see where the background modelling "breaks down". The hypothesis is tested given the output data and the pure background sample. The

Sample tested against a pure background	p-value
Hemisphere mixed background events	0.224
Hemisphere mixed data from mixture of 95% background and 5% signal observations	0.284
Hemisphere mixed data from mixture of 90% background and 10% signal observations	0.005

TABLE 2.5: Obtained p-values for the KDE tests in the permutation framework to verify if the Hemisphere Mixing approach performs according to its purpose. The tests are performed on samples of size $n = 15000$.

resulting p-value of the test is given in the second row of Table 2.5. This also shows no evidence against the null hypothesis at any reasonable significance level α. We then test a 10% signal contamination of the algorithm input data and verify if in that case, the test does reject the null hypothesis also at the significance level $\alpha = 0.01$ – results displayed in the last row of Table 2.5. We have thus verified that a breaking point of the method is reached for signal contamination not larger (but possibly lower) than 10%. For the 5% signal contamination the null hypothesis cannot be rejected, and it is unclear whether it is influenced by the lack of power or the actual assumed performance of the algorithm. The object of future research is to determine the test power for larger subsamples, and until that time, the sensible problem of determining the Hemisphere Mixing algorithm performance is left as an open research question.

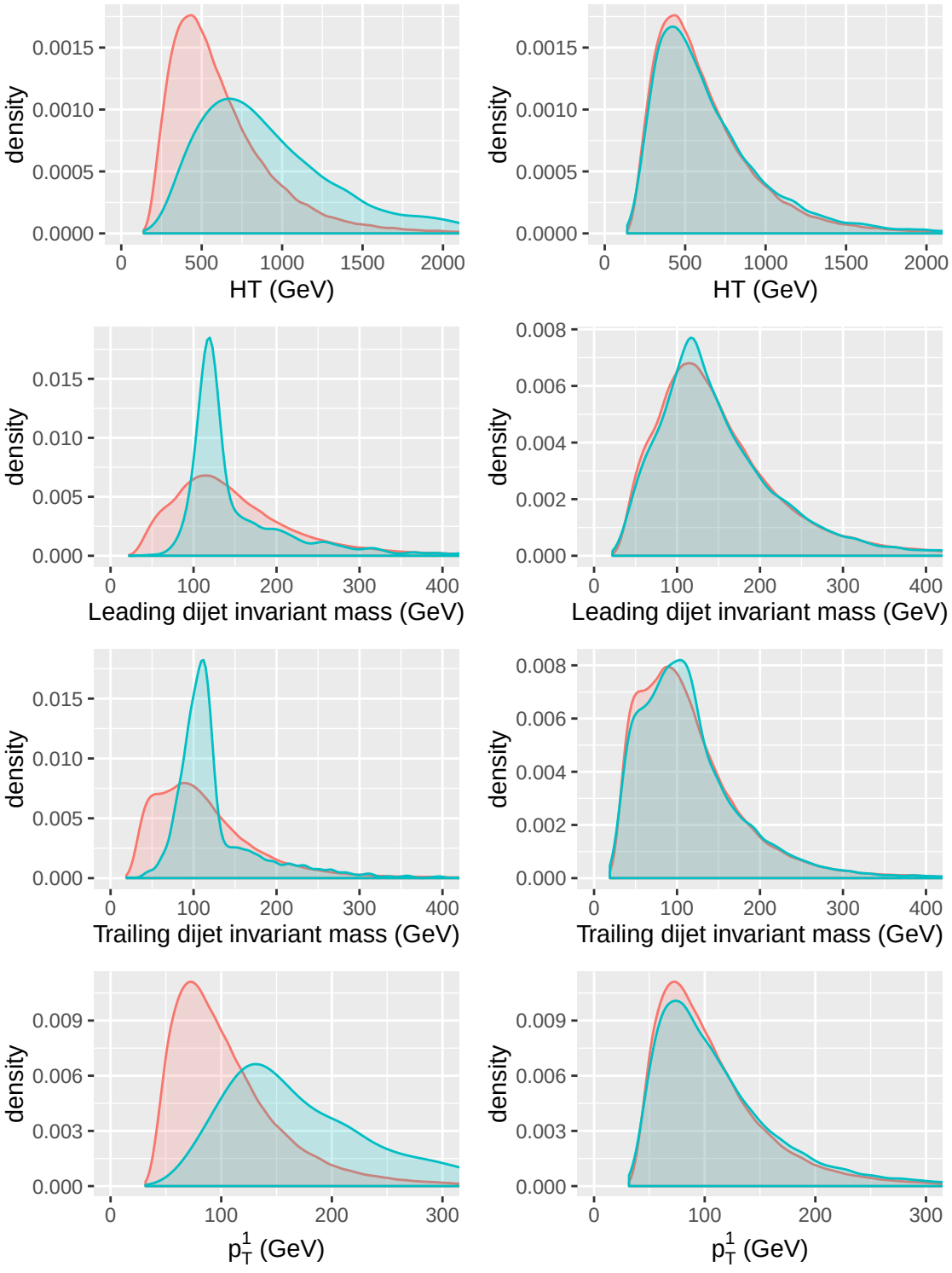


FIGURE 2.4: On the left side are displayed the kernel density estimates of four kinematic variables for background (red) and signal (blue). On the right panel a mixture of 90% background and 10% signal (blue) is compared to the background alone (red). A Gaussian kernel and Silverman's "rule of thumb" for bandwidth selection are used (Silverman, 1986, p.48).

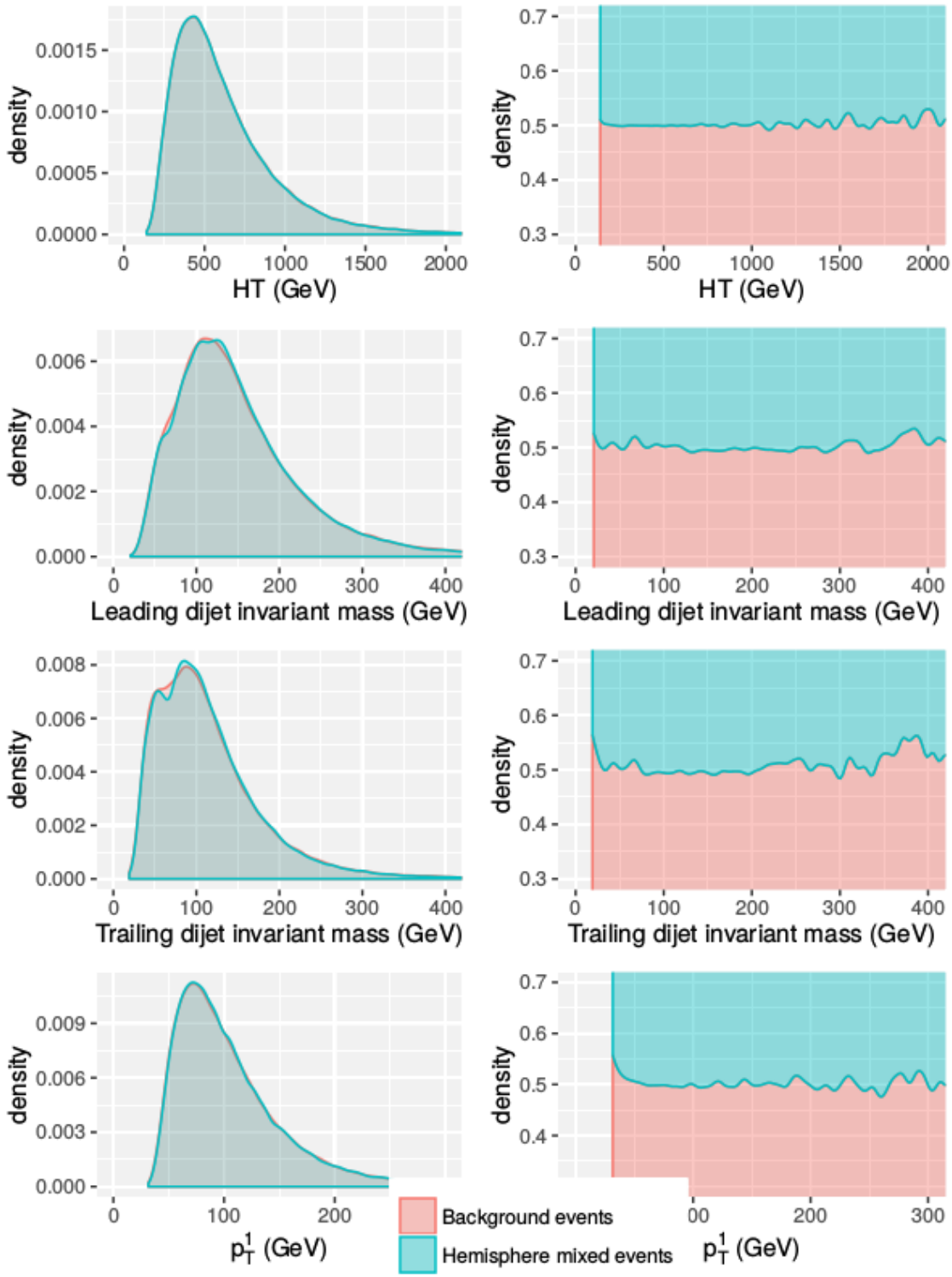


FIGURE 2.5: On the left hand side, it is shown the kernel density estimate of marginal distributions for the chosen kinematic variables (Section 2.4.1) of pure background (red) and their respective hemisphere mixed background data (blue). On the right hand side, the normalised stacked plots of the kernel density estimates.

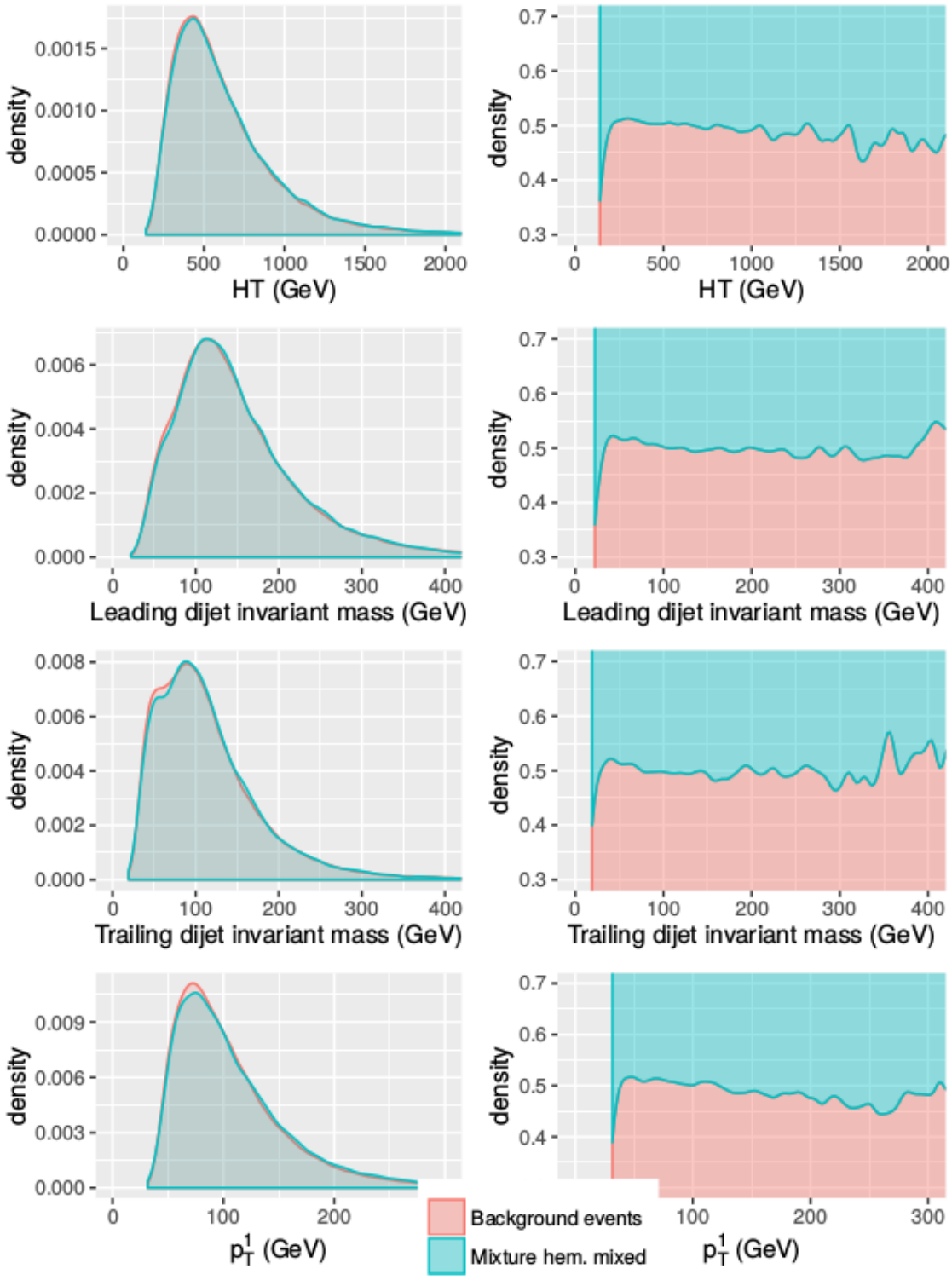


FIGURE 2.6: Left: comparison of the distributions of the four kinematical variables for background alone and the hemisphere mixed data of a sample constituted by 10% signal and 90% background. Right: stacked plots of the estimated densities.

Chapter 3

A penalized likelihood-based approach for new physics searches

3.1 Introduction

Due to the incompleteness of the Standard Model, a relevant strand of research addresses the aim of searching for new possible physics, not accounted for by the current dominant theory. The model-independent approach, followed in this work, is not constrained to any specific physical conjecture. It pursues the aim of new physics searches by exploring the experimental data and looking for any possible signal which behaves as a deviation from the background process, representing, in turn, the known physics.

This assumption clarifies why, from a statistical point of view, the considered problem can be described within the anomaly detection framework. Specifically, the aim of discriminating the background process from a possible signal, requires a classification task, although of a very special nature. While the background process is known and a Monte Carlo sample of virtually infinite size can be generated from it, the signal process is unknown, possibly even missing.

Available data have, consequently, two different sources: a first, labelled, sample from the background class only, and a second, unlabelled sample which might include observations from the signal. A semi-supervised perspective shall be then adopted, either by relaxing assumptions of supervised methods, or by strengthening unsupervised formulations via the inclusion of additional information available from the labelled data.

In the present Chapter we stem from a semi-supervised approach to the problem of signal detection, based on a suitable adjustment of the parametric clustering framework (Vatanen *et al.*, 2012; Kuusela *et al.*, 2012). Due to the high dimensionality of the available data, issues related to computation and accurate parameter estimation emerge.

The formulation is then extended to include a dimension reduction step, via the inclusion of an appropriate penalization of the likelihood associated to the model. The proposed approach is inspired by the contribution of Pan and Shen (2007) introduced in the context of model-based clustering. In that manner, the parameter estimation is performed jointly with the dimensionality reduction.

3.2 Literature overview

Anomaly detection is a relevant topic in Statistics, gathering many methods for a large variety of possible applications (Chandola *et al.*, 2009; Pimentel *et al.*, 2014). The field intersects with other statistical issues as *novelty detection* (Markou and Singh, 2003) or *noise removal* (Dotto *et al.*, 2018). They respectively aim at incorporating a discovered pattern into a normal model, and removal of observations that unnecessarily influence modelling of a prevalent data part.

Depending on the available information, anomaly detection methods are designed within three scenarios. The supervised approach falls into the predictive modelling task, however can require careful adjustments for an imbalanced class effect (Zhai *et al.*, 2017). If observation labels are completely not accessible, unsupervised techniques are widely applied, as for example the one-class Support Vector Machines (Schölkopf *et al.*, 2001; Xu *et al.*, 2017). The considered application to High Energy Physics is framed in a third semi-supervised schema, i.e. only non-anomalous observations are collected (the background data), but anomalies, whenever present, are not labelled - i.e. they are mixed with the background observations forming the experimental dataset. Regarding the unsupervised approach, semi-supervised techniques make use of the available data labels to strengthen the classification power of the considered approach. Similar issues have been found in applications related to spacecraft fault or fraud detections for which no fault is observed, and frauds might be too specific for learning (Dasgupta and Nino, 2000; Hundman *et al.*, 2018).

Specific anomaly detection applications use different definitions of expected anomalies. In general, the approaches are divided into the three intercepting groups: point, contextual and collective anomaly detection. The majority of the research focuses on the point anomaly detection which individually classifies each observation with respect to the rest of the data (Dickerson and Dickerson, 2000). This is the simplest approach which allows for a coarse detection without assuming any specific relationship between the observations. The contextual one is applied usually for time-series and spatial data

where detection aims at defining for a given time or space locally unexpected observations which could be normal if occurred at the different time or site (Benkabou *et al.*, 2018). In the thesis, we consider the collective anomaly detection approach for which separate observations do not need to have anomalous properties, but their common occurrence in a particular region of the data domain is unexpected (Noble and Cook, 2003). In other words, the possible anomalous observations do not necessarily need to appear in regions of low background probability.

Most of the existing literature within the physical community for anomaly detection (new physics searches) exploits naive, typically univariate statistical methods. They rely on comparisons of experimental data histograms with respective ones based on the Monte Carlo background data. The comparison is often performed via some statistical test related to the χ^2 Goodness of Fit, adjusted for multiple testing reasons. By construction, such approaches are not sensitive to potential anomalies that become manifest in multi-dimensional settings. Machine learning approaches based on Neural Networks or Boosting procedures have been applied as well to the context but they suffer from a lack of results interpretation (Baldi *et al.*, 2014).

Another common issue for anomaly detection approaches is a proper signal-oriented variable selection, i.e. possible anomalies can appear only for some of the variables or in a space manifold. The use of a full space statistical model can lead to degraded results or even would not to be feasible to be computed. For example, application arising from genomic studies (but not only), which are characterised by a high dimensional and a low sample size data setting, require variable selection (Tibshirani, 1996; Pierson and Yau, 2015). In case of the anomaly detection, the dimensionality reduction is often employed in a two-step strategy: firstly, an ad hoc method is applied variable-wise for the selection aim, and secondly, modelling is conducted for the preliminarily selected subspace (Alexandridis *et al.*, 2004; Tan *et al.*, 2005). However, such approach suffers from possible error propagation, i.e. the two completely independent steps of model learning do not guarantee that the selected variables are relevant to the subsequent learning task. In fact, they have a tremendous impact on the inference and classification (Fan and Li, 2001; Kabaila, 2005). For this reason, a possible hybrid methods with embedded variable selection techniques are more suitable for such aim (Pan *et al.*, 2006).

3.3 The reference model

In the following we assume that the (Monte Carlo) background data $\mathcal{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip}, ..., x_{iP})'$, $i = 1, ..., n$ are i.i.d. realizations from a probability density function $f_B : \mathbb{R}^P \to \mathbb{R}$. Similarly, experimental data $\mathcal{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_m)'$, $\boldsymbol{y}_l = (y_{l1}, ..., y_{lp}, ..., y_{lP})'$, $l = 1, ..., m$, are supposed to be i.i.d. realizations from a probability density function $f_{BS} : \mathbb{R}^P \to \mathbb{R}$. Since the majority of the experimental data $\mathcal{Y}$ is known to be generated by the background process, and the remaining part may be possibly generated by an unknown signal process with density $f_S : \mathbb{R}^P \to \mathbb{R}$, it seems natural to specify the density $f_{BS}(\cdot)$ as a mixture model

$$f_{BS}(\boldsymbol{y}) = (1 - \lambda) f_B(\boldsymbol{y}) + \lambda f_S(\boldsymbol{y}), \quad \lambda \in [0, 1). \tag{3.1}$$

The specification of model (3.1) complies with the parametric formulation of a clustering problem (Fraley and Raftery, 2002), where groups are associated with the components of a finite mixture of distributions. In the considered setting, each cluster represents a process of interest, namely the background and the signal.

The mixing distributions f_B and f_S are modeled to account for the flexibility required to describe complex collision processes. For this reason, we assume that both f_S and f_B are themselves mixtures of densities, thus somehow departing from the usual model specification in parametric clustering, where the mixture components are assumed to belong to some more elementary family of distributions. In particular, we consider finite mixture of Gaussian distributions, as they have been proven to serve well for density estimation and classification purpose (McNicholas, 2016):

$$f_B(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{y} | \boldsymbol{\mu}_k, \Sigma_k) \tag{3.2}$$

$$f_S(\boldsymbol{y}) = \sum_{q=K+1}^{K+Q} \pi_q \phi(\boldsymbol{y} | \boldsymbol{\mu}_q, \Sigma_q). \tag{3.3}$$

In the Equations above, K and Q denote number of Gaussian components in the mixtures, for $k = 1, ..., K$ and $q = K + 1, ..., K + Q$; π_k and π_q are the mixing proportions (constrained to $\sum_{k=1}^{K} \pi_k = \sum_{q=K+1}^{K+Q} \pi_q = 1$) and $\phi(\boldsymbol{y} | \boldsymbol{\mu}, \Sigma)$ is the P-variate Gaussian density function with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

Mixture model parameters can be estimated via maximum likelihood. Consider, for the sake of simplicity, estimation of the ones involved in the f_B (Equation 3.2).

Conditionally to the background data $\mathcal{X}$, the log-likelihood is formulated as:

$$\log L(\boldsymbol{\theta}; \mathcal{X}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right], \tag{3.4}$$

where $\boldsymbol{\theta} = (\pi_1, ..., \pi_K, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \Sigma_1, ..., \Sigma_K)$ are the model parameters. To find a maximum of the likelihood function, a numeric method has to be used as no explicit solution exists. A common choice is to employ a specifically adjusted Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm iteratively alternates between two steps: computation of the likelihood expectation given current parameter estimates and their following estimation that maximises the former expectation. The local maximum likelihood is found when the algorithm converges. By initialising the EM algorithm from varying starting point, possibly many local likelihood maxima are explored, in the hope that one of them is the global solution to the problem.

Once that the model parameters $\boldsymbol{\theta}$ has been estimated, a posterior probability for an observation $\boldsymbol{x}_i$ for being generated by each component can be determined as:

$$\tau_{il} = \frac{\hat{\pi}_l \phi \left(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_l, \hat{\Sigma}_l \right)}{\sum_{k=1}^{K} \hat{\pi}_k \phi \left(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k \right)}. \tag{3.5}$$

Posterior probabilities τ_{il} are used for proceeding classification in the model-based clustering context. Herein, to some extent the idea is adopted for anomaly detection, however, special care is taken to an identifiability issue since both f_B and f_S are themselves mixtures. In practice, given all the model parameter estimates (Equation 3.1), the posterior probability of being generated by the signal process is determined in analogy as:

$$\tau_{iS} = \frac{\hat{\lambda} \sum_{q=K+1}^{K+Q} \hat{\pi}_q \phi \left(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_q, \hat{\Sigma}_q \right)}{\left(1 - \hat{\lambda} \right) \sum_{k=1}^{K} \hat{\pi}_k \phi \left(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k \right) + \hat{\lambda} \sum_{q=K+1}^{K+Q} \hat{\pi}_q \phi \left(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_q, \hat{\Sigma}_q \right)}. \tag{3.6}$$

With respect to a standard problem of clustering, in the considered framework estimation can be carried on by taking advantage of the additional information available, i.e. by semi-supervising the procedure with the aid of the labelled data $\mathcal{X}$ generated from the background process. To this aim, Vatanen *et al.* (2012) and Kuusela *et al.* (2012) propose the fixed-background model, where parameter estimation is conducted in two steps. First, unsupervised parametric density estimation is performed based on the background data $\mathcal{X}$ and the background model $\hat{f}_B$ is obtained. Afterward, kept fixed the mean and the covariance parameters of $\hat{f}_B$, the weight λ in (1.1) and the parameters

of the new possible component f_S (3.2), are iteratively estimated on the basis of the unlabelled data $\mathcal{Y}$, by maximizing the corresponding likelihood via a suitable adjustment of the EM algorithm.

In order to obtain a solution for the semi-supervised anomaly detection problem given the specific datasets, we stem from the *fixed-background* model introduced in Vatanen *et al.* (2012) and Kuusela *et al.* (2012). The fixed-background model estimation is conducted in two steps. Firstly, unsupervised parametric density estimation is performed based on the background data $\mathcal{X}$ and the background model f_B is obtained. Subsequently, the fixed-background model is produced by extending the mixture with possible signal components not significant for the background data $\mathcal{X}$ (Equation 3.1). This is obtained using the unlabelled data $\mathcal{Y}$ and the previously estimated background parameters via maximisation of the corresponding likelihood.

3.4 Dimensionality reduction methods in mixture models

The fixed-background model is intuitive and makes a sensible use of the two datasets at hand. However, for data of dimension P, a K-component fixed-background model requires the estimation of $K(P+1)(P+2)/2-1$ parameters. Therefore, in high-dimensional spaces an accurate estimation of the parameters is compromised, due to the curse of dimensionality, as well as the aim of finding a global maximum of the likelihood and the subsequent ability to detect a possible signal.

Vatanen *et al.* (2012) propose to perform principal component analysis of the background data $\mathcal{X}$ to circumvent the curse of dimensionality problem. The fixed-background model is fitted in a subspace span by the first two principal components. However, there is absolutely no guarantee that the selected subspace would still exhibit any signal. Alternatively, a subset of the variables could be selected based on criteria related to a divergence between the marginal distributions of the datasets (Alexandridis *et al.*, 2004). However, it is unknown how the prior selection influences the subsequent model parameter estimation.

In the unsupervised context of model-based clustering, the problem of high dimensionality has been frequently addressed. Banfield and Raftery (1993) have proposed parsimonious mixtures to reduce the number of needed parameters. Data modelling depends on the selection of an "optimal" model as a trade-off between the model complexity and its accuracy (Celeux and Govaert, 1995). Bouveyron *et al.* (2007) propose a broader and a more flexible family of Gaussian mixture models where the regularisation

is obtained by constraining some component-specific parameters to be equal across the components.

An alternative approach is to jointly perform parameter estimation and variable selection. In the field of statistical regression such approach is common (Hoerl and Kennard, 1970; Efron *et al.*, 2004). Most methods make use of penalty functions that cause a shrinkage of parameters to fixed values (classical example is the LASSO, Tibshirani, 1996). For example, parameters are then estimated via the maximization of a penalized log-likelihood function

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

were γ is a regularization parameter (strength of the shrinkage) and $p(\boldsymbol{\theta})$ is a penalty function of the model parameters $\boldsymbol{\theta}$.

In the unsupervised context, a similar approach has been introduced by Pan and Shen (2007) within the Gaussian mixture model framework. Assumed the (3.2) to model the underlying standardized data, and constrained the covariance matrices $\Sigma_k = I_P$ to be equal to the identity matrix, the authors propose a penalty of the component mean vectors $\boldsymbol{\mu}_k$ which shrinks their estimates towards $\mathbf{0}$.

$$\log L_p(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i | \boldsymbol{\mu}_k, I_P) \right] - \gamma \sum_{k=1}^{K} \sum_{p=1}^{P} |\mu_{kp}|. \tag{3.7}$$

Maximum Penalized Log-likelihood Estimation (MPLE) is then performed jointly with variable selection using a modified EM algorithm.

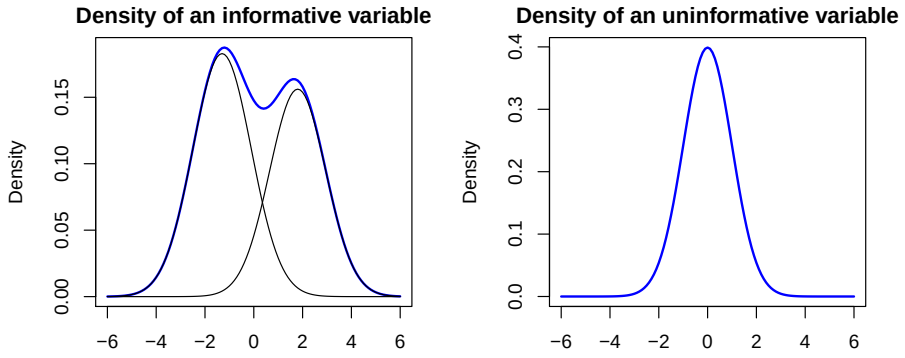
The rational behind the approach, illustrated in Figure 3.1, is the following: since, from the clustering perspective, a variable is considered informative if it can be modelled as a multi-component mixture, variables with mean components far from zero are selected, while the ones with mean close to zero are discarded. A detailed explanation is given in Section 3.5.2.

A similar approach, again based on the use of standardized data, has been introduced by Xie *et al.* (2008). They propose an l_2 penalty of the mean parameters. The penalized log-likelihood is then

$$\log L_p(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i | \boldsymbol{\mu}_k, I_P) \right] - \gamma \sum_{p=1}^{P} ||\mu_{\cdot p}||, \tag{3.8}$$

where $||\mu_{\cdot p}|| = \sqrt{\sum_{k=1}^{K} \mu_{kp}^2}$ for $p = 1, ..., P$. The penalty simultaneously shrinks a whole vector of parameters and results in a better dimensionality reduction performance (in analogy to the grouped LASSO - Yuan and Lin, 2006). Subsequently, Xie (2008) has

FIGURE 3.1: Example of informative and uninformative variables for density estimation. The first one has a more complex density (in blue) and is modelled by the two separated mixture components (in black), while the second one has component means shrunk to 0 and hence is modelled by a single Gaussian.



generalized constraints on the covariance matrices to be diagonal component-specific. This is achieved by proposing a second penalty term which shrinks component variances to 1.

In the semi-supervised context, a penalised algorithm for anomaly detection is introduced by Pan *et al.* (2006). However, the authors consider a slightly different problem and their approach is not capable of detecting novel classes. Hence it cannot be applied in the given context of new physics searches.

3.5 A penalized approach in mixture models

3.5.1 Penalization of the background

Stemming from Xie (2008) and Xie *et al.* (2008), we propose a penalized parametric approach for collective anomaly detection by extending the fixed-background model. With respect to the mentioned methods, our approach relaxes the constraints on component covariance matrices to be arbitrary positive definite. For the sake of simplicity, and since our semi-supervised approach to estimate f_{SB} first requires estimation of f_B in the sense of a standard model-based clustering problem, the proposed penalisation is first illustrated in the unsupervised setting.

The proposed procedure makes use of two penalty functions: one for the component mean parameters and one for the component covariance matrix eigenvalues. Hereafter, the method is referred to as *Mean and Eigenvalue Shrinkage Penalization* (MESP).

The first considered penalty of the MESP is then

$$p_1(\boldsymbol{\theta}) = \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \pi_k \mu_{kp}^2}. \tag{3.9}$$

It borrows the idea of the grouped shrinkage proposed by Xie *et al.* (2008) and takes advantage of the simultaneous shrinkage of component parameters (as illustrated in Equation 3.8). However, for the problem at hand, the approach should be sensitive to a precise estimation of infrequent component parameters. According to Bühlmann and Van De Geer (2011), if true proportions of components differ substantially, penalty function should be appropriately weighted to balance an influence of unequal proportions. For this reason, the proposed penalty (3.9) is also a function of the component proportions π_k which serve as weights. As a consequence, the penalised parameters are appropriately encouraged to the shrinkage, not mostly the one corresponding to the rare components.

In Xie (2008) the covariance matrices are constrained to be component-specific diagonal, and the proposed penalty is a function of the matrix diagonal terms. The second penalty of the MESP depends on the component covariance matrix eigenvalues so that the covariance matrices are not specifically constrained, but they are just component-specific positive definite. Another direction would explore the idea of matrix low-rank approximations by shrinkage of their smallest eigenvalues to 0. However, a matrix with null eigenvalues is not positive definite and would force to use generalised Gaussian distribution and pseudo-determinants in order to write the likelihood of the mixture model. However, optimization of such objective function tends to be unstable and burdensome to perform. In order to circumvent this problem, we propose to shrink the eigenvalues to a component-specific small positive value $\epsilon_k > 0$. In this way, the expected regularisation is performed, the likelihood can be written explicitly, and the EM algorithm is prevented from running into the likelihood singularities. For this approach, if the L_k smallest eigenvalues for the k^{th} component are shrunk to ϵ_k, then the number of model parameters is decreased by $\sum_{k=1}^{K}(L_k - 1)(L_k + 2)/2$.

Let us consider the eigenvalue decomposition for the k^{th} component covariance matrix $\Sigma_k = Q_k D_k Q_k'$ where D_k is a diagonal matrix of eigenvalues and Q_k is composed of orthonormal eigenvectors. Let us denote by δ_{kp} the p^{th} largest value of D_k. The second penalty of the MESP is formulated as:

$$p_2(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{p=1}^{P} \max(\delta_{kp}, \epsilon_k). \tag{3.10}$$

Selection of ϵ_k is performed based on an asymptotic distribution of the eigenvalues (Eaton, 2007). Assuming that the L smallest eigenvalues of the population covariance matrix Σ_k is equal to δ_{const}, the asymptotic distribution of the L smallest unsorted eigenvalues δ_{kl} of the sample covariance matrix is normal with a mean δ_{const} and a variance $\frac{2\delta_{const}^2}{nL}$. Mean of the L smallest eigenvalues of the respective sample covariance matrix $\hat{\epsilon}_k = \frac{1}{n}\sum_{p=P-L+1}^{P}\delta_{kp}$ is then an unbiased estimator of δ_{const}.

The parameter L_k is selected based on sequential tests. The tests partially use the same data between iterations, hence a Bonferroni-like correction is applied to control the type I error (Bonferroni, 1936). Denote by $\bar{\delta}_{k,h}$ an average of $L_k = P - h$ smallest eigenvalues of the k^{th} component sample covariance matrix. Starting from $h = 0$, it is tested in sequence if the $L_k = P - h$ smallest eigenvalues are equal to $\bar{\delta}_{k,h}$ against a general alternative (at least one eigenvalue is different). Rejection regions for the tested hypothesis are determined as

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

where $z_{\frac{\alpha}{2(h+1)}}$ is the $\frac{\alpha}{2(h+1)}$ quantile of the Gaussian random variable. In shorthand, the null hypothesis is rejected if a ratio of the largest eigenvalue and the mean of eigenvalues is too large or a ratio of the smallest eigenvalue and the mean is to low. If there is no reason to reject the null, then we assign the parameter L_k to $P-h$. Otherwise, we take the alternative hypothesis that the eigenvalues are different. Then, in a next iteration, it is assumed that the largest eigenvalue is too large, and the test is performed again with the parameter h increased by 1. With iterations the rejection regions get larger according to the type-I error correction of the sequential test. The iterations are repeated until there is no reason to reject the null.

In summary, the MESP appraoch makes use of two penalties expressed in Equation 3.9 and 3.10. The parameters estimation is performed via optimization of the following penalized likelihood with the specific adjustments for parameter selection

$$\log L_p(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right] - \gamma_1 \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \pi_k \mu_{kp}^2} - \gamma_2 \sum_{k=1}^{K} \sum_{p=1}^{P} \max(\delta_{kp}, \epsilon_k). \tag{3.11}$$

The solution of Equation 3.11 is obtained via a suitable modification of the EM algorithm. With respect to the unpenalized approach, the maximisation step for the penalised parameters is changed. Due to the shrinkage, these estimates are shifted with respect to the MLE toward the fixed values (0 in case of the means and ϵ_k for the eigenvalues). For readability, the pseudo-code for the adjusted EM algorithm is placed

in Appendix A.

3.5.2 Variable selection for the background

Under the assumption of component covariance matrices equal to identity considered by Pan and Shen (2007), the penalty of the mean parameters leads to an automatic variable selection. If for a given p^{th} variable, all the component mean parameters are equal to 0, then the p^{th} variable is uninformative for the cluster classification and the posterior probabilities (3.5) get the following expression:

$$\tau_{il} = \frac{\pi_l \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_l, I_P)}{\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, I_P)} = \frac{\pi_l \phi(\boldsymbol{x}_{ip}|0,1)\phi(\boldsymbol{x}_{i,-p}|\boldsymbol{\mu}_{l,-p}, I_{P-1})}{\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_{ip}|0,1)\phi(\boldsymbol{x}_{i,-p}|\boldsymbol{\mu}_{k,-p}, I_{P-1})} \tag{3.12}$$

where I_P is a P-dimensional diagonal matrix and index $i, -p$ denotes removal of the p^{th} variable from the i^{th} vector. After simplification of the equation it is clear that the data from the p^{th} variable do not contribute to the classification.

For a more general case of the component-specific diagonal covariance matrix (Xie, 2008), in order to remove the p^{th} variable two conditions have to be met. The first one corresponds to mean estimates equal 0 (as previously), while the second relies on the marginal component variances, i.e. for the p^{th} variable all the component variances have to be equal to 1. In such circumstances, the analogue of Equation 3.12 posterior probability can be written and the p^{th} variable does not contribute to the classification.

In a general case of unconstrained covariance matrices (also correlations are modeled), such simple factorization cannot be performed and the conditions for removing variables within the MESP approach need to be derived. Without loss of generality let us divide the variables into two sets -A and B - so that the set A contains the first R variables, B the rest, for any $R \in [1, P-1]$. Denote with $\mathcal{X} = (\mathcal{X}_A, \mathcal{X}_B)$ the consequent partition of the data, with $\boldsymbol{\mu}_k = (\boldsymbol{\mu}_{k,A}, \boldsymbol{\mu}_{k,B})$ the component mean vectors, with $\Sigma_k = \begin{pmatrix} \Sigma_{k,AA} & \Sigma_{k,AB} \\ \Sigma_{k,BA} & \Sigma_{k,BB} \end{pmatrix}$ the component covariance matrices where $\Sigma_{k,AB}$ is a block matrix built from rows in A and columns in B of Σ_k matrix.

Herein we aim at formulating the joint distribution of $f(\mathcal{X}_A, \mathcal{X}_B)$ as a marginal probability of X_B and a conditional probability of X_A given X_B. For the previous cases, it is automatic because uncorrelated Gaussian variables are conditionally independent. We generalize Equation 3.12 by the conditional factorization $f(\mathcal{X}_A, \mathcal{X}_B) = f(\mathcal{X}_A|\mathcal{X}_B)f(\mathcal{X}_B)$

and obtain the following formula:

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

where to ease the notation $\Sigma^{k,AB} = \Sigma_{k,AB}\Sigma_{k,BB}^{-1}$.

The first necessary condition for removing variables belonging to B as uninformative is to have null mean estimates $\hat{\mu}_{kp} = 0$ for all $k = 1, \ldots, K$ and $p \in B$. In that case, the posterior probability of observation membership is

$$\tau_{il} = \frac{\pi_l \phi(\boldsymbol{x}_{iB}|\boldsymbol{0}, \Sigma_{l,BB}) \phi(\boldsymbol{x}_{i,A}|\boldsymbol{\mu}_{l,A} + \Sigma^{l,AB}\boldsymbol{x}_{i,B}, \Sigma_{l,AA} - \Sigma^{l,AB}\Sigma_{l,BA})}{\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_{iB}|\boldsymbol{0}, \Sigma_{k,BB}) \phi(\boldsymbol{x}_{i,A}|\boldsymbol{\mu}_{k,A} + \Sigma^{k,AB}\boldsymbol{x}_{i,B}, \Sigma_{k,AA} - \Sigma^{k,AB}\Sigma_{k,BA})}.$$

which implicitly is a function of parameters from the presumably uninformative variables from set B. Naturally, like in the approach of Xie (2008), Hence, a second necessary condition is necessary. That is to have component-wise equal correlation matrix blocks, i.e. for all $k = 1, \ldots, K$ $\Sigma_{k,BB} = \Sigma_{BB}$ and $\Sigma_{k,AB} = \Sigma_{AB}$, for the fixed Σ_{BB} and Σ_{AB}, where Σ_{BB} is expressed as a weighted average of component specific blocks

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

and Σ_{AB} is a matrix of zeros 0_{AB}. If the two conditions are met then the cluster membership probability is:

$$\begin{aligned} \tau_{il} &= \frac{\pi_l \phi(\boldsymbol{x}_{iB}|\boldsymbol{0}_B, \Sigma_{BB}) \phi(\boldsymbol{x}_{i,A}|\boldsymbol{\mu}_{l,A} + 0_{AB}\boldsymbol{x}_{i,B}, \Sigma_{l,AA} - 0_{AB}\Sigma_{BA})}{\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_{iB}|\boldsymbol{0}_B, \Sigma_{BB}) \phi\left(\boldsymbol{x}_{i,A}|\boldsymbol{\mu}_{k,A} + 0_{AB}\boldsymbol{x}_{i,B}, \Sigma_{k,AA} - 0_{AB}\Sigma_{BA}\right)} \\ &= \frac{\pi_l \phi(\boldsymbol{x}_{i,A}|\boldsymbol{\mu}_{l,A}, \Sigma_{l,AA})}{\sum_{k=1}^{K} \pi_k \phi\left(\boldsymbol{x}_{i,A}|\boldsymbol{\mu}_{k,A}, \Sigma_{k,AA}\right)}. \end{aligned} \tag{3.13}$$

As a result, the variables from set B do not influence the membership probabilities. Hence if the two listed conditions are met, the variables from set B should be removed as the uninformative. While the first condition is obtained automatically by the component mean shrinkage, 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 subsequently the set B consists of potentially uninformative variables. Let us denote by C a set of all the possible subsets of set B ($C = \{C_1, ..., C_{N_B}\}$ for an appropriate N_B value) and $D_i = C_i^C$ is a C_i complement. The Bayesian Information Criterion is then used to select an optimal set C_i of the uninformative variables. Based on the selected model, for all

k in $1, ..., K$ the penalized likelihood estimates $\hat{\Sigma}_{k,C_iC_i}$, $\hat{\Sigma}_{k,D_iC_i}$ and $\hat{\Sigma}_{k,C_iD_i}$ are replaced with the fixed $\Sigma_{C_iC_i}$, $\Sigma_{D_iC_i}$ and $\Sigma_{C_iD_i}$ respectively. The described method might seem computationally expensive, however, there is no need to scan all the $P!$ models. The first necessary condition already filters out most of the true informative variables.

3.5.3 Penalization of the background + signal model

The proposed MESP approach for parameters estimation and simultaneous reduction of dimensionality is designed in the unsupervised context. The specificity of the physical problem of signal detection requires then its embedding in a semi-supervised framework, by suitable extension of the fixed background model proposed by Vatanen *et al.* (2012) and Kuusela *et al.* (2012). Results are enclosed in the proposed Penalized Anomaly Detection method (PAD), discussed in the following.

In the presence of a signal, special care is taken for variable selection because uninformative variables for the background model might be powerful for signal discrimination. Hence, a proper penalty requires to be a function of both the background and signal parameters. This, in turn, causes a dependence between the background and signal estimation. Given the background model $\hat{f}_B$, the penalised log-likelihood of the penalised fixed-background model has the following form:

$$\log L_p(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(\mathcal{X}), \mathcal{Y}) = \sum_{l=1}^{m} \log \left[(1-\lambda) \sum_{k=1}^{K} \hat{\pi}_k(\mathcal{X}) \phi_k(\boldsymbol{y}_l|\hat{\boldsymbol{\mu}}_k(\mathcal{X}), \hat{\Sigma}_k(\mathcal{X})) + \lambda \sum_{q=K+1}^{K+Q} \pi_q \phi_q(\boldsymbol{y}_l|\boldsymbol{\mu}_q, \Sigma_q) \right] - \gamma_1 \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \hat{\pi}_k(\mathcal{X}) \hat{\mu}_{kp}(\mathcal{X})^2 + \sum_{q=K+1}^{K+Q} \pi_q \mu_{qp}^2} - \gamma_2 \sum_{p=1}^{P} \sum_{q=K+1}^{K+Q} \max(\delta_{qp}, \epsilon_q), \tag{3.14}$$

where $\hat{\boldsymbol{\theta}}(\mathcal{X})$ denotes the background model parameter estimates given the data $\mathcal{X}$. Optimization of Equation 3.14 results in obtaining the signal parameter esitmates.

However, as the background parameters influence the signal ones through the penalty function, similarly the signal parameters influence the background ones. In turn, given the signal parameters estimates, we consider the following penalised log-likelihood for the background model:

$$\log L_p(\boldsymbol{\theta}|\boldsymbol{\theta}(\mathcal{Y}), \mathcal{X}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k) \right] - \gamma_1 \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \pi_k \mu_{kp}^2 + \sum_{q=K+1}^{K+Q} \pi_q(\mathcal{Y}) \mu_{qp}(\mathcal{Y})^2} - \gamma_2 \sum_{p=1}^{P} \left(\sum_{k=1}^{K} \max(\delta_{kp}, \epsilon_k) \right), \tag{3.15}$$

where $\boldsymbol{\theta}(\mathcal{Y})$ denotes the fixed signal parameters.

Parameter estimation of the PAD parameters is performed by an adequately modified EM algorithm used to find the maxima of the two Equations (3.14 and 3.15) and an external loop alternating between optimisation of the Equations until convergence. The pseudo-code of the PAD based EM adjustments is reported in Appendix B.

3.6 Experimental analysis on simulated data

3.6.1 Goals of the analysis

To understand the performance of the proposed methodology in terms of classification in the unsupervised setting (i.e. to estimate the background distribution) and in the semi-supervised anomaly detection setting (to estimate the whole process density) the methods are applied to collections of artificially generated data. The simulations are designed to validate different aspects of the approach performance with respect to:

- Different implementations for handling variable selection. Within the penalised model-based clustering approach we consider and test two scenarios. The first (M1) is a result of the MESP fitted to the data of the full dimension P where the penalties serve for the regularisation. The second method (M2), also relies on the MESP but it is constructed in two steps. Firstly, based on M1, informative variables are selected. In the second step, the MESP is fitted again to the reduced-size data. The M2 model should result in a smaller estimate bias because in the reduced space slighter penalty can be applied, as in the reduced space the further model regularisation is not needed. On the other hand, it might suffer from a possible error propagation or a poorer classification performance if the informative variables are incorrectly removed in the first step.

  One might expect a substantial amount of over-fitting for such approach leading to overoptimistic performance measures. However, in this context we work in an unsupervised logic, thus the mentioned risk is not that high with respect to a supervised logic for which information about the signal is used for training the model.

- Performance of competing models, i.e. the parsimonious family of models introduced in Fraley and Raftery (2002) and Bouveyron *et al.* (2007), application of the PCA prior to the parameter estimation according to Kuusela *et al.* (2012) and the penalized approach of Pan and Shen (2007).

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

3.6.2 Simulation settings

Simulated data are generated from a mixture of Gaussian distributions. The true model parameters are chosen in such way that only particular features of the method performance are explored. The true background density is the following:

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

where

- K is chosen in $\{2, 3\}$.
- Mean vectors are set as $\boldsymbol{\mu}_k = (1, ..., 1, 0, ..., 0)' * mult$ with

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

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

- Covariance matrices are block diagonal

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

  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

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

- The data size is n in $\{250, 500\}$ with a dimension $P = 16$.

Given such data generating models, the variables $9 - 16$ are uninformative (according to Equation 3.13).

For simulations with the anomaly detection purpose, the background data are generated using two Gaussian components with the parameters specified above. The signal process is simulated by a single Gaussian component with parameters specified as for the third background component. The experimental data are generated with different proportions of signal events (λ in $\{0.2, 0.1, 0.05\}$). The generated background and experimental data sizes are $n = m = 500$. Additionally, it is tested if background uninformative variables are kept if a signal is present for them. When it is specified, the true signal mean for an arbitrary 14^{th} variable (background uninformative) is non-zero to verify if the algorithm is able to use such additional information for better classification.

3.6.3 Details

1. The simulations were performed in R environment for statistical computations (R Core Team, 2017). The code is available at

   https://github.com/Grzes91/PenalizedAD.

2. Software implementations used for model-based clustering

   - The *mclust* R package (Scrucca *et al.*, 2016) used for the Fraley and Raftery (2002) approach and for the fixed background model after suitable dimensionality reduction using the Principal Component Analysis.
   - The *HDclassif* R package (Bergé *et al.*, 2012) for the Bouveyron *et al.* (2007) approach.

3. An important aspect of the MESP is to determine an optimal number of Gaussian components K and regularisation parameters γ_1 and γ_2. A commonly used approach for the unpenalized mixture models is to use the Bayesian Information Criteria (BIC) defined by Schwarz (1978) as

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

where d is a number of the model parameters, L is the model likelihood and $\hat{\theta}$ are their MLE. According to the criterion, an optimal model minimises BIC value and trades off between the goodness of fit and model complexity. Motivated by Efron *et al.* (2004), Pan and Shen (2007) and Bühlmann and Van De Geer (2011) discuss that the shrunk parameters should not be counted for the model complexity (they are fixed not estimated) and use a modified BIC criterion formulated as

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

where d_{eff} is an effective number of model parameters (parameters not shrunk to the fixed value) and $\hat{\theta}$ are MPLE. Despite the lack of a rigorous theoretical foundation for the modified BIC, in practice, the criterion serves well for the problem of model selection (Bouveyron and Brunet-Saumard, 2014). A minimum BIC value is found based on an extensive grid search over the parameter space. For the considered simulations, we assume that the number of Gaussian components is known a priori so that the results are not dependent on a possible incorrect specification of K.

4. Performance evaluation of the tested methods is based on the classification error and the Adjusted Rand Index (Hubert and Arabie, 1985) as these are the standard evaluation measures in unsupervised problems. The latter index, is a suitable modification of the Rand index, which compare two partitions of n objects by the proportions of pairs of observations which have been allocated either in the same cluster in both the partitions, either in different clusters in both the partitions. The Adjusted Rand Index is the Rand Index accommodated to have 0 mean for random allocation of the observations between the two partitions.

   However, it has been frequently emphasised (He and Garcia, 2008; Menardi and Torelli, 2014) that the use of the aforementioned common performance measures may yield to misleading results as they strongly depend on the class distribution. For instance, in a problem with heavily imbalanced classes, a naive strategy of allocating each example to the prevalent class would achieve a good level of accuracy. However, it is clear that such classification rule is entirely useless. For this reason, the area under the Receiver Operating Characteristics curve (AUC) is also exploited (Egan, 1975) as it is immune to the class imbalanced and can better help in comparing trade-offs arising from the use of distinct classifiers.

5. The EM algorithm for likelihood optimisation is known to suffer from finding local maxima of the objective function instead of the global one. It strongly

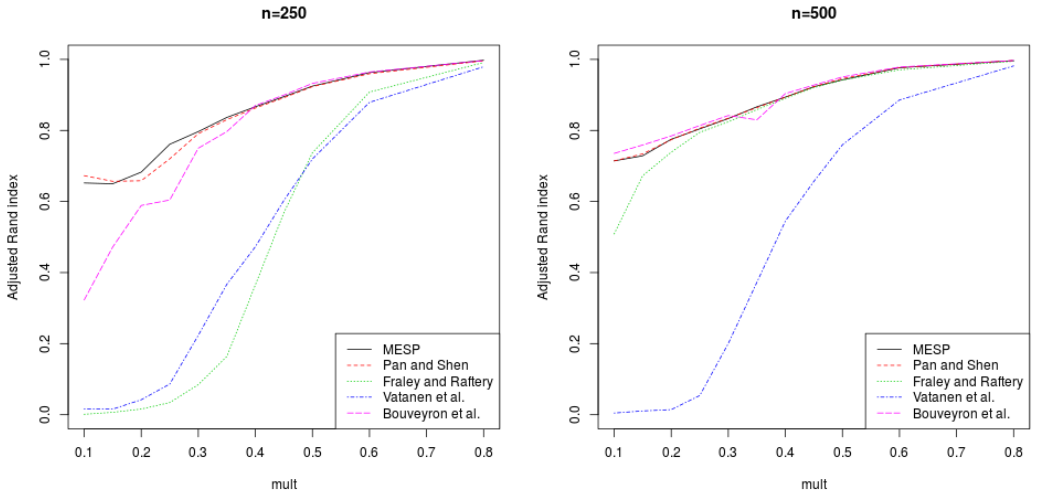


FIGURE 3.2: Performance comparison for the 5 model-based clustering methods given the datasets of size $n = 250$ and $n = 500$ generated from the two Gaussian components for the varying separation (*mult*).

depends on the algorithm initialisation values. Additionally, maximisation of the Gaussian mixture model log-likelihood function is not a well-posed problem, i.e. the likelihood tends to infinity if one of the components becomes a singularity. The modified EM algorithm employed in MESP and PAD share the same problems. However, the singularity problem is surpassed thanks to the use of the covariance matrix based penalty $p_2(\boldsymbol{\theta})$. The EM algorithm is run multiple times from different starting points, to assure that the objective function global maximum is found. In particular, for the model selection and the related grid search, the so-called warm-starts are used as initialisation values, i.e. estimates of models with smaller regularisation parameter γ become the initial values for iterations with the larger γ. Given the warm-starts, the algorithm converges much faster and is more likely to spot the global maximum.

3.6.4 Results and comments

3.6.4.1 Model-based clustering

Let us consider the scenario where data are generated from a Gaussian mixture of two equally weighted components with its parameters previously specified.

For the simple balanced data scenario, the introduced MESP approach has superior performance in comparison to the other methods (Figure 3.2), especially for the most challenging cases ($n = 250$ and the smallest true component separation). With an

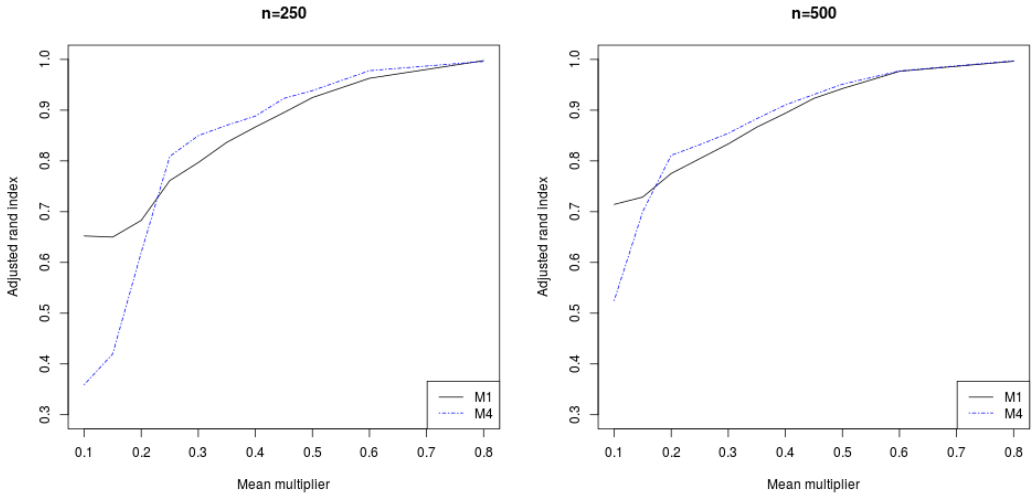


FIGURE 3.3: Performance comparison for the two types of variable selection methods (M1 and M2) for the MESP and the datasets of size $n = 250$ and $n = 500$ with varying separation (*mult*). The results are based on an average model performance on 50 simulated datasets generated from the two Gaussian components.

increase of the separation between the components, the classification is more straightforward, and all but the fixed background model has comparable performance (the first two principal components for the case explain only a small fraction of the data total variability).

Comparison of variable selection methods (Figure 3.3) for MESP suggests that the M2 model performs better in terms of the adjusted Rand index. However, for the small component separation, the variable selection approach incorrectly removes the informative variables which severely decreases the M2 performance.

In the setting with three Gaussian components and unequal weights, the classification difficulty is increased which naturally results in much lower adjusted Rand index values. The hierarchy of model performances changes slightly between the previous balanced and the current unbalanced scenarios, but the MESP, Pan and Shen and Bouveyron et al. are still the best-performing ones (Figure 3.4). For the smaller data size and the smallest separations, the MEAS has far better performance than the competitors. The performance of M1 and M2 variable selection approaches (presented in Figure 3.5) again supports the M2 one.

3.6.4.2 Anomaly detection

For anomaly detection simulations the background and experimental (background + signal) datasets are generated accordingly to the described parametrisation. The

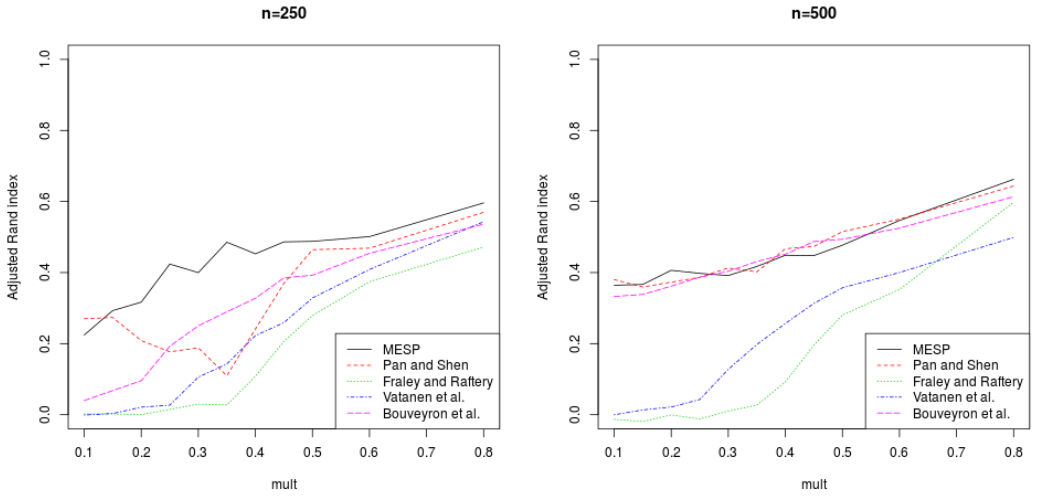


FIGURE 3.4: Performance comparison for the 5 model-based clustering methods given the datasets of size $n = 250$ and $n = 500$ generated from the three Gaussian components for the varying separation.

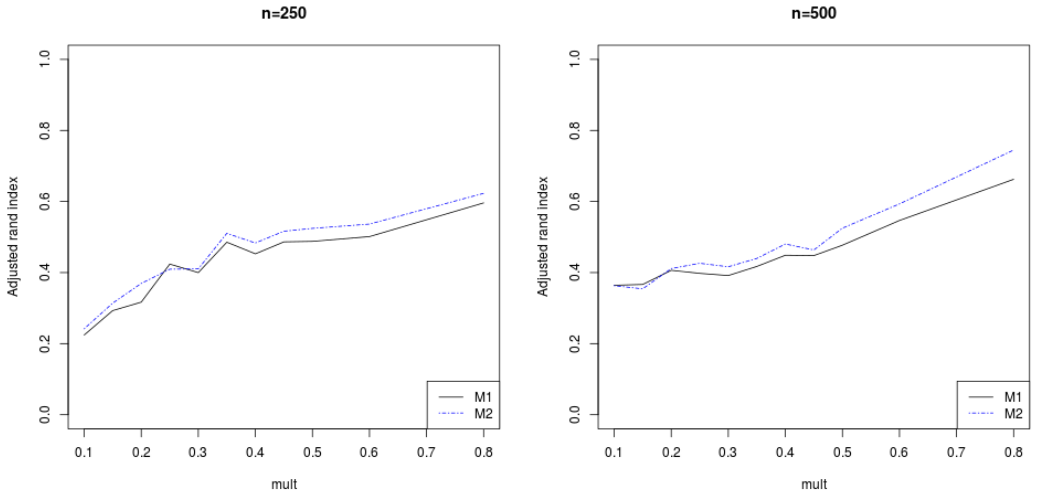


FIGURE 3.5: Performance comparison for the two types of variable selection methods (M1 and M2) for the MESP and the datasets of size $n = 250$ and $n = 500$ with varying separation (*mult*). The results are based on an average model performance on 50 simulated datasets generated from the three Gaussian components.

datasets are standardised with respect to the background data sample mean and variance. For the considered scenario, a simulated signal component lies between the two background components (the background high-density region), which causes difficulties for the signal detection.

The simulations are performed given the following datasets. First, for the signal proportion $\lambda = 5\%$, the competing methods are tested given varying separation of the background components (the first three rows in Table 3.1). By fixing the separation, an

increase of the signal proportion slightly raises the classification performance (rows 4 and 5). An overlap of the true densities disables good classification performance despite the accurate signal component parameter estimation. The next three following scenarios test if for the signal exhibiting as well for a background uninformative variable, the method performance is increased, i.e. if the applied dimensionality reduction method would not falsely remove such variable. For this purpose the true signal mean for the 14^{th} variable is set to 3. The results show that the PAD approach correctly keeps the variable and a much better signal classification is achieved in comparison to the previous scenarios. Consequently, the M2 approach is used as well for the described data scenarios to compare with the M1 (the following 8 rows in Table 3.1). The M2 approach relies on informative variables selected based on the PAD of the M1 type. Almost in all the cases, the adjusted Rand index values for the M2 are higher than for the M1.

For the tested scenarios the PAD was also compared with the Fixed Background model. However, the first two principal components represent only about $20a\%$ of the total data variability that often leads to poor a classification performance. In the last rows of Table 3.1 we see that for the cases with the signal isolated from the background high density regions, the fixed background model has a good performance. For other cases for which the signal is more difficult to be detected, it is more effective to use the PAD approach in terms of the AUC error measure.

3.7 Application to new physics searches

3.7.1 Data description

The proposed approach for anomaly detection has been applied to the Monte Carlo simulated data of high energy physics. The data simulation process has been described in Sections 1.3.1 but specifically for this application, an interest is put on collisions with a two jets final state. The background data were generated according to the Standard Model processes and the signal data according to the RPV-MSSM model with the hypothesised existence of a stop quark with a mass equal to 1000 GeV (Fuks, 2012; Barbier *et al.*, 2005).

As the proposed PAD approach is based on the Gaussian mixtures, it is particularly powerful for modelling elliptically distributed data. The mixtures could be used for the other cases as well, but then a prohibitively large number of components might need to be used. The considered physical data are heavily skewed, hence the Tukey's Ladder of Powers transformation is applied variable-wise (Tukey, 1977; Abdallah *et al.*, 2016), so

TABLE 3.1: Anomaly detection results for the different data generating scenarios and the M1 and M2 dimensionality reduction approaches compared with the fixed background model (FBM). Given 50 simulations for each scenario, the average ARI and AUC measurements are computed based on the training datasets and the AUC based on the independent testing set.

$\mu_{s,14}$	*mult*	λ	model type	ARI training	AUC training	AUC testing
0	2.0	0.05	PAD M1	0.885	0.994	0.948
0	1.5	0.05	PAD M1	0.665	0.906	0.872
0	1.0	0.05	PAD M1	0.392	0.791	0.779
0	1.0	0.10	PAD M1	0.512	0.814	0.788
0	1.0	0.20	PAD M1	0.598	0.850	0.795
3	1.0	0.05	PAD M1	0.567	0.885	0.863
3	1.0	0.10	PAD M1	0.886	0.961	0.958
3	1.0	0.20	PAD M1	0.916	0.973	0.956
0	2.0	0.05	PAD M2	0.964	0.998	0.932
0	1.5	0.05	PAD M2	0.785	0.950	0.926
0	1.0	0.05	PAD M2	0.331	0.715	0.745
0	1.0	0.10	PAD M2	0.602	0.818	0.827
0	1.0	0.20	PAD M2	0.693	0.876	0.822
3	1.0	0.05	PAD M2	0.864	0.943	0.956
3	1.0	0.10	PAD M2	0.912	0.972	0.971
3	1.0	0.20	PAD M2	0.943	0.981	0.956
0	2.0	0.05	FBM	0.906	0.989	0.994
0	1.5	0.05	FBM	0.746	0.918	0.887
0	1.0	0.05	FBM	0.176	0.613	0.592
0	1.0	0.10	FBM	0.373	0.720	0.637
0	1.0	0.20	FBM	0.465	0.842	0.782
3	1.0	0.05	FBM	0.361	0.702	0.664
3	1.0	0.10	FBM	0.502	0.773	0.739
3	1.0	0.20	FBM	0.615	0.844	0.858

that the background date univariate distributions are more Gaussian-like. The datasets are also scaled according to the background sample mean and variance. Despite the transformations, a strong dependency between the variables is still present. A few of the 2-dimensional data scatter plots clearly show complex, not elliptical patterns of the distributions. For this reason, 8 of the 19 variables are removed, and only the other 11 are further considered (listed in Table 3.2).

3.7.2 Method performance

For the different proportion λ of the signal observations in the experimental data, the PAD method of the M2 type is performed for the variable selection. For each λ,

TABLE 3.2: Description of variables used for the application to anomaly detection in context of the high energy physics. The detailed definition of the used variables can be found in (Chen, 2012).

Variable	Description
E1 and E2	jet energies
ϕ_1 and ϕ_2	azimuthal angles of jets
p_{T1} and p_{T2}	transversal momentum of the jets
ΔR	angular measure of the dijet system $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$
M_{jj}	invariant mass of the dijet
MET	missing transverse energy
$S^{1,2}$	sphericity of the dijet system
C_{jj}	centrality of the dijet system

50 experimental datasets of size $n = 4000$ are sampled from the datasets at hand. In following, we search only for a single signal component as it is already a challenging task but also sufficient to provide insight into a possible presence of an anomalous process in the data.

For the data at hand, the true distribution of the simulated signal observations lies within a region of the high background density. Additionally, the background density surpasses the signal one by a large factor (depending on the λ parameter). For this reason, posterior anomaly probabilities obtained based on Equation 3.6 are much lower than the background ones. Unless the threshold level used for the classification is somehow adjusted, the PAD method performance in terms of the correct classification is poor. The threshold level adjustments require to use some ad-hoc method which can additionally influence the performance comparison. For this reason, a different comparison method has to be employed. The area under the Receiver Operating Characteristics curve (AUC) is exploited (Egan, 1975). The results are presented in Table 3.3. As far as shrinkage is concerned, for most the cases, $4-5$ variables are chosen to be uninformative performing the effective dimensionality reduction. The fixed background model suffers from an inability of locating the signal. It is possibly due to the usage of only the first 2 principal components while the signal significantly exhibits for the other components.

TABLE 3.3: Summary of the anomaly detection results performed by the PAD M2 and the fixed background model (FBM) for datasets with different signal proportions λ. For each scenario, 50 datasets are generated to obtain a mean result with the respective standard deviations presented in brackets.

Method	λ	Average estimate $\hat{\lambda}$	Average adjusted Rand ind.	Average AUC
PAD	0.05	0.040(0.012)	0.097(0.133)	0.725(0.109)
PAD	0.10	0.057(0.013)	0.397(0.123)	0.818(0.078)
PAD	0.15	0.086(0.006)	0.507(0.029)	0.876(0.017)
PAD	0.20	0.112(0.006)	0.513(0.022)	0.882(0.012)
FBM	0.05	0.025(0.009)	0.143(0.045)	0.708(0.118)
FBM	0.10	0.046(0.008)	0.174(0.029)	0.764(0.078)
FBM	0.15	0.070(0.006)	0.185(0.018)	0.771(0.073)
FBM	0.20	0.096(0.012)	0.188(0.017)	0.780(0.054)

Chapter 4

On hypothesis testing-based approach for new physics search

4.1 Introduction

The aim to complete the Standard Model calls for advanced new physics search methods. It is not clear how the possible signal exhibits over the background distribution and providing that it does exist, its occurrence among the experimental data will be extremely rare. Hence, many approaches for signal detection have been proposed, in the hope that at least some might be preferable for a signal discovery (Popov, 2011; Baldi *et al.*, 2014). The proposed methods differ in their assumptions, methodology, sensitivity to possible signal appearance and discrimination power.

While the semi-supervised setting induced by the availability of the two data sources and the assumption that the possible signal behaves as a deviation from the background are kept unaltered with respect to the previous Chapter, in following, we consider an alternative formulation to face the model-independent new physics searches. The rationale behind the current approach is to formulate the problem in terms of hypothesis testing, i.e. to test how likely experimental data have been generated by the background distribution, under the null hypothesis that the background and the observed data process (which possibly include a signal) have the same distribution. With different specificities, the hypothesis testing approach has been frequently employed for new physics searches (CMS Collaboration, 2017; ATLAS Collaboration, 2017).

Within this framework, the *Inverse Bagging* algorithm has been recently proposed by Vischia and Dorigo (2017). As it will be highlighted in the following, some of its aspects require to be investigated to validate and possibly improve the general idea of the algorithm, as for example, the influence of the algorithm parameters on its performance

and their optimal selection. We put an effort to explain the algorithm capability based on a probabilistic work and simulation studies. Additionally, as the Inverse Bagging is not immune to issues related to the high dimensionality or correlations of the data at hand, some specific improvements are also discussed, consistently with the previous Chapter.

4.2 Description of the Inverse Bagging

In the following, we adopt notation and definitions introduced in Section 3.3. Recall the reference distribution assumed for the experimental data $\mathcal{Y}$ (3.1):

$$f_{BS}(\boldsymbol{y}) = (1-\lambda)f_B(\boldsymbol{y}) + \lambda f_S(\boldsymbol{y}), \quad \lambda \in [0,1).$$

In contrast with the previous approach, the Inverse Bagging does not rely on the explicit estimation of densities f_B and f_S, and the classification of the observations into the background or signal processes is performed according to a different logic. It is based on the idea of repeatedly testing the following hypothesis

$$H_0: \quad f_{BS}(\cdot) = f_B(\cdot) \quad \Longleftarrow \quad \lambda = 0$$

against the alternative

$$H_1: \quad f_{BS}(\cdot) \neq f_B(\cdot) \quad \Longleftarrow \quad \lambda > 0$$

based on bootstrap samples $\mathcal{Y}^*$ and $\mathcal{X}^*$, of size Q, drawn respectively from $\mathcal{Y}$ and $\mathcal{X}$.

The classical testing approach can, at most, answer to the tested hypothesis but does not explicitly provide a classification of the observations into the two processes of interest. Here we take a more complex perspective to perform multiple tests on different datasets and turn in with observation scores providing information about individual anomalous properties of the observations. In specific, by performing Nb bootstrap iterations, each observation $\boldsymbol{y}_l$ from the experimental data, for $l = 1, ..., m$, is provided with a vector $\boldsymbol{T}_l = (T_{l;1}, ..., T_{l;Tried_l})'$ reporting the test statistics obtained for the bootstrap samples $\mathcal{Y}^*$ including $\boldsymbol{y}_l$. Then based on the collected vector $\boldsymbol{T}_l$, a score for $\boldsymbol{y}_l$ is computed to reflect how likely the observation has been generated from the signal process. The scores can be used for the further classification purpose. For a precise pseudo-code of the Inverse Bagging see Algorithm 2.

Algorithm 2 Pseudo-code of the Inverse Bagging

Input: background data $\mathcal{X}$, experimental data $\mathcal{Y}$
Parameters: number of bootstrap iterations Nb , Size of bootstrap samples Q

$m \leftarrow$ size of the data $\mathcal{Y}$
allocate a list *Results* of size m consisting of empty lists
allocate a *Score* vector of size m
for $b = 1, ..., Nb$ **do**
  sample sets $\mathcal{X}_b^*$ and $\mathcal{Y}_b^*$ both of size Q respectively from $\mathcal{X}$ and $\mathcal{Y}$
  $T_b \leftarrow$ obtained test statistic for the hypothesis test given samples $\mathcal{X}_b^*$ and $\mathcal{Y}_b^*$
  for $q = 1, ..., Q$ **do**
    $\boldsymbol{y}_q^* \leftarrow$ the q^{th} observation from $\mathcal{Y}_b^*$
    to the $\boldsymbol{y}_q^*$ element of the list *Results* append T_b
  end for
end for
for $l = 1, ..., m$ **do**
  $\boldsymbol{T}_l \leftarrow (T_{l;1}, ..., T_{l;Tried_l})$ Assign a list from the l^{th} element of the *Results*
  use any method to combine values from the list $\boldsymbol{T}_l$ and assign results to the l^{th} element of the *Score* vector.
end for
return *Score*

4.3 Research questions

The Inverse Bagging algorithm is constructed based on a simple idea of multiple hypothesis testing, however, several research questions concerning its performance raise. While it seems natural to face the considered problem via hypothesis testing, since the signal fraction λ - if not equal to 0 - is expected to be small, any statistical test will suffer from little power to correctly reject the null hypothesis. Additionally, the resulting test statistics can highly vary due to data sampling, which can further result in a high variance of the produced scores and an incorrect classification. For these and other reasons discussed in the following, the algorithm needs a comprehensive investigation, aimed at validating and, possibly, improving it with respect to the following highlighted aspects.

1. Choices related to the test hypothesis

   Recall that data sampling and the iterative hypothesis testing produce for each observation $\boldsymbol{y}_l$ a respective vector of test statistics $\boldsymbol{T}_l$. A combination of its values is required to obtain an observation score summarising its signal properties. The score computation is crucial for the following classification task because observations with the most extreme scores are classified as the signal. In practice,

there are countless possible methods for the score computation, but for the sake of simplicity, herein we consider the ones which follow.

(a) **Test statistic score** - mean of the obtained test statistics for a given observation

$$R_{Test;l} = \frac{1}{Tried_l} \sum_{k=1}^{Tried_l} T_{l,k}$$

where $Tried_l$ is the number of times the observation $\boldsymbol{y}_l$ is sampled.

(b) **P-value score** - based on the test statistic distribution under H_0, the corresponding p-values are obtained $\boldsymbol{P}_l = (P_{l;1}, ..., P_{l;Tried_i})'$ from the vector $\boldsymbol{T}_l$ and their mean is used as a score value

$$R_{Pvsl;l} = \frac{1}{Tried_l} \sum_{k=1}^{Tried_l} P_{l;k}$$

(c) **Ok score** - proportion of times that for the given l^{th} observation the null hypothesis is rejected with respect to the total number that the observation was sampled

$$R_{Ok;l} = \frac{Ok_l}{Tried_l}$$

where $Ok_l = \sum_{k=1}^{Tried_l} 1\{P_{l,k} < \alpha\}$ is the number of times the obtained p-values are smaller than an arbitrarily chosen significance level α.

Note that the score computation method has a significant impact on the resulting classification. Changing the method influences the observation ranking. Additionally, the last one depends on the additional parameter α which can also have an important influence.

A statistical test used to test the hypothesis has as well its broad relevance. One should select a suitable two-sample test for density equality (some possible choices are briefly summarised in Section 2.3.1). In the following, the two-sample Hotelling's T^2 test is employed (Hotelling, 1931), which simplifies to the Student's T-test for univariate data. It is the simple, multivariate and computationally quick test, which is a great advantage regarding the Inverse Bagging multiple testings procedure. The Hotelling's T^2 test statistic is expressed as

$$T^2 = \frac{Q}{2}(\bar{\mathcal{X}}^* - \bar{\mathcal{Y}}^*)'\hat{\Sigma}^{-1}(\bar{\mathcal{X}}^* - \bar{\mathcal{Y}}^*),$$

$$\hat{\Sigma} = \frac{1}{2}\left(\hat{\Sigma}_{\mathcal{X}^*} + \hat{\Sigma}_{\mathcal{Y}^*}\right),$$

where, respectively for datasets $\mathcal{X}^*$ and $\mathcal{Y}^*$ of size Q, $\hat{\Sigma}_{\mathcal{X}^*}$ and $\hat{\Sigma}_{\mathcal{Y}^*}$ are the sample covariance matrices, and $\bar{\mathcal{X}}^*$ and $\bar{\mathcal{Y}}^*$ are the sample mean vectors. Under the null and assuming that the data are generated from the multivariate normal density, the test statistic has the $T^2_{P,2Q-2}$ distribution. In principle, the physical data are not normally distributed, but the test has been proved for being robust to face the issue (Khan and Rayner, 2003). The other problem is that for the Hotelling's T^2 a different hypothesis is tested, namely the equality of sample means, which, in general, does not mean the equality of their distributions. This might result in poorer performance in respect to another proper test, but despite this, the simple test is sufficient to provide insight into the algorithm performance regarding its other important aspects which might have been more difficult for other complex tests. Advanced statistical tests, as for example the Energy test (Aslan and Zech, 2005) or the kernel density based global two-sample comparison test (Duong and Schauer, 2012), can be certainly used within the algorithm, but their influence and performance are left for future research.

2. Selection of parameters Q and Nb

   The output of the Inverse Bagging algorithm not only depends on the used statistical test or the method to compute observation scores but also on the inherent parameters that drive the algorithm functionality. First, let us consider the parameter Q – the size of the sampled sets. For applications employing bootstrap sampling, it is usually chosen that the sample size Q is equal to the original data size m as it would be some rule of thumb. In contrast, Buja and Stuetzle (2006) show that bootstrap sample size has a direct influence on an error of bootstrap estimators. In this logic, Vischia and Dorigo (2017) suggest choosing the parameter Q to be much smaller than m. By this mean, the proportion of signal observations in some of the sets $\mathcal{Y}^*_b$ can be much higher than λ which can effectively increase the test power and enable the signal detection. This aspect is particularly thoroughly studied in the thesis. The concerns are that a small background sample $\mathcal{X}^*_b$ may contain insufficient information of the background. Consequently, one can ease the sampling procedure, so that for $b = 1, ..., Nb$ only sets $\mathcal{Y}^*_b$ are sampled, and the associated statistic is computed regarding the whole background data $\mathcal{X}$.

   Another fundamental parameter of the Inverse Bagging algorithm is the number of performed bootstrap samplings – the Nb parameter. Naturally, to obtain the best possible performance, it should be infinite, which is prohibited by the limited computational supplies. Hence, the parameter Nb is selected to be reasonably high

in respect to the resources. However, when it comes to comparing the algorithm performance as a function of the parameters Q, the parameter Nb cannot remain fixed across simulations. The expected number of times that each observation is sampled $E(Tried_l)$ changes with Q because

$$E(Tried_l) = \frac{Nb\,Q}{m}.$$

For a fair comparison, observation scores should be computed based on the same number of tests, hence $E(Tried_l)$ is fixed by appropriately adjusting for the parameters Q and Nb.

3. General performance with respect to competitors

   The Inverse Bagging is expected to detect anomalous property locally given unexpected collective behaviour of observations. To verify this property, the algorithm performance is compared with a similar method oriented on the global data properties. A suitable competitor is the well-known Linear Discriminant analysis (Izenman, 2008) – for short the LDA. The LDA is formulated similarly to the Hotelling's T^2 test statistic. On the other hand, it is a supervised method that cannot be directly applied within the context of interest. As mentioned in Section 3.1, a semi-supervised problem can be faced either by strengthening unsupervised formulations via the inclusion of additional information (the idea highlighted for the Penalised Anomaly Detection in Chapter 3) or by relaxing assumptions of supervised methods. Here the LDA is adapted to the semi-supervised setting and is considered herein as a benchmark method.

   The classical LDA approach serves for a two-class classification under the assumption that the two classes are generated according to Gaussian distributions with different mean vectors and common covariance matrix Σ. For classification purpose, a discriminant vector $\boldsymbol{w}$ is computed as

$$w = \hat{\Sigma}^{-1}(\bar{\mathcal{X}}^* - \bar{\mathcal{Y}}^*).$$

   The classification of new unlabelled observations is performed according to a hyperplane perpendicular to the vector $\boldsymbol{w}$. In the supervised setting, the hyperplane location is determined by a threshold computed based on the labelled training data. In the absence of the signal labelled data, the LDA classification cannot be performed. However, if the observations are projected along the discriminant

vector $\boldsymbol{w}$, their position can be used as the observation scores, equivalently to the Inverse Bagging scores. This approach is referred to as the LDA score.

4. Algorithm improvements

   In Section 3.4, we proposed improvements for the PAD approach handling the critical issue of the curse of dimensionality which appears in the multivariate setting. That approach employs variable selection for the more accurate parameter estimation, for avoiding problems with optimisation of the objective function, and for removal of not relevant variables for further signal discrimination. Within the Inverse Bagging algorithm, highly dimensional data can be an issue depending on the used statistical test, i.e. the Hotelling's T^2 is quite robust to deal with high dimensional data but, for instance, the KDE test is not. Hence, we include some variable selection technique for the Inverse Bagging algorithm, extending it to be a more general-purpose method. However, any variable selection must be carefully performed because a possible signal which does not necessarily exhibit in the univariate distributions of the variables, might present in their multivariate structure.

   Secondly, if the tests are performed given the same variables for all the bootstrap samples, the obtained test statistics are correlated, especially highly for bootstrap samples containing many common observations. Consequently, the produced scores have a high variance which cannot be reduced even by increasing the number of bootstrap iterations Nb. On the other hand, if the tests are performed given different subsets of variables, the between tests correlation is reduced which can produce scores with a smaller variance. Hence, it appears that possible improvements of the Inverse Bagging algorithm can be made if the observations sampling is subsequent with a specific feature sampling approach exhibiting signal informative variables.

   Lately, we propose some algorithm extensions regarding highly correlated data. For such cases, the previously proposed improvements might fail. We consider a specific data transformation that de-correlates the data prior to the algorithm application.

Given the above discussion, we can more specifically describe how the research questions have been faced. Firstly, the focus is put on finding an optimal method to select the parameter Q and a best-performing method for the scores computation. The goals are attained by a probabilistic work and suitably designed simulations. The Inverse Bagging algorithm is compared with another anomaly scoring method– the LDA score

– to find its strong and weak points. Subsequently, some extensions are proposed aiming at the algorithm performance improvement to circumvent the issues related to the curse of dimensionality and high correlations of the test results.

4.4 Optimal parameter selection, choices related to the test hypothesis, comparison with competitors

4.4.1 Optimal parameter selection

In this section, we perform a probabilistic work to validate if the use of parameter $Q < m$ has any reasonable foundation. Let us consider, for the sake of simplicity, the first method of the score computation, i.e. the test statistic score. Denote by $\psi(Q|s)$ an expected value of the used test statistics for bootstrap samples of size Q given that accurately s anomalous observations are included in $\mathcal{Y}_b^*$. Without loss of generality, assume that the test statistic is higher when more evidence against the null is observed. Naturally, the function $\psi(Q|s)$ is then monotonically increasing with s as potentially more evidence against the null hypothesis rises the expected test power.

Denote by ϕ_b and ϕ_s an expected Inverse Bagging score for respectively for an observation generated by the background and signal process. By using the law of total probability we express the expectations of the scores as

$$\phi_b(Q) = \sum_{s=0}^{Q-1} \psi(Q|s) f_{Bin}(s, Q-1, \lambda)$$

$$\phi_s(Q) = \sum_{s=0}^{Q-1} \psi(Q|s+1) f_{Bin}(s, Q-1, \lambda)$$

where $f_{Bin}(s, Q-1, \lambda)$ is the Binomial probability mass function of getting exactly s anomalous observations in $Q-1$ trials with the probability of selecting an anomalous observation equals to λ (the signal fraction).

Note that $\phi_b \leq \phi_s$ because from the monotonicity $\psi(Q|s) \leq \psi(Q|s+1)$ hence

$$\begin{aligned} \phi_b(Q) &= \sum_{s=0}^{Q-1} \psi(Q|s) f_{Bin}(s, Q-1, \lambda) \\ &\leq \sum_{s=0}^{Q-1} \psi(Q|s+1) f_{Bin}(s, Q-1, \lambda) = \phi_s(Q). \end{aligned} \tag{4.1}$$

From the above, the expectation of the Inverse Bagging score for the anomalous observation is higher than for the background one. On the other hand, for fixed $Q > 2$ there exists an $\epsilon > 0$ that for all $s = 1, ..., Q-1$

$$\psi(Q|s+1) < \psi(Q|s) + \epsilon$$

because the expectation of the test statistic is monotone and bounded given the data $\mathcal{Y}$. Hence from equation 4.1 we have

$$\begin{aligned} \phi_b(Q) &= \sum_{s=0}^{Q-1} \psi(Q|s) f_{Bin}(s, Q-1, \lambda) \\ &> \sum_{s=0}^{Q-1} \left(\psi(Q|s+1) - \epsilon\right) f_{Bin}(s, Q-1, \lambda) \\ &= \sum_{s=0}^{Q-1} \psi(Q|s+1) f_{Bin}(s, Q-1, \lambda) - \epsilon \sum_{s=0}^{Q-1} f_{Bin}(s, Q-1, \lambda) = \phi_s(Q) - \epsilon \end{aligned}$$

Consequently $\phi_b(Q) \in (\phi_s(Q) - \epsilon, phi_s(Q)]$ where ϵ is a fixed value given the data $\mathcal{Y}$, parameter Q and the statistical test. It is expected that for the larger sample sizes Q, the smaller ϵ value can be found because influence of a single anomalous observation gets smaller as the sample size increases. Hence the selection of the smaller Q parameter might, in principle, allow better discrimination. However, it is not certain if usage of different values for the parameter Q does that cause a bias or an increase of the scores variance influencing the classification performance. For this reason the above reasoning does not give as a final solution.

4.4.2 Numerical work scenarios

The Inverse Bagging algorithm is applied to collections of artificially generated datasets to answer the stated research questions. The simulations are designed to address the indicated algorithm aspects individually. The data are generated from the known distributions to ease the results interpretation. The computed Inverse Bagging and LDA

scores are not directly comparable, i.e. at most, they can be used for the classification but for this aim an additional method is required to choose a suitable threshold for the score-based classifications. The methods comparison is then based on the more convenient Receiver Operating Characteristics (ROC) curve or the related Area Under the ROC Curve (AUC) (Egan, 1975).

In specific, the simulated background and experimental data have size $n = m = 2000$. To verify the impact of sampling small bootstrap samples, we check several values of the parameter $Q \in \{25, 50, 100, 250, 500, 1000, 2000\}$. For each case, we adjust the number of bootstrap iterations Nb so that on average each observation is sampled $E(Tried_l) = 10^4$ times. Furthermore, we consider the Hotelling's T^2-test to test the hypothesis (simplified to the Student's T-test for the univariate cases). Within each simulation scenario, the datasets are generated 50 times, and the averaged results are shown.

For first, the background and signal distributions are specified to be the Gaussian ones (either uni or multivariate) with a common variance parameter (or respectively covariance matrix). Such settings are then optimal for the LDA scores which are put as a challenge for the Inverse Bagging performance. Later the signal distribution is specified to be non-Gaussian to compare the Inverse Bagging performance in settings not optimal for the LDA. For the last cases, the test statistic method for scores computation is used. The experimental data distribution is specified to be a mixture according to the reference mixture model (3.1) with the signal fraction $\lambda = 0.05$. In specific, for simulation scenarios the parameters of the underlying distributions are:

- For a univariate case

$$f_B(x) = \mathcal{N}(0, \sigma^2), \quad \text{and} \quad f_S(x) = \mathcal{N}(\mu, \sigma^2)$$

  where $\mu = 1$ and $\sigma^2 = 1$.

- For a multivariate normal case

$$f_B(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \Sigma_1), \quad \text{and} \quad f_S(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma_1)$$

  where $\boldsymbol{\mu} = (1, 1, 1, 1)'$ and $\Sigma_1 = I_4$.

- For a spherical signal case

$$f_B(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \Sigma_2)$$

  where $\Sigma_2 = I_5$, and signal observations are uniformly distributed on the 5-dimensional sphere centered at $\mathbf{0}$ and with a radius equal to 3.

- For a hemispherical signal case

$$f_B(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \Sigma_2)$$

and signal observations are uniformly distributed on the 5-dimensional sphere centered at $\mathbf{0}$ with a radius 3 and for the first variable absolute value is taken, so that form the hypersphere a hyper-hemisphere is created.

4.4.3 Simulation results

4.4.3.1 Univariate Data

The simulation results for the univariate scenario, are presented in Table 4.1. In is displayed a comparison of the algorithm performance for different score computation methods are the varying parameter Q. It is apparent that the test statistic score method leads to the best performance. For the selection of the parameter Q, the optimal results are achieved for $Q = 100$ across all the methods. An interesting observation is that for $Q = 100$ the ratio $Q/m = \lambda$ – the true signal proportion. Performance of the algorithm for the most optimal parameter selection is equivalent to the LDA scores.

In Figure 4.1 the Inverse Bagging scores are presented for one of the generated datasets and for $Q \in \{100, 1000, 2000\}$. Note that they align in parallel lines with a significant slope and enjoy small variance around the lines. Additionally, the higher is the sample set size Q, the higher is the variance of the scores around the lines.

For the specified data scenario, we test the influence of the algorithm performance on the number of used bootstrap samples Nb as a function of the parameter $E(Tried_l)$. The results are present in Figure 4.2. We observe that for the large number of sampling the classification performance is equal across various values of the parameter Q, however, for the smallest Q a decent performance is reached sooner. We presume that such behavior is caused by the higher variance of scores computed for the larger parameter Q (see Figure 4.1) but while increasing the Nb the variance converges to a common value across all values of the parameter Q.

4.4.3.2 Multivariate data

In table 4.2 the mean AUC values for the Inverse Bagging performance are presented for the considered score computation methods and the varying parameter Q. In analogy to the univariate case, the test statistics score serves as the best method for the classification. Selection of the parameter $Q = 100$ is also the most preferable. In both cases,

TABLE 4.1: The mean performance of the Inverse Bagging regarding the AUC for the different methods of scores computations and diverse sample size Q shows a superior performance for the method based on the test statistics. The mean AUC for the LDA score is 0.760.

Parameter Q	Test statistic	P.value	Ok
25	0.760	0.750	0.743
50	0.760	0.751	0.749
100	0.760	0.753	0.752
250	0.758	0.750	0.748
500	0.756	0.749	0.744
1000	0.750	0.741	0.739
2000	0.738	0.720	0.714

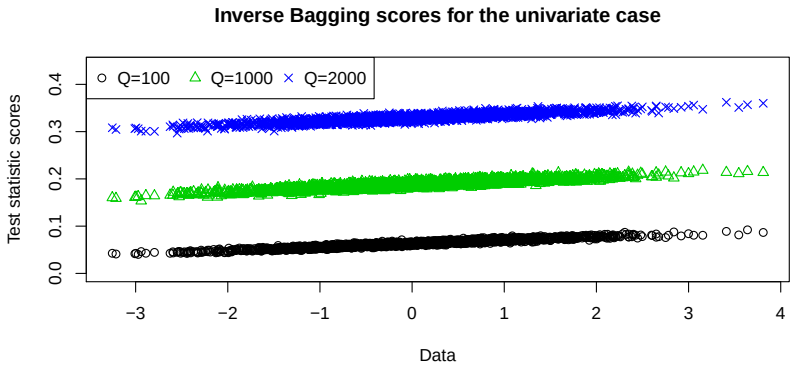


FIGURE 4.1: The Inverse Bagging scores computed based on the test statistics for the varying sample size Q plotted against the simulated univariate data. In legend the respective AUC values are shown.

the mean LDA score has equal performance to the Inverse Bagging algorithm employing the most suitable combination of the parameters.

4.4.3.3 Multivariate spherical signal

From the previous simulations, it is apparent that the Inverse Bagging based on the Hotelling's T^2-test has similar performance to the LDA score for the data generated from the Gaussian distributions with common covariance matrices which is an optimal case for the LDA performance. This time, it is validated that the Inverse Bagging is a more general method which works well for cases where the LDA score does not.

In Figure 4.4 performance of both methods is shown for the spherically distributed signal defined in Section 4.4.2. The LDA score approach has the mean AUC just slightly above 0.5, not being able to detect any anomalies. The Inverse Bagging approach has a much better result for the small parameter Q but quickly decrease to 0.5 for the larger Q. It seems that the parameter Q is not only related to a number of bootstrap iterations

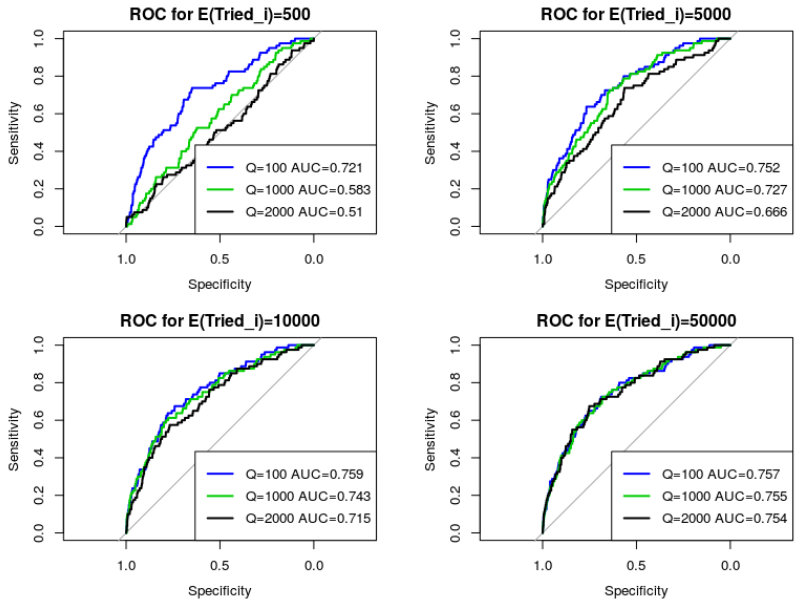


FIGURE 4.2: The ROCs and their corresponding AUC values in the legend for the Inverse Bagging scores computed based on the Ok scores computation method for the varying sample size Q and a number of the performed sampling (expressed by the parameter $E(Tried_l)$) for one of the simulated datasets.

TABLE 4.2: The mean performance of the Inverse Bagging regarding the AUC for the different methods of scores computations and varying samples size Q. The mean AUC of the LDA score is 0.904.

Score type	Test statistic	P.value	Ok
25	0.895	0.892	0.898
50	0.902	0.899	0.902
100	0.904	0.901	0.902
250	0.899	0.896	0.897
500	0.895	0.889	0.890
1000	0.891	0.870	0.871
2000	0.881	0.775	0.747

at which the optimal solution is obtained but also to a flexibility of the approach, that is to reflect respectively local and global data properties.

4.4.3.4 Multivariate hemispherical signal

The sample mean of the spherically distributed signal observation is close to **0** hence naturally the LDA scores based on the location difference of signal and background samples is going to have poor performance. Herein, we compare the two approaches for the signal with the hemisphere distribution (see Section 4.4.2). Such scenario is meant to validate if the Inverse Bagging can use both global and local properties of the data.

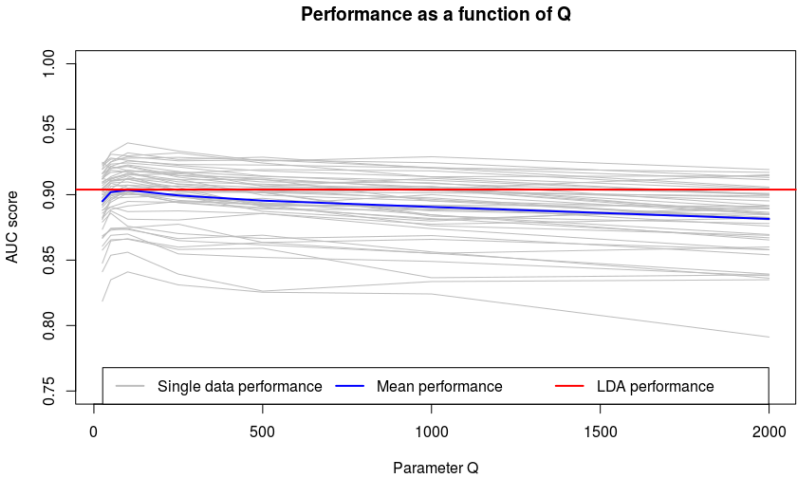


FIGURE 4.3: The AUC performance of the Inverse Bagging for the different parameter Q and the simulated multivariate datasets. In blue it is denoted the mean Inverse Bagging performance which in its maximum reaches the performance of the mean LDA score (in red).

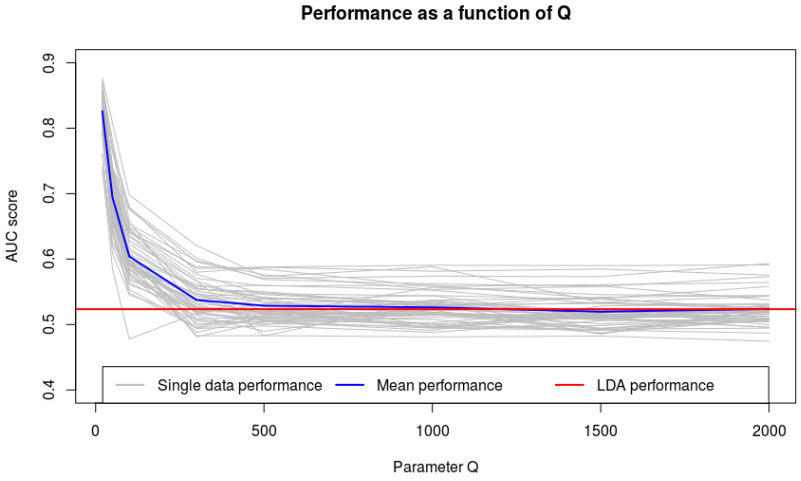


FIGURE 4.4: The AUC performance of the Inverse Bagging for the varying parameter Q and the simulated datasets with a spherically distributed signal. In blue it is denoted the mean Inverse Bagging performance and in red the average performance of the LDA score.

The performance is presented in Figure 4.5. There exists an analogy between the current results and the one of the spherical signal, i.e. the Inverse Bagging has better performance for the small values of the parameter Q while for the larger ones the performance converges to the LDA score approach.

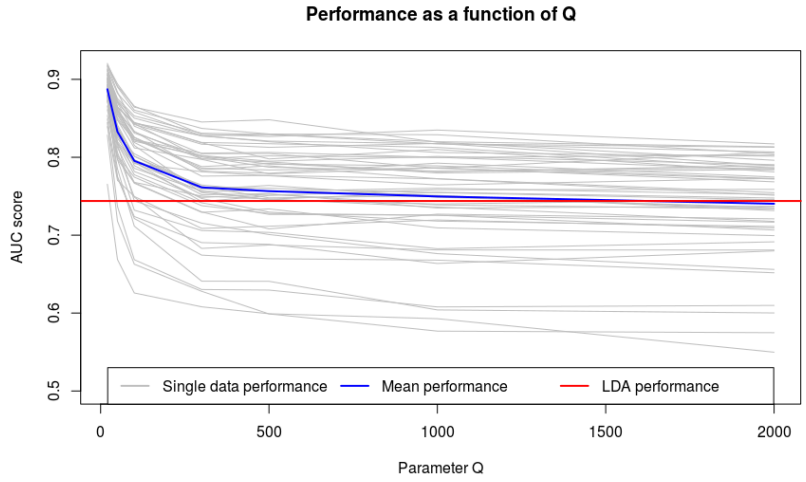


FIGURE 4.5: The AUC performance of the Inverse Bagging for the different parameter Q and the simulated datasets with signal uniformly distributed on a hemisphere. In blue it is denoted the mean Inverse Bagging performance and in red the one of the LDA score.

4.4.4 Comments

The Inverse Bagging algorithm performance has been computed based on specific simulation scenarios. It is clear that the test statistic method for scores computation is optimal from the three considered methods. Furthermore, it has an advantage, that does not depend on selection of additional parameters (like the parameter α for the Ok score) and on distribution of test statistic, which can be known only asymptotically (like for the KDE test of Duong and Schauer, 2012) or requires bootstrapping to obtain it (the Energy test of Aslan and Zech, 2005).

Selection of the optimal parameter Q is not as obvious as the selection of the score computation method. We observe that for the Gaussian data the best value of Q is about $m\lambda$ - but in practical application parameter λ is unknown - more precisely we test if $\lambda > 0$. Simulations for the spherical and hemispherical signal confirm that selection of small Q allows for discovering local data properties.

For the optimal selection of the Inverse Bagging parameters, the algorithm has comparable results to the LDA score for scenarios preferable to the LDA, and much better performance for situations not preferred to the LDA. This is an important finding, as the Hotteling's T^2-test on which the algorithm is based is very similar to the LDA, but the multiple hypothesis testing framework enables the proposed algorithm to outperform the LDA.

4.5 Algorithm improvements

4.5.1 Different score computation methods

The described methods for the scores computation are based on the averaging of various forms of test results. Although the methods - especially the test statistic one - have been proven to perform well, it is worth exploring if other methods can lead to improvements. In specific, we want to use a quantile score, based on the idea that many of the tests are performed on bootstrap samples containing a small fraction of signal observations, and hence it might be more suitable to use quantiles of high orders. In specific, the score for the observation $\boldsymbol{y}_l$, $l = 1, .., m$ is computed as a quantile of the q^{th} order from the respective test statistics vector $\boldsymbol{T}_l$. Here q is an additional parameter which needs to be arbitrarily selected.

4.5.2 Dimensionality reduction-like approach

Herein we search for possible improvements of the Inverse Bagging algorithm given the above findings considering the parameters selection. Let us refer to the original algorithm proposed by Vischia and Dorigo (2017) as the standard Inverse Bagging.

The first attempts focus on the curse of dimensionality which can emerge for some tests. Secondly, an effort is put on multivariate problems for which only a few variables are discriminative, and the others are redundant (uninformative) for the signal detection. Such cases are often encountered in applications for High Energy Physics or genomic analysis.

We propose a variable selection-like approach embedded within the Inverse Bagging. Its idea stems from the random forest approach for which variable sampling takes place when searching for optimal variables for splitting. As an Inverse Bagging extension, we propose for each bootstrap sample Y_j^* to simultaneously draw a set of L variables and perform the following statistical test in the reduced sub-space spanned by the L sampled variables. This method is referred to as the variable sub-sampling. In this manner, we allow to use tests sensitive to the curse of dimensionality; secondly, if the tests are performed on a subspace with many informative variables, the obtained information can be more reliable than from the test performed based on all the variables. Finally, correlations between the test results are decreased if the tests are performed in different subspaces which leads to the smaller variance of the scores, hence to a better classification.

The second more advanced idea of the variable sub-sampling is that for a given set of the sampled observation Y_j^*, G tests are performed on G sampled variable sets each of size L. From the G test results the maximum is taken and is used as a single step result for a sample Y_j^*. In this manner, only the highest test statistic values are saved, likely obtained based on a feature set rich in informative variables. Let us refer the method as a variable max-G-sampling (where G is the used parameter). Note that the variable sub-sampling is equivalent in notation to the variable max-1-sampling.

Further improvements refer to sampling method itself, i.e. so far in all the considered scenarios, the sampling of both observations and variables has been carried out using equal probabilities (weights). Our idea is to use specifically adjusted weights so that it is more likely to obtain signal reacher samples and perform tests on variables more likely to be informative. The idea is that in this manner, less bootstrap iteration is needed to obtain equivalent results. Two types of sampling weights are proposed:

- Feature weights - At the beginning of the algorithm employment, the weights are set to be equal as a priori is not known which variables are informative. As with each bootstrap iteration, the highest scoring variables are selected due to taking the maximum of the G tests, knowledge about the variables discrimination power is gathered. After a certain number of steps (we take 50% of Nb) this information is used to adjust the weights for the feature sampling, so that sets rich in informative variables are more likely to be sampled.

- Observation weights - As no prior information about experimental data labels is given, at the beginning the observations need to be sampled with equal probabilities. However, after a decent number of iteration, we can come up with weights based on the inverse bagging score themselves to be updated with the subsequent iterations. The higher sampling probabilities are assigned to observations with high scores, so that bootstrap samples consist of sets with a higher proportion of anomalous observations. Additionally, more tests are performed for the high scoring observations to decrease their variance by the costs of the lowest scoring observations, for which accurate computation is not the priority.

4.5.3 Highly correlated data

Until this point, for all the analysed scenarios the generated data are uncorrelated. It needs to be verified if and how the correlations might influence the Inverse Bagging performance and all the proposed extensions as correlations are present in many real datasets.

The data correlations can decrease the performance of the proposed dimensionality reduction-like technique. We suggest performing a specific transformation of the data to uncorrelated it. Let us denote by H a lower triangular matrix obtained from the Cholesky decomposition of the sample covariance matrix $\hat{\Sigma}_{\mathcal{X}}$ based on the background data $\mathcal{X}$. If we multiply the original data by the inverse of the H, then it has a diagonal covariance matrix. Such transformation does not influence the LDA score performance, however, enables to use the adopted dimensionality reduction technique.

4.5.4 Numerical work scenarios

The simple multivariate Gaussian scenario described in Section 4.4.2 are used for validation of the quantile method for the scores computation performance.

For the improvements aiming at dimensionality reduction, specific datasets are generated. In similarity to the previous case, the generated samples are of size $m = 2000$. The test statistic is used for the scores computation and parameter $Q = 100$ as they appear to be the optimal selection based on the previous simulations. For each method, 50 simulations are performed. The experimental data are drawn from the mixture density (3.1) where

$$f_B(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \Sigma_3), \quad \text{and} \quad f_S(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma_3)$$

where $\boldsymbol{\mu} = (\mathbf{0}_{20}, \mathbf{1}_5)$ and $\Sigma_3 = I_{25}$. Such distribution generated anomalies which exhibit only for the last 5 variables. In turn, the first 20 uninformative variables should be removed as they do not increase the classification power and their presence is burdensome.

In order to verify the algorithm performance for correlated data, we specify different distribution. Let the background and signal density be expressed respectively as

$$f_B(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \Sigma_4), \quad \text{and} \quad f_S(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma_4)$$

TABLE 4.3: Results of the Inverse Bagging using specific quantiles of test statistics for the scores computation. The 4 different quantiles are considered and compared with the averaging method.

	Used method for the scores computation				
	Test	Quantile score of orders			
Sample size	statistic	0.50	0.75	0.90	0.95
Q=100	0.904	0.901	0.907	0.899	0.896
Q=1000	0.891	0.872	0.867	0.850	0.799
Q=2000	0.881	0.856	0.825	0.806	0.784

where $\boldsymbol{\mu} = (\mathbf{0}_7, \mathbf{1}_3)^T$ and $\Sigma_4 = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 & -0.3 & 0.0 & 0.0 & 0.0 & 0.3 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.4 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.1 & 1.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.1 \\ -0.3 & 0.2 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.5 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.1 \\ 0.3 & 0.0 & 0.0 & 0.0 & 0.5 & 0.5 & 0.0 & 0.0 & 1.0 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1.0 \end{pmatrix}$.

The covariance matrix Σ_4 cause the generated data to be correlated. The correlation, in turn, can negatively influence the proposed dimensionality reduction-like technique, because at this point it is not clear which variables are informative.

4.5.5 Simulation results

4.5.5.1 Quantile score method

The results for application of the quantile score method are presented in Table 4.3. With no doubts, the test statistic method performs better than any of the proposed quantile methods for all the sample sizes Q and the consider quantiles q. Additionally, the higher parameter q is selected, the worse performance is obtained. It is presumably due to the fact that the scores based on the quantiles have a higher variance in respect to the scores based on the averaging leading to a more accurate classification of the second method.

4.5.5.2 Extensions for sampling

The results of the described in Section 4.5.2 algorithm extensions, namely the variable max-G-sampling approaches, are presented in table 4.4. For the parameter $G \leq 5$ the

TABLE 4.4: Mean performance of the different variable sampling approaches regarding the AUC. The parameter L controls the number of selected variables for testing. The mean AUC of the LDA scores is equal to 0.865 and of the standard Inverse Bagging with $Q = 50$ is 0.862.

Method	L=5	L=10	L=15	L=20
Variable max-1-sampling	0.843	0.855	0.859	0.861
Variable max-5-sampling	0.859	0.862	0.863	0.862
Variable max-10-sampling	0.862	0.863	0.863	0.863
Variable max-20-sampling	0.864	0.864	0.863	0.862

TABLE 4.5: Mean AUC of the weighted sampling approaches for the varying parameter L given different combinations of using weights withing sampling approaches denoted by T -True - and F - False. The mean AUC of the LDA scores for the generated datasets is equal to 0.865.

Method	Var. weights	Obs. weights	L=5	L=10	L=15	L=20	L=25
Var. max-1-sampling	F	F	0.843	0.855	0.859	0.861	0.862
Var. max-10-sampling	F	F	0.864	0.864	0.863	0.862	0.862
Var. max-10-sampling	T	F	0.892	0.882	0.875	0.869	0.862
Var. max-10-sampling	F	T	0.864	0.866	0.866	0.865	0.865
Var. max-10-sampling	T	T	0.889	0.886	0.880	0.874	0.870

standard approach has better performance than variable sampling methods for any parameter L. However, when higher values of G are used, then the performance for small parameter L is increased to be slightly higher than for the standard approach. Hence, variable max-G-sampling has a potential of improving the standard approach performance using small sets of features with moderately high parameter G (number of tests from which the maximum is taken).

Given the same simulation scenario, the Inverse Bagging scores for the different weighted sampling approaches are computed, and the comparison is presented in Table 4.5. We conclude that using weights for the variable sampling could effectively increase the Inverse Bagging performance for the cases where anomalies exhibit only for some of the variables. Especially, usage of weights for variable sampling highly increase the algorithm performance, while the use of the observation weights is of a questionable profit.

4.5.5.3 Correlated data results

In Table 4.6 the mean performance for the simulations for correlated data is shown. If no data transformation is applied (parameter Rotation is F - False) then the weighted variable sampling approach does not increase the performance of the standard Inverse

TABLE 4.6: Mean AUC of the Inverse Bagging approaches on the correlated data for the varying parameter L. The mean AUC of the LDA scores is 0.907.

Method	Rotation	L=3	L=5	L=8	L=10
Variable max-10-sampling	F	0.807	0.848	0.887	0.913
Variable max-10-sampling	T	0.920	0.917	0.914	0.913
Variable max-10-sampling with var. weights	F	0.803	0.819	0.838	0.913
Variable max-10-sampling with var. weights	T	0.924	0.921	0.916	0.913
Variable max-10-sampling with obs. and var. weights	F	0.764	0.801	0.849	0.921
Variable max-10-sampling with obs. and var. weights	T	0.921	0.924	0.921	0.921

Bagging as it was the case for the uncorrelated data. In essence, the correlations prevent the weighted sampling from performing the essential variable selection. In Table 4.6 we see that the proposed transformation is beneficial to the weighted sampling performance (parameter Rotation T - True).

4.5.6 Comments

Concerning the new method for the scores computation (the quantile score), no improvements regarding the algorithm performance have been made. It seems that the score computation based on averaging has a lower variance in respect to the quantiles. Hence it serves better for the following signal discrimination.

In respect to the dimensionality reduction approaches, the improvements have been made. Firstly, we allowed to perform the tests in lower dimensional spaces and combine their results in the way that boost the algorithm performance. It is especially useful for cases that the signal exhibits only for some of the variables. Next, we investigated observation and variable sampling using unequal probabilities, which even further improved the algorithm performance. However, it was shown that for the correlated data the proposed extension could decrease the algorithm performance regarding the standard version. To circumvent the problem, the specific data transformation was proposed so that the introduces extension again can be used as the improvements.

4.6 Applications

4.6.1 Spam data

In this section, the performance of the Inverse Bagging algorithm is validated on the well-known spam data (Friedman *et al.*, 2001). The spam data became a baseline dataset to compare supervised classifiers. They consist of 57-dimensional observations labelled either as "spam" or "non-spam" messages.

For the anomaly detection purpose, the data are transformed to fit the semi-supervised context of this Chapter. From the original spam data, by sampling two sets are obtained: a background set with pure non-spam labelled observations, and a mixed set containing a mixture of non-spam and of $\lambda = 4\%$ spam observations. Both datasets are normalised according to the sample mean and variance of the background set. The resulting background set has 2761 observations and the mixed one 1174. The parameter Q is set to be equal 100. Additionally, we test the algorithm on the produced datasets and as well after performing the proposed Cholesky transformation. For variable sampling we consider the parameter $L \in \{7, 12, 22, 32, 42, 52\}$.

In Table 4.7 the performance of the Inverse Bagging algorithm is shown. As the data variables are highly correlated, it appears that the proposed data transformation is beneficial, which is in line with the previous simulation studies. The LDA score performance is also worse than the Inverse Bagging with the prior data transformation. However, in contrast to the simulation studies, for the application feature sampling method results in the most miserable performance. In general, the proposed algorithm improvements have the comparable performance with the standard algorithm setting. In absence on prior knowledge of the number of possible informative variables, selection of the parameter L is unclear, but it has relevant consequences on the algorithm performance. The unsatisfactory performance of the proposed extensions for the application is presumably due to the fact that the data variables were explicitly selected to be good discriminants for the following spam classification, i.e. all the data features are informative. This property is in contradiction to the feature sampling idea proposed as the Inverse Bagging improvement, but despite it turns out that the algorithm performance remains comparable to the standard approach even if the parameter L is misspecified (the lowest values of L).

TABLE 4.7: The Inverse Bagging classification performance using the AUC for the 4% mixed data with and without the prior Cholesky data transformation (column "Rotation"). The AUC for the LDA score is 0.839 and the one of the standard Inverse Bagging is 0.851.

Method	Rot.	L=7	L=12	L=22	L=32	L=42	L=52
Variable sub-sampling	F	0.721	0.785	0.841	0.834	0.819	0.820
Var. max-10-sampling	F	0.751	0.799	0.819	0.829	0.827	0.822
Feature weights	F	0.703	0.747	0.826	0.839	0.833	0.825
Variable sub-sampling	T	0.865	0.863	0.852	0.856	0.848	0.853
Var. max-10-sampling	T	0.847	0.859	0.852	0.850	0.850	0.851
Feature weights	T	0.846	0.851	0.844	0.841	0.849	0.850

4.6.2 Application to the high energy physics

The Inverse Bagging algorithm has been initially formulated in the context of high energy physics and consistently with the thesis framework, its application to collision data is herein performed. In following, the same data as described in Section 3.7 are used, to additionally allow for comparison of the two proposed methods, which indeed have the same purpose and settings but are designed according to entirely different ideas.

Let us shortly recall a description of the data at hand. They are simulated to mimic detector measurements for proton-proton collisions with a two jets final state at the LHC. Naturally, the background reflects the Standard Model process and the signal refers to the exotic concept beyond the current theory. To allow a fair comparison between the proposed methods, the same 11 variables used for the former approach are adopted herein. Recall that the PAD application in Chapter 3 determines (depending on scenarios) 7 or 8 informative variables out of the 11 used; hence the usage of the proposed improvement within the Inverse Bagging can be beneficial.

For the application of the Inverses Bagging, the parameter Q is set to 200 which is 5% of the experimental data size. We apply the standard Inverse Bagging and the proposed variable max-10-sampling improvement with the parameter $L \in \{5, 8\}$. Before the algorithm application, the introduced de-correlating transformation is performed. To verify the performance, the AUC measure is used (Egan, 1975) because it has not been yet understood how a suitable threshold should be chosen for scores on which the classification is performed. Four simulation scenarios are considered, namely four scenarios of generating the experimental data $\mathcal{Y}$ with the signal proportion $\lambda \in \{0.05, 0.10, 0.15, 0.20\}$. Application results are based on 50 generated datasets within each simulation scenario. The results are displayed in Table 4.8.

In line with the previous results for the artificially simulated data, in this physical

TABLE 4.8: Performance of the Inverse Bagging regarding the AUC on the physical data for several considered anomaly detection approaches; from left-hand side for the standard Inverse Bagging, the improved version of the Inverse Bagging – the max-10-sampling with the parameter $L \in \{5, 8\}$, the LDA score and the PAD from Chapter 3. Results are obtained for varying signal proportion λ based on 50 generated datasets.

	Performance of the different methods				
	Variable max-10-sampling		Standard	LDA	PAD
λ	$L = 5$	$L = 8$	Inv. Bag.	score	
0.05	0.815	0.815	0.811	0.814	0.725
0.10	0.851	0.848	0.845	0.845	0.818
0.15	0.840	0.847	0.846	0.847	0.876
0.20	0.856	0.852	0.851	0.850	0.882

application, the Inverse Bagging algorithm performance is similar to the LDA score. For this case, the proposed variable max-10-sampling improvement is comparable with the standard approach – for some scenarios presents a slight improvement. The penalised approach compares differently for various scenarios, i.e. for small signal contamination λ the performance is worse presumably due to issues with finding the global maximum of the penalised likelihood; however, for the higher signal proportion, the penalised approach has a better performance by taking advantage from the collective-anomaly detection idea. In conclusion, the two proposed approaches are based on entirely different concepts and their performance is respectively diverse.

Conclusions

Particle Physics is an appealing field for statistical science. The research at the physical experiments embraces several disciplines, ranging from engineers who build and maintain the infrastructure, through IT specialist to theorists. Their collaborative effort provides an insight into the development of the science and understanding of the universe. Especially for large experiments as for the LHC, statistics play a pivotal role and significantly contributes to the research.

This thesis has been inspired and conducted by facing real problems encountered in physical analysis, i.e. on finding answers for two general questions of the Particle Physics. With the due specifications, the focus has been put on the knowledge improvement within the Standard Model framework, and on empirical searches of unknown phenomena beyond the current theory. In following, we shortly discuss the thesis contribution, its possible impact and future work directions both within the Statistics and applications to Physics.

The effort to complete the knowledge within the current theory refers to a more accurate estimation of the Standard Model parameters. The Hemisphere Mixing algorithm is designed to facilitate such research for multi-jet events, which had not been possible using a more classical approach based on Monte Carlo simulations. In Chapter 2, the algorithm has been validated to perform according to its design purpose, provided that the input data contain at most a small proportion of signal observations. From the physical perspective, the future direction is to apply the algorithm to the real physical data within the suitable framework and to find estimates of the wanted Standard Model parameters. However, it should also be tested if the algorithm performance extends to data with different final states, presumably containing a signal with another distribution and enjoying distinct properties. From the statistical side, the introduced KDE permutation test *per se* contributes to the family of two-sample comparison tests and can be applied for complex non-Gaussian multivariate data. However, a more thorough study of the test performance is preferable, especially to optimally strike the trade-off between its power, bootstrap samples and the computation time. It is advisable to compare the test with other competitors, possibly even in less restricted settings in which also other

tests are feasible to be applied.

The second considered physical goal refers to searches of unknown physics beyond the Standard Model. They are specified within the anomaly detection framework, and two approaches are proposed. The Penalised Anomaly Detection approach (Chapter 3) originates from the family of model-based clustering. It contributes to the field by incorporating the penalised methodology that circumvents the curse of dimensionality issue and allows flexible data modelling. The approach has few weak points, for example, Gaussian mixtures are insufficient to model complex multivariate data. Future research can focus on finding a solution for it by the usage of other distribution mixtures or, for instance, factor analysers. Secondly, the selection of the number of components or regularisation parameters is performed by an extensive grid search and regarding the modified BIC information criterion. There is no theoretical explanation that such criterion is optimal for the parameters selection or additionally if it is more suitable than other criteria. Finding a solution for these drawbacks would definitely bring a break-out, not only for this application but to the whole family of the penalised methods. Finally, possible research directions can focus on a broader study of the approach performance given various penalties, on finding a more stable solution to optimise the objective function or on adopting an additional method selecting a threshold for the further classification given the signal probabilities.

In line with the new physics search framework, the second proposed anomaly detection method – the Inverse Bagging introduced in Chapter 4 – requires a more in-depth study, especially considering its performance provided that a more sophisticated multivariate test is chosen instead of the used Hotteling's T^2-test. It is not clear if the discussed optimal parameter selection holds as well for the usage of different tests. In such setting, it would also be interesting to check the algorithm performance, also regarding for the proposed improvements. We find it also important, to introduce a method for statistical inference on the possibly found signal, i.e. the algorithm produces observation scores describing the data likelihood to be generated by an unknown signal process, but so far no methodology has been proposed to statistically infer about the scores significance, or more generally, about the signal existence.

Appendix A

Pseudo-code for the MESP approach

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| $> \nu$ AND $r <$n_iter)

   (a) Covariance matrices
   For (k in 1:K)
   $$\hat{\Sigma}_k^{(r)} = \hat{Q}_k^{(r)} \hat{D}_k^{(r)} \left(\hat{Q}_k^{(r)}\right)'. \tag{A.1}$$

   (b) Posterior probability
   For (i in 1:n, l in 1:K)
   $$\tau_{il}^{(r)} = \frac{\pi_l^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_l^{(r)}, \hat{\Sigma}_l^{(r)}\right)}{\sum_{k=1}^{K} \pi_k^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k^{(r)}, \hat{\Sigma}_k^{(r)}\right)} \tag{A.2}$$

   (c) Components proportions
   For (k in 1:K)
   $$\hat{\pi}_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(r)}. \tag{A.3}$$

   (d) MLE estimates
   For (k in 1:K)
   $$\tilde{\mu}_{kp}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} x_{ip}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{A.4}$$
   $$\tilde{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} * (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})'}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{A.5}$$

   (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
   $$\tilde{\Sigma}_k^{(r+1)} = \hat{Q}_k^{(r+1)} \tilde{D}_k^{(r+1)} \left(\hat{Q}_k^{(r+1)}\right)'. \tag{A.6}$$

   (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)

- $$\bar{D}_{k,pp}^{(r+1)} = \frac{-n\hat{\pi}_k^{(r+1)} + \sqrt{\left(n\hat{\pi}_k^{(r+1)}\right)^2 + 8\gamma_2 n \hat{\pi}_k^{(r+1)} \tilde{D}_{k,pp}^{(r)}}}{4\gamma_2}. \tag{A.7}$$

- 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. ϵ_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.

Appendix B

Pseudo-code for the PAD approach

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| $> \nu$ AND $r <$n_iter)

   (a) Covariance matrices
   For (k in K+1:K+L)
   $$\hat{\Sigma}_k^{(r)} = \hat{Q}_k^{(r)} \hat{D}_k^{(r)} \left(\hat{Q}_k^{(r)}\right)'. \tag{B.1}$$

   (b) Posterior probability
   For (i in 1:n, l in 1:K+L)
   $$\tau_{il}^{(r)} = \frac{\pi_l^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_l^{(r)}, \hat{\Sigma}_l^{(r)}\right)}{\sum_{k=1}^{K} \pi_k^{(r)} \phi\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k^{(r)}, \hat{\Sigma}_k^{(r)}\right)} \tag{B.2}$$

   (c) Components proportions
   For (k in K+1:K+L)
   $$\hat{\pi}_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(r)}. \tag{B.3}$$

   (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)
   $$\tilde{\mu}_{kp}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} x_{ip}}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{B.4}$$
   $$\tilde{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} * (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(r)})'}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{B.5}$$

   (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
   $$\tilde{\Sigma}_k^{(r+1)} = \hat{Q}_k^{(r+1)} \tilde{D}_k^{(r+1)} \left(\hat{Q}_k^{(r+1)}\right)'. \tag{B.6}$$

   (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

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

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

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

- $$\bar{D}_{k,pp}^{(r+1)} = \frac{-n\hat{\pi}_k^{(r+1)} + \sqrt{\left(n\hat{\pi}_k^{(r+1)}\right)^2 + 8\gamma_2 n\hat{\pi}_k^{(r+1)}\tilde{D}_{k,pp}^{(r)}}}{4\gamma_2}. \tag{B.7}$$

- 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. ϵ_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

$$||\hat{\mu}_{.p}^{(r)}|| = \sqrt{\sum_{k=1}^{K+L} \hat{\mu}_{kp}^{(r)^2}}$$

insted of used previously

$$||\hat{\mu}_{.p}^{(r)}|| = \sqrt{\sum_{k=1}^{K} \hat{\mu}_{kp}^{(r)^2}}$$

(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)

Bibliography

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

Alexandridis, R., Lin, S. and Irwin, M. (2004) Class discovery and classification of tumor samples using mixture modeling of gene expression data - a unified approach. *Bioinformatics* **20**(16), 2545–2552.

Alwall, J., Frederix, R., Frixione, S., Hirschi, V., Maltoni, F., Mattelaer, O., Shao, H. S., Stelzer, T., Torrielli, P. and Zaro, M. (2014) The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *Journal of High Energy Physics* **07**, 79.

AMVA4NewPhysics ITN (2017) Report of the performance of algorithms for data-driven background shape modeling. *https:userswww.pd.infn.it dorigowp4-d1.pdf* .

Aslan, B. and Zech, G. (2005) New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation* **75**(2), 109–119.

ATLAS Collaboration (2012) Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B* **716**(1), 1–29.

ATLAS Collaboration (2014a) A general search for new phenomena with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV. *ATLAS Conference notes* **2014-006**.

ATLAS Collaboration (2014b) 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**(5).

ATLAS Collaboration (2016) Dijet production in $\sqrt{s} = 7$ TeV pp collisions with large rapidity gaps at the ATLAS experiment. *Physics Letters B* **754**, 214–234.

ATLAS Collaboration (2017) A model independent general search for new phenomena with the ATLAS detector at $\sqrt{s} = 13$ TeV. *ATLAS Conference notes* **2017-001**.

Azzalini, A. and Scarpa, B. (2012) *Data analysis and data mining: An introduction*. Oxford University Press.

Baldi, P., Sadowski, P. and Whiteson, D. (2014) Searching for exotic particles in high-energy physics with deep learning. *Nature communications* **5**, 4308.

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics* pp. 803–821.

Barbier, R., Bérat, C., Besançon, M., Chemtob, M., Deandrea, A., Dudas, E., Fayet, P., Lavignac, S., Moreau, G. and Perez, E. (2005) R-parity violating supersymmetry. *Physical Review* **420**, 1–202.

Benkabou, S. E., Benabdeslem, K. and Canitia, B. (2018) Unsupervised outlier detection for time series by entropy and dynamic time warping. *Knowledge and Information Systems* **54**(2), 463–486.

Bentley, J. L. (1975) Multidimensional binary search trees used for associative searching. *Communications of the Association for Computing Machinery* **18**(9), 509–517.

Bergé, L., Bouveyron, C. and Girard, S. (2012) HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software* **46**(6), 1–29.

Bibby, J., Kent, J. and Mardia, K. (1979) *Multivariate analysis*. Academic Press, London.

Bolker, B. M. (2008) *Ecological models and data in R*. Princeton University Press.

Bonferroni, C. (1936) Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* **8**, 3–62.

Bouveyron, C. and Brunet-Saumard, C. (2014) Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis* **71**, 52–78.

Bouveyron, C., Girard, S. and Schmid, C. (2007) High-dimensional data clustering. *Computational Statistics & Data Analysis* **52**(1), 502–519.

Bühlmann, P. and Van De Geer, S. (2011) *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Buja, A. and Stuetzle, W. (2006) Observations on bagging. *Statistica Sinica* pp. 323–351.

CDF Collaboration (1995) Observation of top quark production in pp collisions with the Collider Detector at Fermilab. *Physical Review Letters* **74**(14), 2626.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern recognition* **28**(5), 781–793.

Chacón, J. E. and Duong, T. (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test* **19**(2), 375–398.

Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly detection: A survey. *Association for Computing Machinery computing surveys* **41**(3), 15.

Chen, C. (2012) New approach to identifying boosted hadronically decaying particles using jet substructure in its center-of-mass frame. *Physical Review D* **85**(3), 034007.

CMS Collaboration (2009) Particle-flow event reconstruction in CMS and performance for jets, taus and MET. Technical report, CMS-PAS-PFT-09-001.

CMS Collaboration (2010) First measurement of Bose-Einstein correlations in proton-proton collisions at $\sqrt{s} = 0.9$ and 2.36 TeV at the LHC. *Physical Review Letters* **105**(3), 032001.

CMS Collaboration (2011a) Model unspecific search for new physics in pp collisions at $\sqrt{s} = 7$ TeV. *CMS Physics Analysis Summaries* **EXO-10-021**.

CMS Collaboration (2011b) Measurement of Bose-Einstein correlations in pp collisions at $\sqrt{s} = 0.9$ and 7 TeV. *Journal of High Energy Physics* **2011**(5), 29.

CMS Collaboration (2012) Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B* **716**(1), 30–61.

CMS Collaboration (2016a) 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**(6), 317.

CMS Collaboration (2016b) The CMS trigger system. *arXiv preprint* **1609**(02366).

CMS Collaboration (2017) Model unspecific search for new physics, in pp collisions at $\sqrt{s} = 8\,\text{TeV}$. *CMS Physics Analysis Summaries* **EXO-14-016**.

Dall'Osso, M., de Castro Manzano, P., Dorigo, T., Finos, L., Kotkowski, G., Menardi, G. and Scarpa, B. (2017) Hemisphere mixing: A fully data-driven model of QCD multijet backgrounds for LHC searches. *Proceedings of the European Physical Society Conference on High Energy Physics* **314**(370).

Dasgupta, D. and Nino, F. (2000) A comparison of negative and positive selection algorithms in novel pattern detection. *2000 Institute of Electrical and Electronics Engineers International Conference on systems, man, and cybernetics* **1**, 125–130.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B* pp. 1–38.

Dickerson, J. E. and Dickerson, J. A. (2000) Fuzzy network profiling for intrusion detection. *Proceedings of the 19th international conference on Fuzzy Information Processing Society* pp. 301–306.

DONUT Collaboration (2001) Observation of tau neutrino interactions. *Physics Letters B* **504**(3), 218–224.

Dotto, F., Farcomeni, A., García-Escudero, L. A. and Mayo-Iscar, A. (2018) A reweighting approach to robust clustering. *Statistics and Computing* **28**(2), 477–493.

Duong, T. Goud, B. and Schauer, K. (2012) Closed-form density-based framework for automatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences* **109**(22), 8382–8387.

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

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. *et al.* (2004) Least angle regression. *The Annals of statistics* **32**(2), 407–499.

Egan, J. P. (1975) *Signal detection theory and ROC analysis*. Academic Press.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**(456), 1348–1360.

de Favereau, J., Delaere, C., Demin, P., Giammanco, A., Lemaître, V., Mertens, A. and Selvaggi, M. (2014) DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics* **02**, 057.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97**(458), 611–631.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The elements of statistical learning*. Volume 1. Springer series in statistics New York.

Friedman, J. H. and Rafsky, L. C. (1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* pp. 697–717.

Frixione, S. and Webber, B. R. (2002) Matching next leading order QCD computations and parton shower simulations. *Journal of High Energy Physics* **2002**(06), 029.

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

Grossman, Y. and Rakshit, S. (2004) Neutrino masses in R-parity violating supersymmetric models. *Physical Review D* **69**(9), 093002.

He, H. and Garcia, E. A. (2008) Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering* (9), 1263–1284.

Heard, N. A. and Rubin-Delanchy, P. (2018) Choosing between methods of combining-values. *Biometrika* **105**(1), 239–246.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67.

Hotelling, H. (1931) The economics of exhaustible resources. *Journal of political Economy* **39**(2), 137–175.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of classification* **2**(1), 193–218.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I. and Soderstrom, T. (2018) Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. *arXiv preprint* **1802**(04431).

Izenman, A. J. (2008) *Modern multivariate statistical techniques: regression, classification and manifold learning*. Springer.

Justel, A., Peña, D. and Zamar, R. (1997) A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters* **35**(3), 251–259.

Kabaila, P. (2005) On the coverage probability of confidence intervals in regression after variable selection. *Australian & New Zealand Journal of Statistics* **47**(4), 549–562.

Karunamuni, R. J. and Alberts, T. (2005) On boundary correction in kernel density estimation. *Statistical Methodology* **2**(3), 191–212.

Khan, A. and Rayner, G. D. (2003) Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics & Decision Sciences* **7**(4), 187–206.

Kramer, G. and Spiesberger, H. (2018) λ_b^0-baryon production in pp collisions in the general-mass variable-flavour-number scheme and comparison with CMS and LHCb data. *arXiv preprint arXiv:1803.11103* .

Kuusela, M., Vatanen, T., Malmi, E., Raiko, T., Aaltonen, T. and Nagai, Y. (2012) Semi-supervised anomaly detection - towards model-independent searches of new physics. *Journal of Physics: Conference Series* **368**(1), 012032.

Markou, M. and Singh, S. (2003) Novelty detection: a review - part 1: statistical approaches. *Signal processing* **83**(12), 2481–2497.

McNicholas, P. D. (2016) Model-based clustering. *Journal of Classification* **33**(3), 331–373.

Menardi, G. and Torelli, N. (2014) Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery* **28**(1), 92–122.

Noble, C. C. and Cook, D. J. (2003) Graph-based anomaly detection. *Proceedings of the ninth Association for Computing Machinery international conference on Knowledge discovery and data mining* pp. 631–636.

Pan, W. and Shen, X. (2007) Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8**(5), 1145–1164.

Pan, W., Shen, X., Jiang, A. and Hebbel, R. P. (2006) Semi-supervised learning via penalized mixture model with application to microarray sample classification. *Bioinformatics* **22**(19), 2388–2395.

Pesarin, F. and Salmaso, L. (2010) *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Pierson, E. and Yau, C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology* **16**(1), 241.

Pimentel, M. A., Clifton, D. A., Clifton, L. and Tarassenko, L. (2014) A review of novelty detection. *Signal Processing* **99**, 215–249.

Popov, A. (2011) Searches for new physics at Tevatron: Most recent results. *Physics of Atomic Nuclei* **74**(3), 477–486.

R Core Team (2017) R: A language and environment for statistical computing. *R Foundation for Statistical Computing* .

Razali, N. M., Wah, Y. B. *et al.* (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics* **2**(1), 21–33.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C. (2001) Estimating the support of a high-dimensional distribution. *Neural computation* **13**(7), 1443–1471.

Schwarz, G. (1978) Estimating the dimensions of a model. *Annals of Statistics* **6**, 461–464.

Scott, D. W. (2015) *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8**(1), 205–233.

Sheskin, D. J. (2003) *Handbook of parametric and nonparametric statistical procedures*. Chemical Rubber Company Press.

Silverman, B. W. (1986) *Density estimation for statistics and data analysis*. Volume 26. Chemical Rubber Company Press.

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

Sjőstrand, T., Mrenna, S. and Skands, P. Z. (2006) PYTHIA 6.4 Physics and Manual. *Journal of High Energy Physics* **05**, 026.

Tan, P. N., Steinbach, M. and Kumar, V. (2005) Association analysis: basic concepts and algorithms. *Introduction to Data mining* pp. 327–414.

Thomson, M. (2013) *Modern particle physics*. Cambridge University Press.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B* pp. 267–288.

Tinsley, H. and Brown, S. (2000) *Handbook of applied multivariate statistics and mathematical modeling*. Academic Press.

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

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

Vischia, P. and Dorigo, T. (2017) The Inverse Bagging algorithm: Anomaly detection by inverse bootstrap aggregating. *European Physical Journal Web of Conferences* **137**, 11009.

Wand, M. and Jones, M. (1995) Kernel smoothing. *Chapman & Hall* **1**(2), 6.

Xie, B. (2008) *Variable selection in penalized model-based clustering*. University of Minnesota Press.

Xie, B., Pan, W. and Shen, X. (2008) Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics* **64**(3), 921–930.

Xu, G., Cao, Z., Hu, B.-G. and Principe, J. C. (2017) Robust support vector machines based on the rescaled hinge loss function. *Pattern Recognition* **63**, 139–148.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68**(1), 49–67.

Zhai, J., Zhang, S. and Wang, C. (2017) The classification of imbalanced large data sets based on mapreduce and ensemble of elm classifiers. *International Journal of Machine Learning and Cybernetics* **8**(3), 1009–1017.

Grzegorz Kotkowski

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +48 533 217 874
e-mail: grzegorzmichal.kotkowski@studenti.unipd.it

Current Position

Since October 2015; (expected completion: March 2019)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Advanced statistical methods for data analysis in particle physics
Supervisor: Prof. Giovanna Menardi Co-supervisor: Prof. Bruno Scarpa, Prof. Livio Finos

Research interests

- Anomaly detection
- Statistical/machine learning
- Application of statistical methods to physics

Education

10.2013 – 07.2015
Master degree in Mathematics with specialization in Statistics.
Wrocław University of Technology, Faculty of Fundamental Problems of Technology, Statistical Department
Title of dissertation: "Random Matrices Theory in Statistical Multidimensional Data Analysis"
Supervisor: Prof. M. Bogdan
Final mark: 5 (in scale from 2 to 5.5)

10.2010 – 07.2013
Bachelor degree in Mathematics.
Wrocław University of Technology, Faculty of Fundamental Problems of Technology, Statistical Department
Final mark: 4.5 (in a scale from 2 to 5.5)

10.2011 – 08.2015
Bachelor degree in Physics.
University of Wrocław, Faculty of Theoretical Physics
Final mark: 5 (the highest possible)

Visiting periods

02.2018 – 03.2018
Universidad de Oviedo, Spain.
Supervisors: prof. Pietro Vischia

05.2017 – 06.2017
Université Blaise-Pascal, Clermont-Ferrand, France.
Supervisors: prof. Julien Donini

10.2016 – 12.2016
CERN, Geneva, Switzerland.
Supervisors: prof. Tancredi Carli and prof. Tommaso Dorigo

Work experience

11.2017 – 01.2018
SDG Group, Milano, Italy.
PhD secondment in private sector as a trainee

04.2015 – 00.2015
Datarino (Hicron Group), Wrocław, Poland.
Data analyst

Awards and Scholarship

10.2015 – 10.2018
Marie Sklodowska-Curie scholarship for a PhD position financed by the H2020 program of the European Commission under the AMVA4NewPhysics project.

Computer skills

- Advanced in R and SQL
- Intermediate Python and C++
- Basics in Apache Spark, Java, Matlab,
- Familiar with Jira, SVN, Microsoft package, Linux environment (Vim, Sed, Awk)

Language skills

- Polish: native
- English: fluent
- Italian: basic
- German: basic

Publications

Working papers

Dall'Osso, M and de Castro Manzano, P and Dorigo, T and Finos, L and Kotkowski, G and Menardi, G and Scarpa, B (2017). Hemisphere Mixing: a fully data-driven model of QCD multijet backgrounds for LHC searches. *Proceedings of the European Physical Society Conference on High Energy Physics* 314(370).

Conference presentations

Kotkowski, G. and Jiménez, F. (2017). Anomaly detection for generic searches. *ATLAS Machine Learning Workshop*, CERN, Geneva, Switzerland, 08-09.06.2017.

Kotkowski, G. (2018). Model independent searches for new physics via parametric anomaly detection. *XIIIth Quark Confinement and the Hadron Spectrum Conference*, Maynooth University, Dublin, Ireland, 01-03.08.2018.

References

Available on request