Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Inference for Clustering and Anomaly Detection
Presented by: Purvasha Chakravarti

Accepted by: Department of Statistics & Data Science

Readers:

Larry Wasserman, Advisor

Sivaraman Balakrishnan

Mikael Kuusela

Andrew B. Nobel

Rebecca Nugent

Alessandro Rinaldo

Approved by the Committee on Graduate Degrees:

Richard Scheines, Dean Date

Carnegie Mellon University

Inference for Clustering and Anomaly Detection

A Dissertation Submitted to the Graduate School

in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy

in

Statistics

by

Purvasha Chakravarti

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

May 2020

To my family who give me the confidence to fly...

Acknowledgements

First, I would like to thank Larry Wasserman, for being my advisor, providing me with invaluable guidance over the years, and for being the best role model one could ever hope for. I could not have completed this dissertation without his insights, advice and various anecdotes. The numerous hours of discussions on the board, on paper and on Skype with him have taught me that creativity, diligence and patience can solve any problem that one faces in research. His enthusiasm for getting to the bottom of things has made working on this thesis all the more exciting and fun.

My many thanks to my collaborators - Siva Balakrishnan and Mikael Kuusela, who have been invaluable to this thesis. The thesis was shaped by the many hours Siva spent proof-reading the numerous proofs and correcting them, and by Mikael's amazing insights and guidance. I have learned a lot about clarity in research, writing and presentation through my fruitful interactions with them. Their great ideas and helpful feedback helped guide me through the Ph.D. process.

Special thanks to my other committee members, Andrew Nobel, Rebecca Nugent, and Alessandro Rinaldo, for their invaluable help, comments and guidance through my PhD. Particularly, I would like to thank Rebecca for going out of her way to help me present my work, get feedback on it at various stages, as well as help me network with other academics.

The CMU Statistics & Data Science Department has one of the most friendly and helpful faculty, who have made my time at CMU all the more productive. I would like to thank all the faculty members including Bill Eddy, Chris Genovese, Ann Lee, Brian Junker, Aaditya Ramdas, Peter Freeman, Ryan Tibshirani, Valerie Ventura, Joel Greenhouse and Robin Mejia, for their advice, guidance, insightful conversations and encouragement. I would also like to thank the incredibly helpful staff in the Department of Statistics & Data Science. In particular, a shout out to Margie Smykla, Beth Dongilli, Christopher Peter Makris, Jess Paschke, and Carl Skipper.

Thanks to my colleagues and friends in the department, who make the department one of the best and friendliest! For the wonderful memories, experiences and advice, thank you Amanda Luby, Lee Richardson, Francesca Matano, Niccolo Dalmasso, Maria Cuellar, Jining Qin, Nil-Jana Akpinar, Natalia Lombardi,

Octavio Mesner, Shamindra Shrotriya, Jackie Mauro, Zach Branson, Manjari Das, Maria Jahja, Theresa Gebert, Robin Dunn, Pratik Patil, Ilmun Kim, Jaehyeok Shin and Jacqueline Liu.

Thank you Kayla Frisoli, for being one of the most enthusiastic friends I have had, for organizing all the events that kept us sane and together. Thank you Mikaela Meyer, for being such a great support and always injecting cheer into my day. Thank you Taylor Pospisil, Benjamin LeRoy, Ciaran Evans and Neil Spencer for hosting the many board-game nights that I used to look forward to, on a regular basis.

Thank you Brendan McVeigh, Collin Politsch, Kevin Lin, Natalie Klein and Yotam Hechtlinger for being the best office-mates and wonderful friends; for the brain-storming sessions on the white board, for the late-night homework solving, for the Karaoke nights, the late-night hangouts and so much more.

Thank you Shannon Gallagher for being my movie buddy, my last minute practice-talk-listener, my proof-reader, my strongest supporter and most importantly my friend no matter the hour or the day.

The majority of the credit for making Pittsburgh my home away from home goes to the Indian Graduate Student Association, in particular, to the members of "The Peace Band". You have filled my life with music, love and cheer. Out of this group were born several sub-groups that I would also like to thank individually. First, special thanks go to "The Club" that provided me with extraordinary inter-disciplinary feedback on my research as well as my presentation skills. Second, I would like to thank "Spring Gamez" and "Burgh Gamez" for becoming my family and supporting me in every way possible for the past six years.

Honorable mention goes to Saurabh Kadekodi, who went above and beyond, by helping me with tech support, indulging in constructive arguments at all times of the day and providing me with lunches and dinners in times of need. Special thanks also to Arushi Vyas and Dipan Pal. Arushi, for bringing Zumba into my life and always being there for me. Dipan, for being an amazing friend, for the in-depth discussions and philosophies, and for the never-ending support.

Throughout my Ph.D. at CMU, my friends from my undergraduate years, Surya Teja, Anushree Barjatya, Avi Marathe and Srilikhitha Patel, never let me feel their absence. They have been a support through every step and without their help and encouragement, this thesis would just not have been possible.

Finally to my family, thank you Anwesha Chakravarti, for being the best sister in the world and always having my back in anything I do. I look forward to multiple collaborations in the future! Thank you mom and dad, Madhumita and Bhudeb Chakravarti, for supporting me from my childhood to pursue Mathematics and then Statistics, and for always encouraging me to think like a researcher. You have both been, and always will be, my inspiration to pursue my dreams. To my grandmothers, Mina Chakravarti and Bela Das, thank you for always believing that I can reach the stars if only I try. I will always love you!

CMU has also given me my best friend and partner Siddharth Singh, who never stops believing in me, supports me unconditionally, provides academic insight whenever I need it, and brings general cheer and joy into my day to day life. Thank you so much for enriching my Ph.D. life and making it an adventure! Looking forward to many more such adventures to come!

"It is a capital mistake to theorize before one has data.
Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

$\sim$ Sherlock Holmes (Arthur Conan Doyle), A Scandal in Bohemia

Abstract

This thesis focuses on developing scalable clustering and anomaly detection methods, with realistic assumptions and theoretically-sound guarantees, for analyzing high-dimensional data. It also studies the theory behind the performance of the proposed methods. Specifically, this thesis takes an inferential approach to searching for evidence that indicates the presence of two or more collections of data, with different distributions, in a single data set. It addresses two fundamental questions relating to this: (a) How can we perform clustering that results in statistically significant clusters? (b) In high energy physics, how can we detect new signals in experimental data, that are not explained by known physics models, without assuming a model for the new signal?

In order to answer the first question, we consider clustering based on significance tests for Gaussian Mixture Models (GMMs). Our starting point is the SigClust method developed by Liu et al. (2008), which introduces a test based on the k-means objective (with k = 2) to decide whether the data should be split into two clusters. When applied recursively, this test yields a method for hierarchical clustering that is equipped with significance guarantees. We study the limiting distribution and power of this approach in some examples and show that there are large regions of the parameter space where the power is low. We then introduce a new test based on the idea of *relative fit*. Unlike prior work, we test for whether a mixture of Gaussians provides a better fit relative to a single Gaussian, without assuming that either model is correct. The proposed test has a simple critical value and provides provable error control. We then develop several different versions of the test, one of which provides exact type I error control without requiring any asymptotic approximations. We show how the test can be applied recursively to obtain a hierarchical clustering of the data with significance guarantees. We also construct a sequential, non-hierarchical version of the approach that can additionally be used for model selection. We conclude with an extensive simulation study and a cluster analysis of a gene expression dataset.

To answer the second question, we search for new signals that appear as deviations from known Standard Model physics in experimental particle physics data. To do this, we determine whether there is any significant difference between the distribution of background samples alone (generated from an assumed Monte Carlo model according to the Standard Model) and the distribution of the actual experimental observations, which

could be a mixture of background and signal samples. Traditionally, model-dependent methods are used to train a supervised classifier to detect hypothesized signals expected under models of new physics. In this thesis, we propose a model-independent method, that does not make any assumptions about the signal and uses a semi-supervised classifier to detect the presence of a signal in the experimental data. We use a test based on the likelihood ratio test statistic as well as one based on the area under the curve (AUC) statistic. The second test is based on the assumption that if the experimental data does not contain any signal then the classifier should find the experimental data indistinguishable from the background data. Additionally, we explore active subspace methods to interpret the proposed semi-supervised classifier tests in order to understand properties of the signal detected in the experimental data. We conclude by studying the performance of the methods on a data set related to the search for the Higgs Boson provided by the ATLAS experiment at the Large Hadron Collider (LHC) at CERN.

Contents

List of Tables

List of Figures

Part I

Introduction

Introduction

Clustering is the intuitive action of partitioning a set of objects into a collection of clusters so that objects in the same cluster are similar, while objects in different clusters are dissimilar. This notion arises very naturally in many fields, whenever there is a heterogeneous set of objects that need to be grouped based on some underlying similarity measure. Clustering is used as a data analysis tool across very diverse disciplines such as image segmentation, marketing, genetics and bioinformatics. The wide use of clustering is perhaps unsurprising as it can be used during different stages of the data analysis process, starting from exploratory data analysis to discovering new sub-classes, sub-types or partitions in the data.

Despite its popularity, a fundamental question that still requires answering is: How many clusters are there in the data? Furthermore, when should a clustering even be applied and when is a particular cluster statistically significant? These are the questions that we want to answer in this dissertation.

Under some circumstances, clustering is just used as an exploratory data analysis tool to understand the spread of the data or to visualize high-dimensional data. For this task, clustering provides useful information irrespective of whether any inherent cluster structure is present in the data or not. But in some cases, as illustrated below, clustering is used to uncover real groupings inherent in the data. In this case, if the data is split into more clusters than the real groupings in the data, the resulting clusters could be arbitrary and consequently potentially misleading (Adolfsson et al., 2019).

For example in medical sciences, especially in bioinformatics, scientists look for actual groupings in the patients to develop different new treatments for the different groups. An example of this is Glioblastoma multiforme (GBM), which is the most common form of malignant brain cancer in adults. Patients with GBM have a uniformly poor prognosis, with a median survival of one year. This makes it particularly important to understand the prognosis of GBM better. One of the things that has been found promising for prognosis or prediction of response to therapy, is to use clustering methods on gene sequences of patients to find molecular subclasses of GBM (Verhaak et al., 2010).

A similar need arises for different kinds of breast cancers as well. Genomic studies have established four major breast cancer intrinsic subtypes (luminal A, luminal B, HER2-enriched, basal-like) and a normal breast-like group that show significant differences in incidence, survival and response to therapy (Prat et al.,

2010). This diversity calls for the need to have different tumor classes that are clinically useful with respect to prognosis or prediction. For example, Metaplastic breast cancers (MBC) are treated in the same fashion as basal-like or triple receptor-negative ductal cancers even though MBCs are usually chemoresistant. Therefore over the last few years as gene expression studies have evolved, further subclassification of breast tumors into new molecular entities has occurred. Clustering methods have been extensively used in this front. For example, (a) new subclasses of Metaplastic breast cancers (MBC) (Hennessy et al., 2009) have been found using clustering; (b) subtypes luminal A (LumA), luminal B (LumB), HER2-enriched, basal-like, and normal-like of breast cancer have been extensively studied by microarray and hierarchical clustering analysis (Parker et al., 2009); and (c) stable subtypes as well as subtype-specific molecular targets have been identified for triple-negative breast cancers (Burstein et al., 2015) using clustering techniques.

These examples demonstrate the need for significant clusters that identify the *real* groupings inherent in the data. In this direction, Liu et al. (2008) propose an approach called SigClust. They define a cluster as data coming from a single Gaussian distribution and formulate the problem of assessing statistical significance of clustering as a testing procedure. Their test statistic is based on the k-means cluster index, with $k = 2$. If the test rejects, then the data is split into two clusters. This test can be applied recursively leading to a top-down hierarchical clustering (Kimes et al., 2017). This approach attempts to distinguish clusters which are actually present in the data from the natural sampling variability. The method is appealing because it is simple and because, as we further elaborate in Chapter 3, it provides certain rigorous error control guarantees. This procedure has already been extensively used in bioinformatics for finding cancer sub-types by analyzing gene-sequence data sets (Parker et al., 2009; Hennessy et al., 2009; Prat et al., 2010; Burstein et al., 2015).

However SigClust has two disadvantages. First, it assumes that a cluster is generated from a single Gaussian and second, as we demonstrate in Chapter 3 of this dissertation, there are large regions of the parameter space where the method has poor statistical power. To address these two issues with SigClust, we propose a different approach. We test whether a single multivariate Gaussian is closer to the true distribution than a mixture of two Gaussians without assuming that either model is true. We call this a test of *relative fit* or Rift (Relative Information Fit Test). The result is a test with a simple limiting distribution, that makes no assumptions about the true distribution of the clusters. Following Kimes et al. (2017), we also apply the test recursively to obtain a hierarchical clustering of the data with significance guarantees as well as a sequential, non-hierarchical version, of the approach.

Similar to the life sciences, the physical sciences also face a relatable fundamental problem. The problem is one of detecting new physics phenomena that are not explained by the Standard Model, which describes our current understanding of fundamental particles and how they interact with each other. Here the task of searching for signals that behave anomalously to the known background processes (those that are explained by the Standard Model) poses as an anomaly detection problem. However, in this specific case, the anomalous

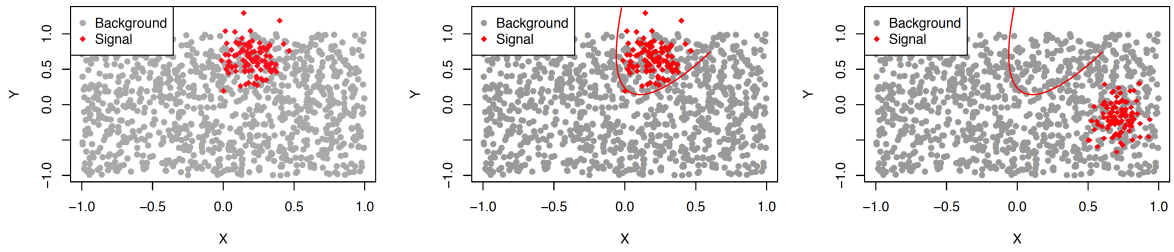


Figure 1.1: Example of detecting the signal (red) from the background (grey) as a collective anomaly detection process. (a) Signal as an anomalous cluster on top of the background. (b) The boundary of the classifier when trained on signal generated from the assumed signal model. (c) When there is a systematic error in the training signal data, the test completely misses the actual signal data.

signals lie in the domain of the background. Hence each signal data point individually looks like it could have been generated from the background. So instead of searching for anomalous data points individually, we need to search for their occurrence together as a collection that seems anomalous. The detection of such anomalies is called *collective anomaly detection* (Chandola et al., 2009). This is in a similar spirit to the previous problem of significant clustering, because in this second problem, we are searching for a significant cluster of signals (an anomalous collection) over the background.

The objective can be described as determining if there is any significant difference between the distribution of background samples alone (generated from an assumed Monte Carlo model) and the distribution of the actual experimental observations, which could be a mixture of background and signal samples. A simple example of how the signal might look with respect to the background can be found in Figure 1.1(a). An approach to detect this signal can be designed by using a version of RIFT where we first fit the background using a multivariate Gaussian mixture model. We then fit a mixture of this background model and a number of additional Gaussians to the experimental data. The test of significance can then be performed by using a version of RIFT where we test if the experimental mixture model fits the experimental data significantly better than the background mixture model. The procedure is similar to what was proposed by Kuusela et al. (2012) and Vatanen et al. (2012). They assume the mixture models to be true and use a likelihood ratio test to compare the fits. On the other hand, we do not assume the models to be true. In either case, since the signal strength in this setup can be really small and mixture model fits in higher dimensions perform poorly, we expect both of these methods to have very little power in detecting the signal.

Typically the search for high-dimensional new physics signals is performed in a model-dependent fashion using supervised classifiers. Such an approach assumes a signal model for the new signal that is being sought out and generates training signal samples using a Monte Carlo (MC) event generator. It then trains a supervised classifier on the generated background and signal samples and performs a test structured as a likelihood ratio test (Williams, 2010; Cowan et al., 2011; ATLAS Collaboration and CMS Collaboration,

2011). There are two main drawbacks of this type of approach. Firstly, this approach cannot be used to search for signals that we are not specifically looking for or have not considered yet. Secondly, systematic errors can be very influential on supervised classifiers and so any error or imprecision in the signal model will adversely affect the method. Figure 1.1 illustrates the problem more clearly. If a classifier is trained on training signal data as generated in Figure 1.1(b), it gives the classification boundary as shown. But what if the signal data actually looks like Figure 1.1(c)? Then the classifier ends up misclassifying the signal as background. So an algorithm trained on a wrong signal model might completely miss the actual signal.

In contrast, in Part III of this thesis, we propose tests to search for new physics signals in a model-independent fashion, without assuming any model for the signal (Kuusela et al., 2012; Vatanen et al., 2012; Casa and Menardi, 2018; Aaboud et al., 2019). Specifically, in the experimental data (as collected from particle detectors) we search for any signal that deviates from the background process (as explained by the Standard Model). We use a semi-supervised approach that trains a classifier to differentiate the background data from the experimental data. We then propose two different tests to detect the presence of signal in the experimental data. The first test is based on a likelihood ratio test statistic that is estimated using the classifier output. The second test is based on the performance of the classifier measured using the area under the curve test statistic. Both the tests are based on the argument that in the absence of signal events in the experimental data, a classifier should not be able to differentiate the experimental samples from the background samples. This approach is better than the mixture modelling methods because classifiers work better in high-dimensional spaces.

We further propose using active subspace methods (Constantine, 2015) to identify the characteristics of variables and their dependencies on each other, that differ between the background and the experimental data affecting the outcome of the classifier. This can be used to characterize the signal region in the experimental data.

In summary, the questions we aim to answer in this thesis are:

1. **Clustering:** How can we perform clustering that results in significant clusters?
2. **Anomaly detection:** How can we detect collective anomalies in a model-independent semi-supervised fashion, in a high-dimensional space, when the anomalies are really small in proportion?

In the following sections, we provide a road-map of the entire thesis, highlighting the chapter-wise contributions made in the thesis. We then introduce notations that we use throughout this thesis.

1.1 Contributions and the Road-Map of the Thesis

This thesis is organized into four main parts. Part I is the introduction of the thesis. In Part II, we introduce clustering algorithms that come equipped with significance guarantees. In Part III we introduce

the model-independent methods to detect new physics signals in the experimental data using tests based on semi-supervised classifiers. Finally, in Part IV we draw the final conclusions from all of our studies and outline our vision for future work. Table 1.1 highlights the main contributions that are presented in this thesis.

Table 1.1: Contributions of the thesis

	Chapter	Contribution	Description
Clustering (Part I)	3	Power of SigClust	Asymptotic power derivation for SigClust to detect two clusters that shows the existence of regions where the power is low.
	4	RIFT	A test based on relative fit to detect two clusters.
		M-RIFT	A robust version of RIFT with exact type I error control.
	5	Hierarchical RIFT	Top-down and bottom-up hierarchical versions of RIFT, to detect more than two clusters.
		Sequential RIFT	Sequential RIFT to detect more than two clusters.
Anomaly Detection (Part II)	8	Semi-supervised LRT	Model-independent test based on likelihood-ratio test statistic to detect signal events.
		Semi-supervised AUC test	Model-independent test based on area under the curve (AUC) test statistic to detect signal events.
		Signal region characterization	Identification of the active subspace of the classifier which detects the signal.
	9	Higgs boson detection	Experiments that compare the performance of model-dependent methods and model-independent methods in detecting the Higgs boson particle.

1.1.1 Contributions in Part I: Clustering

In Chapter 3, we review the SigClust procedure (Liu et al., 2008) and we derive its power in some cases. We show that SigClust can have poor power against certain alternative hypotheses. This chapter also has results on the geometric properties of k-means clustering in a special case, which is a prelude to finding the power. In Chapter 4, we describe our new procedure and its different versions. In Chapter 5, we demonstrate the use of our new tests in a hierarchical framework as well as a sequential framework which can be used as a model selection tool for the GMM. We compare the proposed algorithms with SigClust and other methods

using some simulations studies in Section 6.1 and an analysis of a gene expression data set in Section 6.2 within Chapter 6. We defer the technical details of most proofs to the Appendix.

1.1.2 Contributions in Part II: Anomaly detection

In Chapter 8 we first mathematically set up the problem of detecting the new physics signals in the experimental data observed by the particle detectors. We then describe comparable supervised methods in Section 8.2. The proposed model-independent semi-supervised methods are introduced in Section 8.3. We introduce the two tests, one based on the likelihood ratio test statistic and the other based on the area under the curve (AUC) test statistic. In Section 8.5 we describe active subspace methods to understand the subspace most strongly affecting the classifier, leading to an understanding of the signal region. Finally in Chapter 9, we demonstrate the performance of the proposed methods and compare them to the supervised approaches as well as nearest neighbor two sample tests introduced in Schilling (1986) and Henze (1988).

1.2 Notation

For Part II of this dissertation we assume that the dimension d is fixed and the sample size n is increasing. In contrast, Liu et al. (2008) and Kimes et al. (2017) focus on the large d, fixed n case which requires dealing with challenging issues such as estimating the covariance matrix in high dimensions (see also Vogt and Schmid (2017)). However, because of the challenges of high dimensional estimation, these prior works only establish results about power in very specific cases. In contrast, we provide a more detailed understanding of the power in the fixed-d case.

Throughout this dissertation we use $\|\cdot\|$ to denote the Euclidean norm, i.e., for $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\| := \sqrt{\sum_{i=1}^d x_i^2}$. We use the symbols $\xrightarrow{p}$ and $\rightsquigarrow$ to denote the standard stochastic convergence concepts of convergence in probability and in distribution, respectively.

Part II

Inference for Clustering: Gaussian Mixture Clustering Using Relative Tests of Fit

Chapter 2

Introduction to Significant Clustering

Gaussian mixture models (GMMs) are a commonly used tool for clustering. A major challenge in using GMMs for clustering is in adequately answering inferential questions regarding the number of mixture components or the number of clusters to use in data analysis. This task typically requires hypothesis testing or model selection. However, deriving rigorous tests for GMMs is notoriously difficult since the usual regularity conditions fail for mixture models (Ghosh and Sen, 1984; Dacunha-Castelle et al., 1999; Gassiat, 2002; McLachlan and Peel, 2004; McLachlan and Rathnayake, 2014; Chen, 2017; Gu et al., 2017).

In this direction, Liu et al. (2008) proposed an approach called SigClust. Their method starts by fitting a multivariate Gaussian to the data. Then a significance test based on k-means clustering, with $k = 2$, is applied. If the test rejects, then the data is split into two clusters. This test can be applied recursively leading to a top-down hierarchical clustering (Kimes et al., 2017). This approach roughly attempts to distinguish clusters which are actually present in the data from the natural sampling variability. The method is appealing because it is simple and because, as we further elaborate on in the sequel, it provides certain rigorous error control guarantees.

In this dissertation we study the power of SigClust and show that there are large regions of the parameter space where the method has poor power. A natural way to fix this would be to use another statistic designed to distinguish "a Gaussian" versus "a mixture of two Gaussians" such as the generalized likelihood ratio test. However, such an approach has two problems: first, as mentioned above, mixture models are irregular and the limiting distribution of the likelihood ratio test (and other familiar tests) is intractable. Second, such tests assume that one of the models (Gaussian or mixture of Gaussians) is correct. Instead from a practical standpoint, for the purposes of clustering, we only regard these models as approximations.

So we consider a different approach. We test whether one model is closer to the true distribution than the other without assuming either model is true. We call this a test of *relative fit*. Our test is based on data splitting. Half the data are used to fit the models and the other half are used to construct the test. The

result is a test with a simple limiting distribution which makes it easy to determine an appropriate cutoff for it. In fact, we provide several versions of the test. One version provides exact type I error control without requiring any asymptotic approximations.

Following Kimes et al. (2017), we also apply the test recursively to obtain a hierarchical clustering of the data with significance guarantees. We develop a bottom-up version of mixture clustering which can be regarded as a linkage clustering procedure where we first over-fit a mixture and subsequently combine elements of the mixture. We also construct a sequential, non-hierarchical version, of the approach. We call our procedure RIFT (Relative Information Fit Test).

Throughout this dissertation we assume that the dimension d is fixed and the sample size n is increasing. In contrast, Liu et al. (2008) and Kimes et al. (2017) focus on the large d, fixed n case which requires dealing with challenging issues such as estimating the covariance matrix in high dimensions (see also Vogt and Schmid (2017)). However, because of the challenges of high dimensional estimation, these prior works only establish results about power in very specific cases. In contrast, we provide a more detailed understanding of the power in the fixed-d case.

2.1 Overview of the Related Literature

Estimating the number of clusters has been approached in many ways (Bock, 1985; Milligan and Cooper, 1985; McLachlan and Peel, 2004). A common approach is to find the optimal number of clusters by optimizing a criterion function, examples of which are the Hartigan index (Hartigan, 1975), the silhouette statistic (Rousseeuw, 1987) or the gap statistic (Tibshirani et al., 2001).

Another approach to estimating the number of clusters is to assess the statistical significance of the clusters. McShane et al. (2002) proposed a method to calculate p-values by assuming that the cluster structure lies in the first three principal components of the data. Tibshirani and Walther (2005) use resampling techniques to quantify the prediction strength of different clusters and Suzuki and Shimodaira (2006) assess the significance of hierarchical clustering using bootstrapping procedures. More recently, Maitra et al. (2012) proposed a distribution-free bootstrap procedure which assumes that the data in a cluster is sampled from a spherically symmetric, compact and uni-modal distribution. Engelman and Hartigan (1969) considered the maximal F-ratio that compares between group dispersions with within group dispersions. Lee (1979) proposed a subsequent multivariate version and a robust version was recently proposed by Garcia-Escudero et al. (2009). Another example is a statistical test proposed by Vogt and Schmid (2017). They develop a fairly general significance test but it relies on assuming that the number of covariates tends to infinity and that the clusters are, in a certain sense, well-separated (i.e. can be consistently estimated as the number of features increases).

Alternatively, and closer to our approach, Gaussian mixture models can be used for cluster analysis. See for instance, the works Fraley and Raftery (2002); McLachlan and Peel (2004); McLachlan and Rathnayake (2014) for overviews. There is much prior work for testing the order of a Gaussian mixture. For example, the works Ghosh and Sen (1984); Hartigan (1985) used the likelihood ratio test with the null hypothesis that the order is one. Hartigan (1985) explored the impact of non-regularity of the mixture models and Ghosh and Sen (1984) used a separation condition in order to find the asymptotic distribution of the likelihood ratio test statistic.

Since finite normal mixture models are irregular and the limiting distribution of the likelihood ratio test statistic is difficult to derive, deriving a general theory for testing the order of a mixture is hard. Instead most of the algorithms test for homogeneity in the data. The works Charnigo and Sun (2004); Liu and Shao (2004); Chen et al. (2009) among others, are examples of this approach. More recently, Li and Chen (2010) and Chen et al. (2012) constructed a new likelihood-based expectation-maximization (EM) test for the order of finite mixture models that uses a penalty function on the variance to obtain a bounded penalized likelihood. Further developments can be found in the works Dacunha-Castelle et al. (1999); Gassiat (2002); Chen (2017); Gu et al. (2017). Our approach differs in three ways: we use a test that avoids the irregularities, it avoids assuming that the mixture model is correct and it is valid for multivariate mixtures. We only treat the mixture model as an approximate working model.

Liu et al. (2008) proposed a Monte Carlo based algorithm (SigClust) that defines a cluster as data generated from a single multivariate Gaussian distribution. The distribution of the test statistic under the null hypothesis SigClust depends on the eigenvalues of the null covariance matrix. Huang et al. (2015) proposed a soft-thresholding method that provides an estimate of these eigenvalues, and this soft-thresholding method leads to a modified version of SigClust that is better suited to high-dimensional problems.

2.2 Organization of Part I

In Chapter 3 we review the SigClust procedure and we derive its power in some cases. We show that SigClust can have poor power against certain alternatives. This chapter also has results on the geometric properties of k-means clustering in a special case, which is a prelude to finding the power. In Chapter 4 we describe our new procedure. In Chapter 5 we show how to use our new tests in a hierarchical framework as well as a sequential framework which can be used as a model selection tool for the GMM. We consider some simulations in Section 6.1, and additionally analyze a gene expression data set in Section 6.2 within Chapter 6. We defer the technical details of most proofs to the Appendix. We also describe several other tests that are used for comparison in Section 4.5.

Chapter 3

Inference Using k-means Clustering: The SigClust Procedure

In this chapter, we study the SigClust procedure which was introduced by Liu et al. (2008) in detail. Liu et al. (2008) define a cluster as a population sampled from a single Gaussian distribution. They then propose SigClust, a testing procedure, that directly targets this Gaussian definition of a cluster to assess the statistical significance of a clustering. To capture the non-Gaussianity due to the presence of multiple clusters, they use the k-means cluster index (CI) as the test statistic, which is the within-class sums of squares about the mean, divided by the total sum of squares about overall mean, in the case where $k = 2$. The Gaussian null distribution allows a direct formulation of the p value which can quantify the significance of a given clustering.

Now we introduce some mathematical notation in order to formally define the SigClust testing procedure. Let $X_1, X_2, \ldots, X_n \sim \mathbb{P}$ be i.i.d. observations from some distribution with probability measure $\mathbb{P}$ on $\mathbb{R}^d$. The objective of the k-means clustering algorithm is to choose cluster centers $\mathbf{b_n} = (b_{n1}, \ldots, b_{nk}) \in \mathbb{R}^{d\times k}$ that minimize the within-cluster sum of squares,

$$W_n(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n} \min_{1\le j\le k} \|X_i - a_j\|^2 \tag{3.1}$$

as a function of $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^{d\times k}$. For each center a_j, we can also associate a convex polyhedron A_j which contains all points in $\mathbb{R}^d$ closer to a_j than to any other center. The sets $\{A_1, \ldots, A_k\}$ are the Voronoi tessellation of $\mathbb{R}^d$. The tessellation defines the clustering. We also define,

$$W(\mathbf{a}) = \mathbb{E}[W_n(\mathbf{a})],$$

and we let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^{d \times k}$ denote the minimizer of $W(\mathbf{a})$. When the minimizers are not unique we let $\mathbf{b_n}$ and $\boldsymbol{\mu}$ denote arbitrary minimizers of $W_n(\mathbf{a})$ and $W(\mathbf{a})$ respectively.

Then the SigClust test statistic T_n, defined to be the ratio between the within-class sum of squares and the total sum of squares, is given by

$$T_n = T_n(\mathbf{b_n}) = \frac{W_n(\mathbf{b_n})}{\frac{1}{n}\sum_{i=1}^n \|X_i - \overline{X}\|^2} = \frac{\sum_{i=1}^n \min_{1 \le j \le 2} \|X_i - b_{nj}\|^2}{\sum_{i=1}^n \|X_i - \overline{X}\|^2},$$

where $\mathbf{b_n} = (b_{n1}, b_{n2})$ is the vector of optimal centers chosen by the 2-means clustering algorithm and $\overline{X}$ is the sample mean of the data. We note in passing that in their extension of this method to hierarchical clustering, Kimes et al. (2017) also consider other statistics that arise in hierarchical clustering.

The test rejects the null for small values of this statistic. In order to estimate the p-value we can use a version of the parametric bootstrap. The (estimated) p-value is an estimate of $\mathbb{P}_{\widehat{\mu},\widehat{\Sigma}}(T_n^* < T_n)$ where T_n^* is computed on the bootstrap samples from $\mathbb{P}_{\widehat{\mu},\widehat{\Sigma}}$, where $\mathbb{P}_{\widehat{\mu},\widehat{\Sigma}} = N(\widehat{\mu}, \widehat{\Sigma})$ and where $\widehat{\mu} = \overline{X}$ and $\widehat{\Sigma}$ is the sample covariance matrix. We note that in the high dimensional case, as discussed earlier, Liu et al. (2008) use a regularized estimator of Σ.

3.1 Limiting Distribution of SigClust Under the Null

In order to analytically understand the SigClust procedure and to develop results regarding its power we first find the limiting distribution of the test statistic under the null in a simplified setup.

We focus in this and subsequent sections on the case when under the null hypothesis, we obtain samples from $\{X_1, \ldots, X_n\} \sim N(0, \Sigma)$ where Σ is a diagonal matrix. We assume that the two leading eigenvalues are distinct which ensures that, under the null, the k-means objective at the population-level has a unique optimal solution whose optimal value is tractable to analyze in closed-form. For notational convenience, we will assume that, $\sigma_1^2 > \sigma_2^2 \ge \sigma_3^3 \ldots \ge \sigma_d^2 > 0$.

Our results extend in a straightforward way to the general non-spherical, axis-aligned case with minor modifications. These results in turn are easily generalized to the non-spherical, not necessarily axis-aligned, case by noting the invariance of the test statistic to orthonormal rotations under the null. The spherical case is more challenging since the population optimal k-means solution is not unique and the limiting distribution is more complicated. To illustrate some of the difficulties, we derive the limiting distribution of the SigClust statistic, under the null, for the two-dimensional case in Appendix A.1.3, but do not consider the power of the test in that setting.

Recall, that $\boldsymbol{\mu}$ denotes the (unique) population optimal k-means solution, and we use $\{A_1, A_2\}$ to denote the corresponding Voronoi partition. Our results build on the following result from Pollard (1982) and Bock (1985):

Lemma 3.0.1 (Corollary 6.2 of Bock (1985))**.** *The minimum within cluster sum of squares* $W_n(\mathbf{b_n})$ *has an asymptotically normal distribution given by,*

$$\sqrt{n}(W_n(\mathbf{b_n}) - W(\boldsymbol{\mu})) \rightsquigarrow N(0, \tau^2), \quad as \quad n \to \infty,$$

where

$$\tau^2 := \sum_{i=1}^{2} \mathbb{P}(A_i)\mathbb{E}\left[\|X - \mu_i\|^4 | X \in A_i\right] - \left[W(\boldsymbol{\mu})\right]^2.$$

To analyze the power of the SigClust procedure, and to better understand its limiting distribution, we need to calculate $W(\boldsymbol{\mu}), \tau^2$ and the mass of the Voronoi cells. It is easy to verify that under the null the probability of each of the Voronoi cells corresponding to $\boldsymbol{\mu}$ is $1/2$. In Appendix A.1.1, we establish the following claims:

$$W(\boldsymbol{\mu}) = \sum_{i=1}^{d} \sigma_i^2 - \frac{2\sigma_1^2}{\pi} \tag{3.2}$$

$$\tau^2 = 2\sum_{i=1}^{d} \sigma_i^4 - \frac{16\sigma_1^4}{\pi^2}. \tag{3.3}$$

As a consequence of these calculations, we obtain the limiting distribution of the SigClust statistic under the null:

Theorem 3.1. *For* $W(\boldsymbol{\mu})$ *and* τ^2 *defined in* (3.2) *and* (3.3) *we have that,*

$$\sqrt{n}\left(T_n(\mathbf{b_n}) - \frac{W(\boldsymbol{\mu})}{\sum_{i=1}^{d}\sigma_i^2}\right) \rightsquigarrow N\left(0, \left[\frac{\tau}{\sum_{i=1}^{d}\sigma_i^2}\right]^2\right), \quad as \quad n \to \infty.$$

Remark: Leveraging this result, we are able to characterize the rejection region of the test and in Theorem 3.4 we analyze its power. The proof of Theorem 3.1 is quite long and technical. Most of the work is done in the Appendix. Here is a brief proof that leverages Lemma 3.0.1 which contains most of the technical details.

Proof for theorem 3.1. From Lemma 3.0.1 we have that,

$$\sqrt{n}(W_n(\mathbf{b_n}) - W(\boldsymbol{\mu})) \rightsquigarrow N(0, \tau^2), \quad \text{as} \quad n \to \infty.$$

Furthermore by the Weak Law of Large Numbers we have that,

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \overline{X}\|^2 \xrightarrow{p} \sum_{i=1}^{d} \sigma_i^2.$$

Putting these together yields the desired claim. □

3.2 Geometry of k-means Under the Alternative

Our goal is to find special cases where we are able to explicitly calculate SigClust's power and understand cases in which it has high power and cases where it has low power. In order to find the power, we first need to understand the behaviour of 2-means clustering under the alternative. In particular, we need to understand what the optimal split is and what the optimal within sum of squares is, if the data was indeed generated from the alternative.

We focus on the case when the data, under the alternative, is generated from a mixture of two Gaussian distributions of the form

$$\{X_1, \ldots, X_n\} \sim \frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D), \tag{3.4}$$

where $\theta_1 = (a/2, 0, \ldots, 0) \in \mathbb{R}^d$, a is a non-zero constant and D is a diagonal matrix,

$$D = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \ldots & \sigma_d^2 \end{bmatrix}.$$

In this section, we will consider cases where $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$, allowing in some cases σ_2^2 to be larger than σ_1^2. We treat the case when $a > 0$ and $0 < \sigma_d^2 \leq \ldots \leq \sigma_2^2, \sigma_1^2 < \infty$, are fixed (do not vary with the sample-size).

For technical reasons, we make a small modification to 2−means clustering. We consider 2−means clustering with symmetric centers. That is, we consider $\mathbf{b_n}^{(0)}$ that minimizes the within-cluster sum of squares,

$$W_n^{(0)}(t) := W_n(t, -t) = \frac{1}{n}\sum_{i=1}^{n} \min\{||X_i - t||^2, ||X_i + t||^2\}, \tag{3.5}$$

as a function of $t \in \mathbb{R}^d$.

We also introduce notation for the optimal split by considering a symmetric population version of the 2-means clustering for the following theorems and lemmas. We define the following terms to be used in the lemmas. Let

$$\mu^* = \begin{pmatrix} \mu_1^* \\ \mu_2^* \end{pmatrix} = \begin{pmatrix} \mu_1^* \\ -\mu_1^* \end{pmatrix},$$

where μ_1^* and $-\mu_1^*$ denote the optimum cluster centers that minimize the within sum of squares when symmetric 2-means clustering is performed on the data. The corresponding minimum within sum of squares

is denoted by $W(\mu^*)$. That is,

$$W^{(0)}(\mu_1^*) := W(\mu^*) = \inf_{t \in \mathbb{R}^d} E\left[\min\{||X-t||^2, ||X+t||^2\right] = E\left[\min\{||X-\mu_1^*||^2, ||X+\mu_1^*||^2\}\right].$$

We conjecture that this symmetric assumption has no practical effect on SigClust, since the samples are drawn from a symmetric distribution and in practice the optimum 2-means cluster centers are close to being symmetric. Moreover, to consider the limiting distribution of $W_n(\mathbf{b_n})$, given by Theorem 6.4 (b) of Bock (1985, pp. 101), we need the population optimal centers to be unique. This is guaranteed only if the population optimal centers are symmetric about the origin, since if (μ_1^*, μ_2^*) minimizes the population within sum of squares, then due to the symmetry of the distribution, $(-\mu_1^*, -\mu_2^*)$ also minimizes the population within sum of squares. Therefore for the minimizer to be unique, $\mu_2^* = -\mu_1^*$.

Therefore we state a result analogous to Theorem 6.4 (b) of Bock (1985, pp. 101) for symmetric 2-means clustering for our population as follows:

Theorem 3.2. *Let the data be generated from $\frac{1}{2}N(-\theta_1, D)+\frac{1}{2}N(\theta_1, D)$, as defined above, and $\mathbf{b_n}^{(0)}, \mu_1^*, \mu^*, W_n^{(0)}(t)$ and $W^{(0)}(\mu_1^*)$ are as defined above. Suppose*

(i) the vector μ_1^* that minimizes $W^{(0)}(\mu_1^*)$ is unique upto relabeling of its coordinates;

(ii) the matrix G is positive definite, where G as defined in Pollard (1982) (as Γ) is a matrix made up of $d \times d$ matrices of the form,

$$G_{ij} = \begin{cases} 2\mathbb{P}(A_i)\mathbf{I}_d - 2r_{ij}^{-1}\int_{M_{ij}} f(x)(x-\mu_i^*)(x-\mu_i^*)^T \, d\sigma(x) & \text{for} \quad i=j \\ -2r_{ij}^{-1}\int_{M_{ij}} f(x)(x-\mu_i^*)(x-\mu_j^*)^T \, d\sigma(x) & \text{for} \quad i \neq j, \end{cases} \tag{3.6}$$

for $i, j \in \{1, 2\}$ where $r_{ij} = \|\mu_i^* - \mu_j^*\|$, $f(\cdot)$ is the corresponding density function, $\sigma(\cdot)$ is the $(d-1)$ dimensional Lebesgue measure, A_1 is the convex polyhedron that contains all points in $\mathbb{R}^d$ that are closer to μ_1^* compared to $-\mu_1^*$ and A_2 is vice-versa and M_{ij} denotes the face common to A_i and A_j, and $\mathbf{I}_d$ denotes the $d \times d$ identity matrix.

Then as $n \to \infty$,

$$\sqrt{n}(W_n^{(0)}(\mathbf{b_n}^{(0)}) - W(\mu^*)) \rightsquigarrow N(0, \tau^{*2}),$$

where

$$\tau^{*2} = \sum_{i=1}^{2} P(A_i) E[||X - E[X|X \in A_i]||^4 | X \in A_i] - [W(\mu^*)]^2.$$

Since μ_1^* and $-\mu_1^*$ denote the optimum cluster centers, the corresponding optimal separating hyperplane passes through the origin. We denote the corresponding optimal separating hyperplane by

$$\mathcal{H}(b^*) = \left\{ y \in \mathbb{R}^d : b^{*T} y = 0 \right\}, \quad \text{where} \quad \sum_{i=1}^{d} b_i^{*2} = 1.$$

Then the corresponding within sum of squares can be written as:

$$\begin{aligned} W(b^*) := W(\mu^*) &= \inf_{b \in \mathbb{R}^d} \big\{ P(b^T X > 0) E[||X - E[X|b^T X > 0]||^2 | b^T X > 0] \\ &\qquad + P(b^T X < 0) E[||X - E[X|b^T X < 0]||^2 | b^T X < 0] \big\} \\ &= \inf_{b \in \mathbb{R}^d} E[||X - E[X|b^T X > 0]||^2 | b^T X > 0], \qquad (\text{Since, } -X \stackrel{d}{=} X) \\ &= E[||X - E[X|b^{*T} X > 0]||^2 | b^{*T} X > 0]. \end{aligned}$$

The following theorem gives the optimal separating hyperplane and the optimal within sum of squares under the alternative.

Theorem 3.3. *For data generated from $\frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D)$, where $\theta_1 = (a/2, 0, \ldots, 0) \in \mathbb{R}^d$, $a > 0$ is fixed and D is a diagonal matrix with elements $D_{jj} = \sigma_j^2$, such that $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$ are fixed.*

1. *When*

$$\sigma_2^2 < \sigma_1^2 + \frac{a^2}{4}, \tag{3.7}$$

the unique optimal separating hyperplane which gives the minimum within sum of squares $W(b^*)$ is given by $\mathcal{H}(b) = \{y \in \mathbb{R}^d : y_1 = 0\}$, that is, the unique optimal b^* is such that $b_1^* = 1$ and $b_i^* = 0$ for every $i \neq 1$. The corresponding optimal within sum of squares is given by

$$W(\mu^*) = W(b^*) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \left(\sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\sigma_1^2}} + \frac{a}{2}\ P\left(|Z| < \frac{a}{2\sigma_1} \right) \right)^2. \tag{3.8}$$

2. *When*

$$\sigma_2^2 > \max \left\{ \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4 + \frac{a^4}{64}}}{2\sigma_1^2}, \frac{\pi}{2} \left(\sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\ \sigma_1^2}} + \frac{a}{2}\ P\left(|Z| < \frac{a}{2\ \sigma_1} \right) \right)^2 \right\}, \tag{3.9}$$

the unique optimal separating hyperplane which gives the minimum within sum of squares $W(b^*)$ is given by $\mathcal{H}(b) = \{y \in \mathbb{R}^d : y_2 = 0\}$, that is, the unique optimal b^* is such that $b_2^* = 1$ and $b_i^* = 0$ for

every $i \neq 2$. The corresponding optimal within sum of squares is given by

$$W(\mu^*) = W(b^*) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_2^2. \tag{3.10}$$

In simpler words, the theorem implies that when the condition in (3.7) holds, i.e. when the variance along the second covariate is small, the optimal symmetric 2-means split at the population-level splits the data along the first covariate. On the other hand when the condition in (3.9) holds, i.e. when the variance along the second covariate is large, the optimal symmetric 2-means split at the population-level is along the second covariate.

We conjecture that even for 2-means clustering without the symmetric assumption, as long as the data is generated from $\frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D)$, the above statement holds. That is, when the condition in (3.7) holds, the optimal 2-means split at the population-level is along the first covariate and on the other hand when the condition in (3.9) holds, the optimal 2-means split at the population-level is along the second covariate.

Additionally we also have the following lemma:

Lemma 3.3.1. *In both the cases mentioned in Theorem 3.3, the matrix G given by equation (3.6) is positive definite.*

Therefore Theorem 3.3 and the above Lemma 3.3.1 combined together with Theorem 3.2 give the limiting distribution under the alternative.

3.3 Power of the SigClust Procedure

In this section we derive the asymptotic power of the test using the previous results on the limiting distribution. Since in the previous section we assumed using a symmetric $2-$means clustering we now consider the test statistic for the symmetric $2-$means clustering. We define

$$T_n^{(0)} := T_n^{(0)}\left(\mathbf{b_n}^{(0)}\right) = \frac{W_n^{(0)}(\mathbf{b_n}^{(0)})}{\frac{1}{n}\sum_{i=1}^{n} \|X_i - \overline{X}\|^2}. \tag{3.11}$$

Let

$$\text{Power}_n(a) = \mathbb{P}(T_n^{(0)} > t_{\alpha,n}),$$

denote the power of the test where $t_{\alpha,n}$ denotes the α-level critical value. Building once again on the result in Lemma 3.0.1 and additionally on Theorem 3.2, we show the following result:

Theorem 3.4. *Suppose that samples are generated according to the model described in* (3.4) *and let* $Z \sim N(0,1)$ *then:*

1. **Consistent:** *If,*

$$\sigma_2^2 < \sigma_1^2 + \frac{a^2}{4}, \tag{3.12}$$

then SigClust is consistent, i.e. $\text{Power}_n(a) \to 1$ *as* $n \to \infty$.

2. **Inconsistent:** *On the other hand if,*

$$\sigma_2^2 > \max\left\{ \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4 + \frac{a^4}{64}}}{2\sigma_1^2}, \frac{\pi}{2}\left(\sqrt{\frac{2}{\pi}}\ \sigma_1\ \exp\left(-\frac{a^2}{8\sigma_1^2}\right) + \frac{a}{2}\ P\left(|Z| < \frac{a}{2\ \sigma_1}\right)\right)^2 \right\} \tag{3.13}$$

then SigClust is inconsistent, i.e. $\text{Power}_n(a) < 1$ *as* $n \to \infty$.

Remarks:

1. In order to roughly understand the result, as we show more precisely in the Appendix for small values of $a > 0$:

$$\frac{\pi}{2}\Big(\sqrt{\frac{2}{\pi}}\sigma_1 \exp\left(-\frac{a^2}{8\sigma_1^2}\right) + \frac{a}{2}\mathbb{P}\Big(|Z| < \frac{a}{2\sigma_1}\Big)\Big)^2 \approx \sigma_1^2 + \frac{a^2}{4},$$

where we use $\approx$ to mean equal up to a small error of size roughly a^4/σ_1^2. As a consequence, in our setup we see that when the variance of the second covariate is sufficiently large SigClust has no power in detecting departures from Gaussianity along the first covariate.

2. We observe a phase-transition in the power of SigClust, and we provide a precise characterization of this phase-transition. We highlight that the low power of SigClust is a persistent phenomenon, i.e. there is a large, non-vanishing part of the parameter space where *the test is not consistent.* We see that the power of SigClust is very sensitive to the particular values of the variances in the matrix D. In the next chapter we consider alternative tests based on relative-fit that address these drawbacks of SigClust.

3. The proof of this result is quite technical and we defer the details to Appendix A.3. At a high-level, the proof follows from Theorem 3.3 which characterizes the optimal 2-means split at the population-level, and uses it to study the distribution of the test statistic under the alternate. We then leverage our previous characterization of the distribution of the test statistic under the null to study the power of SigClust.

4. Despite the technical nature of the proof, the intuition behind the phase-transition is quite natural. As shown in Theorem 3.3, when the condition in (3.12) holds, the optimal 2-means split at the population-level splits the data along the first covariate and as a result the test is able to detect the non-Gaussianity of the first covariate. On the other hand when the condition in (3.13) holds, the optimal 2-means split at the population-level is along the second covariate and the resulting test is asymptotically inconsistent.

5. Finally, we note in passing that in the case when

$$\frac{\pi}{2}\Big(\sqrt{\frac{2}{\pi}}\sigma_1 \exp\left(-\frac{a^2}{8\sigma_1^2}\right) + \frac{a}{2}\mathbb{P}\Big(|Z| < \frac{a}{2\sigma_1}\Big)\Big)^2 = \sigma_2^2,$$

the 2-means solution is no longer unique, and we are unable to use our techniques to characterize the power of the test. However, we conjecture that SigClust remains inconsistent even in this case.

Chapter 4

A Test for Relative Fit of Mixtures (RIFT)

A natural way to improve the low power of SigClust is to formally test for whether the data are generated from a Gaussian versus a mixture of Gaussians. There is a long history of research on this problem; see, for example, Dacunha-Castelle et al. (1999); Gassiat (2002); Chen (2017); Gu et al. (2017) and references therein. As we mentioned earlier, the mixture model is irregular and there has been little success in deriving a practical, simple test with valid type I error control. Furthermore, and more importantly, such tests ignore the fact that we are only using the parametric model as an approximation; we don't expect that the true distribution is exactly Gaussian or a mixture of Gaussians. This motivates our new approach where we test the relative fit of the models without assuming that either model is correct. Also, our test is valid for multivariate mixtures whereas many of the existing tests are for the univariate case.

4.1 The Basic Test: RIFT

Let $\mathcal{P}_1$ denote the set of multivariate Gaussians and let $\mathcal{P}_2$ denote the set of mixtures of two multivariate Gaussians. We are given a sample $X_1, \ldots, X_{2n} \sim P$ but we do not assume that P is necessarily in either $\mathcal{P}_1$ or $\mathcal{P}_2$. Note that, for notational simplicity, we denote the total sample size by $2n$.

We randomly split the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$. Assume each has size n. Using $\mathcal{D}_1$, fit a Gaussian $\widehat{p}_1$ and a mixture of two Gaussians $\widehat{p}_2$. Any consistent estimation procedure can be used; in our examples we use the Expectation Maximization (EM) algorithm. Understanding precise conditions under which EM yields a global maximizer is an area of active research (Balakrishnan et al., 2017), but we do not pursue this further in this thesis.

Instead of testing $H_0 : P \in \mathcal{P}_1$ versus $H_1 : P \in \mathcal{P}_2$ we test whether $\widehat{p}_2$ is a significantly better fit for the data than $\widehat{p}_1$. This is a different hypothesis from the usual one, but, arguably, it is more relevant since it is $\widehat{p}_1$ or $\widehat{p}_2$ that will be used for clustering. Furthermore, this does not require that the true distribution be in either $\mathcal{P}_1$ or $\mathcal{P}_2$.

To formalize the test, let

$$\Gamma = K(p, \widehat{p}_1) - K(p, \widehat{p}_2), \tag{4.1}$$

where $K(p, q) = \int p \log(p/q)$ is the Kullback-Leibler distance and p is the true density. Note that Γ is a random variable. Formally, we will test, conditional on $\mathcal{D}_1$,

$$H_0 : \Gamma \leq 0 \quad \text{versus} \quad H_1 : \Gamma > 0. \tag{4.2}$$

Since Γ is a random variable, these are random hypotheses. Let

$$\widehat{\Gamma} = \frac{1}{n} \sum_{i \in \mathcal{D}_2} R_i \tag{4.3}$$

where $R_i = \log\left(\widehat{p}_2(X_i)/\widehat{p}_1(X_i)\right)$. Below, we show that, conditionally on $\mathcal{D}_1$,

$$\sqrt{n}(\widehat{\Gamma} - \Gamma) \approx N(0, \tau^2) \qquad \text{as} \quad n \to \infty,$$

where $\tau^2 \equiv \tau^2(\mathcal{D}_1) = \mathbb{E}[R_i^2] - \Gamma^2$. The quantity τ^2 can be estimated by $\widehat{\tau}^2 = \frac{1}{n} \sum_{i \in \mathcal{D}_2} (R_i - \overline{R})^2$. We reject H_0 if

$$\widehat{\Gamma} > \frac{z_\alpha \widehat{\tau}}{\sqrt{n}},$$

and we refer to this as the RIFT (Relative Information Fit Test). For technical reasons, we make a small modification to the test statistic. We replace R_i with $\widetilde{R}_i = R_i + \delta Z_i$ where $Z_1, \ldots, Z_n \sim N(0, 1)$, $\{Z_i : i = 1 \ldots n\}$ are independent of the observed data and δ is some small positive number, for example, $\delta = 0.00001$. This has no practical effect on the test and is only needed for the theory.

For the following result, let the fitted Gaussian density be given by $\widehat{p}_1 = N(\widehat{\mu}, \widehat{\Sigma})$ and the fitted mixture of two Gaussians be given by $\widehat{p}_2 = \widehat{\alpha}\widehat{f}_1 + (1 - \widehat{\alpha})\widehat{f}_2$, where $\widehat{f}_1 = N(\widehat{\mu}_1, \widehat{\Sigma}_1)$ and $\widehat{f}_2 = N(\widehat{\mu}_2, \widehat{\Sigma}_2)$. For technical reasons, we restrict the parameter estimates to lie in a compact set. Formally, we assume that each $\widehat{\mu}_i$ is restricted to lie in a compact set $\mathcal{A}$ and that the eigenvalues of $\widehat{\Sigma}$ and $\widehat{\Sigma}_i$ lie in some interval $[c_1, c_2]$ for $i = 1, 2$, where $c_1, c_2 > 0$. As a consequence of data splitting, the test of relative fit has a simple limiting distribution unlike the usual tests for mixtures which have intractable limits.

Theorem 4.1. *Let* $Z \sim N(0, \tau^2)$ *where* $\tau^2 = \mathbb{E}[(\widetilde{R}_i - \Gamma)^2 | \mathcal{D}_1]$. *Then, under* H_0

$$\sup_t \left| \mathbb{P}(\sqrt{n}(\widehat{\Gamma} - \Gamma) \leq t \,|\, \mathcal{D}_1) - \mathbb{P}(Z \leq t) \right| \leq \frac{C}{\sqrt{n}} \tag{4.4}$$

where $C = \frac{C_0}{\delta^3}\left[8C_1^3 + \delta\left(12C_1^2\sqrt{\frac{2}{\pi}} + 6C_1\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right)\right]$, $C_0 = 33/4$ *and* C_1 *is a constant.*

Remark: It is also possible to consider the normalized version of the statistic $\widehat{\Gamma}$. Formally, under the conditions of the above result conditional on $\mathcal{D}_1$:

$$\sup_t \left| P\left(\sqrt{n}\left(\frac{\widehat{\Gamma}}{\widehat{\tau}} - \frac{\Gamma}{\tau}\right) \leq t\right) - \mathbb{P}(Z \leq t) \right| \leq \frac{C}{\sqrt{n}}$$

where $Z \sim N(0,1)$. We note that since the constant C does not depend on $\mathcal{D}_1$ this result also holds unconditionally.

We now turn our attention to the power of RIFT. Suppose that we consider a distribution p such that,

$$\Gamma^* = \inf_{p_1 \in \mathcal{P}_1} K(p, p_1) - \inf_{p_2 \in \mathcal{P}_2} K(p, p_2) > 0, \tag{4.5}$$

i.e. p is a distribution for which the class of mixtures of two Gaussians provides a strictly better fit than a single Gaussian. Then we have the following result characterizing the power of RIFT:

Theorem 4.2. *Suppose that* Γ^* *in* (4.5) *is strictly positive, then* RIFT *is asymptotically consistent, i.e. as* $n \to \infty$.

$$Power_n(\text{RIFT}) = \mathbb{P}(\widehat{\Gamma} > z_\alpha \widehat{\tau}/\sqrt{n}) \to 1.$$

Remark: A consequence of this result is that RIFT is consistent against any fixed distribution $p \in \mathcal{P}_2 \backslash \mathcal{P}_1$. In other words, the power deficiency of SigClust observed in Theorem 3.4 does not happen for our test.

4.2 Variants of RIFT

In this section we introduce and study a few variants of RIFT that can be advantageous in various applications.

4.2.1 A Robust, Exact Test

The Kullback-Leibler (KL) distance between two densities p and q is $K(p,q) = \mathbb{E}_p[W]$ where $W = \log(p(X)/q(X))$. This distance can be sensitive to the tail of the distribution of W. For this reason we also consider a robustified version of the KL distance, namely, $\widetilde{K}(p,q) = \text{Median}_P[W]$, that is, the median

of W under p (we will assume for convenience that the median is unique). In this case, the sample median of $W_1, \ldots, W_n$ is a consistent estimator of $\widetilde{K}(p,q)$, where $W_i = \log\left(p(X_i)/q(X_i)\right)$.

For relative fit we define

$$\widetilde{\Gamma} = \text{Median}_p[R] \tag{4.6}$$

where $R = \log \widehat{p}_2(X)/\widehat{p}_1(X)$. A point estimate is the sample median based on $\mathcal{D}_2$. To test $H_0 : \widetilde{\Gamma} \leq 0$ versus $H_1 : \widetilde{\Gamma} > 0$ we use the sign test. Hence, under H_0, $\mathbb{P}(\text{rejecting } H_0) \leq \alpha$. We will refer to this as median-RIFT or M-RIFT. This approach has two advantages: it is robust and it does not require any asymptotic approximations.

4.2.2 ℓ_2 Version

The test does not have to be based on Kullback-Leibler distance. We can also use the ℓ_2 distance as we now explain. Define the ℓ_2-relative fit by $\Theta = \int (p - \widehat{p}_1)^2 - \int (p - \widehat{p}_2)^2$. We test, conditional on $\mathcal{D}_1$,

$$H_0 : \Theta \leq 0 \quad \text{versus} \quad H_1 : \Theta > 0.$$

To estimate Θ, note that we can write $\Theta = \int \widehat{p}_1^2 - \int \widehat{p}_2^2 - 2\int p(\widehat{p}_1 - \widehat{p}_2)$ which can be estimated by

$$\widehat{\Theta} = \int \widehat{p}_1^2 - \int \widehat{p}_2^2 - \frac{2}{n}\sum_{i \in \mathcal{D}_2} U_i$$

where $U_i = \widehat{p}_1(X_i) - \widehat{p}_2(X_i)$. To evaluate the integrals, we use importance sampling. We sample $Y_1, \ldots, Y_N \sim g$ from a convenient density g (such as a t-distribution) and then use

$$\int \widehat{p}_1^2 \approx \frac{1}{N}\sum_j \frac{\widehat{p}_1^2(Y_j)}{g(Y_j)}, \quad \int \widehat{p}_2^2 \approx \frac{1}{N}\sum_j \frac{\widehat{p}_2^2(Y_j)}{g(Y_j)}.$$

Again, for technical reasons, we make a small modification to the test statistic. We replace U_i with $\widetilde{U}_i = U_i + \delta Z_i$ where $Z_1, \ldots, Z_n \sim N(0,1)$, $\{Z_i : i = 1 \ldots n\}$ are independent of the observed data and δ is some tiny positive number, for example, $\delta = 0.00001$. Again this has no practical effect on the test and is only needed for the theory. Recall that the Gaussian density is given by $\widehat{p}_1 = N(\widehat{\mu}, \widehat{\Sigma})$ and the mixture of two Gaussians is given by $\widehat{p}_2 = \alpha\widehat{f}_1 + (1-\alpha)\widehat{f}_2$, where $\widehat{f}_1 = N(\widehat{\mu}_1, \widehat{\Sigma}_1)$ and $\widehat{f}_2 = N(\widehat{\mu}_2, \widehat{\Sigma}_2)$. Once again, we assume that $\widehat{\mu}_i$ are restricted to lie in a compact set $\mathcal{A}$ and that the eigenvalues of $\widehat{\Sigma}$ and $\widehat{\Sigma}_i$ lie in the interval $[c_1, c_2]$ for $c_1, c_2 > 0$ and for $i = 1, 2$.

Theorem 4.3. *Let $Z \sim N(0, a^2)$ where $a^2 = var(\widetilde{U}_i)$. Then, under H_0,*

$$\sup_t |\mathbb{P}(\sqrt{n}(\widehat{\Theta} - \Theta) \leq t | \mathcal{D}_1) - \mathbb{P}(Z \leq t)| \leq \frac{\widetilde{C}}{\sqrt{n}}, \tag{4.7}$$

where $\widetilde{C} = \frac{C_0}{\delta^3}\left[8C_2^3 + \delta\left(12C_2^2\sqrt{\frac{2}{\pi}} + 6C_2\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right)\right]$, $C_0 = 33/4$ and C_2 is a constant.

4.3 Aside: A Test for Mixtures Using RIFT

Our focus is on the relative fit as described in the previous section. However, it is possible to modify our test so that it tests the more traditional hypotheses

$$H_0 : P \in \mathcal{P}_1 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_2$$

where $\mathcal{P}_1$ are Normals and $\mathcal{P}_2$ are the mixtures of two Normals. There is currently no available test that is simple, asymptotically valid and has easily computable critical values in the multivariate case. But we can use our split test for this hypothesis if we modify the test using the idea of Ghosh and Sen (1984) where we force the fit under the alternative to be bounded away from the null. When combined with data splitting, this results in a valid test. Specifically, when we fit H_1, we will constrain the fitted density $\widehat{p}_2$ to satisfy $K(p, \widehat{p}_2) > \Delta$ for all $p \in \mathcal{P}_1$. Here, Δ is any small, positive constant.

Theorem 4.4. *If $P \in \mathcal{P}_1$ then $\mathbb{P}(\widehat{\Gamma} > z_\alpha \widehat{\tau}/\sqrt{n}) \leq \alpha + o(1)$. Indeed, it can be shown that, $\mathbb{P}(\widehat{\Gamma} > z_\alpha \widehat{\tau}/\sqrt{n}) = o(1)$ for any fixed $\alpha \in (0, 1)$.*

Hence, combining data splitting with the Ghosh-Sen separation idea yields an asymptotically valid test for mixtures with a simple critical value. To the best of our knowledge, this is the first such test.

4.4 Truncated RIFT

If the support of the data is a truncated space, the null hypothesis will be a truncated Normal rather than a Normal. For example, if we use RIFT for top-down hierarchical clustering, as described later in Chapter 5.1, then after the first split, the test is now applied to the data in a cluster which is a truncated space. So instead of comparing the fit of a Normal $\widehat{p}_1$ and a fit of a mixture of two Normals $\widehat{p}_2$, we need to compare the fit of a truncated normal to a fit of a truncated mixture of two Normals. We can use exactly the same test except that $\widehat{p}_j$ should be replaced with $\widehat{p}_j / \widehat{P}_j(S)$ where S denotes the subset of $\mathbb{R}^d$ corresponding to the cluster being tested. We can estimate $P_j(S)$ as follows. First, generate $Z_1, Z_2, \ldots, Z_m \sim \widehat{P}_j$ for some large m. Then set $\widehat{P}_j(S) = \frac{1}{m}\sum_{i=1}^m \mathbb{I}(Z_i \in S)$. Then replace $\widehat{p}_j$ with with $\widehat{p}_j / \widehat{P}_j(S)$ in the test.

4.5 Other Tests for Significance of a Cluster

Another way to decide whether to split a cluster or not is to use a goodness-of-fit test for Normality. In this section we describe two such tests. Note that such tests can only be used for the first split in the clustering problem. We include them in our study because they are simple and they provide a point of comparison. We also note that it is possible to use tests for goodness-of-fit with minimax-optimal power against neighborhoods defined in particular metrics, based on binning and the χ^2-test, but these tests are complex and have tuning parameters that need to be carefully chosen.

4.5.1 Mardia's Multivariate Kurtosis Test

Mardia (1974) proposed using the Kurtosis measure to test for normality. If X is a d-dimensional random (column) vector with expectation $\mu = \mathbb{E}[X]$ and non-singular covariance matrix $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$, Mardia (1970) defines the multivariate Kurtosis as

$$\beta_2 = \mathbb{E}\left[\left\{(X-\mu)^T\Sigma^{-1}(X-\mu)\right\}^2\right].$$

The proposed test uses the Kurtosis measure to test for multivariate normality. If $X_1, \ldots, X_n \in \mathbb{R}^d$ are independent observations from any multivariate normal distribution, then the sample analogue of Kurtosis is given by,

$$b_{2,d} = \frac{1}{n}\sum_{j=1}^{n}\left\{\left(X_j - \overline{X}\right)^T S_n^{-1}\left(X_j - \overline{X}\right)\right\}^2,$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_j, \quad S_n = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)\left(X_j - \overline{X}\right)^T$$

are the sample mean vector and the sample covariance matrix. Mardia (1970) shows that $b_{2,d}$ has a distribution under the null hypothesis, H_0, given by

$$\frac{\sqrt{n}\left(b_{2,d} - d(d+2)\right)}{\sqrt{8d(d+2)}} \xrightarrow{d} N(0,1)$$

as $n \to \infty$. So we reject the null hypothesis for both large and small values of $b_{2,d}$. This multivariate normality test is consistent if, and only if,

$$\mathbb{E}\left[\left\{\left(X_i - \mu\right)^T \Sigma^{-1}\left(X_i - \mu\right)\right\}^2\right] \neq d(d+2).$$

For detecting clusters, the method starts by fitting a multivariate Gaussian to the data. We then perform the multivariate normality test using the Kurtosis measure and if the test gets rejected then the data is split

into two clusters. We reject H_0 at level α if

$$\left| \frac{\sqrt{n}\,(b_{2,d} - d(d+2))}{\sqrt{8d(d+2)}} \right| > z_{\alpha/2}.$$

4.5.2 Nearest Neighbor Goodness of Fit Tests

Nearest neighbor (NN) goodness of fit tests were developed by Bickel and Breiman (1983) and Zhou and Jammalamadaka (1993). Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be samples from P with a density function $p(x)$. We want to test $H_0 : P = P_0$ where P_0 has density p_0.

In the clustering framework, we consider the null hypothesis that the data is drawn from a single multivariate Gaussian distribution. That is, we consider p_0 to be the multivariate Gaussian distribution, with some mean μ and covariance matrix Σ. To implement these tests, we first split the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$ and use $\mathcal{D}_1$ in order to estimate the μ and Σ. Therefore in our setting, $P_0 = N(\widehat{\mu}, \widehat{\Sigma})$ is the estimated null.

Let $R_i = \min_{j \neq i} \|X_i - X_j\|$. The first version of this test uses

$$W_i = \exp(-nD_i) := \exp(-np_0(X_i)V(R_i))$$

where $V(r) = K_d r^d$ is the volume of a ball of radius r and $D_i = p_0(X_i)V(R_i)$. Under H_0, the W_i's are approximately Uniform on $[0, 1]$ and hence we can use the Kolmogorov-Smirnov test.

For the second version, we consider the test proposed by Zhou and Jammalamadaka (1993) that uses

$$T_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(nD_i) - \mathbb{E}_0[h(nD_i)]]$$

where h is a bounded function on $[0, \infty)$. The authors show that $\sqrt{n}\, T_n^* \xrightarrow{d} N(0, \sigma^2(h))$ which is independent of the null distribution P_0, where $\sigma^2(h)$ only depends on the function h.

We consider $h(x) = \exp(-x)$ and calculate the test statistic in terms of W_i as

$$T_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\exp(-nD_i) - \mathbb{E}_0[\exp(-nD_i)]] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [W_i - \mathbb{E}_0[W_i]]\,.$$

Since under the null distribution P_0, $W_i \approx U(0, 1)$, $\mathbb{E}_0[W_i] = 0.5$. Therefore, we reject H_0 at level α if

$$\left| \frac{\sqrt{n}\, T_n^*}{\widehat{\sigma}(h)} \right| > z_{\alpha/2}$$

where $\widehat{\sigma}^2(h)$ is the estimated variance of the W_i's.

Chapter 5

Hierarchical and Sequential Clustering

5.1 Hierarchical Clustering

To propose a hierarchical version of RIFT, we apply the procedure recursively. We propose both top-down and bottom-up versions of the algorithm. In both the cases we begin by splitting the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$. The first half $\mathcal{D}_1$ is used to fit the Gaussian and the mixture of two Gaussians whose fits we compare for the test RIFT, and to recursively split the clusters forming a cluster tree. The second half $\mathcal{D}_2$ is used to conduct the significance tests. In the top-down approach, the tests are applied from the top of the tree downwards and we stop when H_0 is not rejected. In the bottom-up approach we start at the bottom of the tree and combine leaves until the test rejects.

In the top-down case we start with the whole space $\mathbb{R}^d$ as the root node and split a node into two nodes every time we reject the null hypothesis of the test (any version of RIFT) using the data at that node. Note that root node we use RIFT, but for subsequent nodes, the nodes represent truncated spaces and hence we need to use the truncated versions of RIFT. This way, by recursively applying the test to each node, we build a binary tree. The final clustering is given by the leaf nodes of the tree derived by the algorithm.

We detail the algorithm below in Algorithm 1. But before we do that, we define some notation.

Definition 5.1. A collection of sets $\mathcal{P} = \{P_1, \ldots, P_m\}$ is said to be a partition of a set A, if the sets

(i) are mutually disjoint, $P_i \cap P_j = \phi$ and

(ii) have as union the entire set, $\cup_{i=1}^{m} P_i = A$.

Definition 5.2. For a sample $\mathcal{D} = \{X_1, \ldots, X_n\}$ of random variables in $\mathbb{R}^d$, we define the partition of $\mathcal{D}$ with respect to a partition $\mathcal{P} = \{P_1, \ldots, P_m\}$ of $\mathbb{R}^d$ as $\mathcal{P}_\mathcal{D} = \{P_{1\mathcal{D}}, \ldots, P_{m\mathcal{D}}\}$, where

$$P_{i\mathcal{D}} = \{X_j : X_j \in P_i\} \quad \forall\, i = 1, \ldots, m.$$

Algorithm 1: Top-down Hierarchical RIFT

Result: Leaf nodes of a binary tree that give the hierarchical clustering at significance level α ($\mathcal{B}$).

Set of "to be split leaf nodes" $= \mathcal{A} = \{\mathbb{R}^d\}$.

Set of "not to be split leaf nodes" $= \mathcal{B} = \phi$, the null set.

Partition of $\mathbb{R}^d = \mathcal{P} = \mathcal{A} \cup \mathcal{B}$.

Initializing node labels: $i = 0, T_0 = \mathbb{R}^d$.

Split the data $\mathcal{D}$ into two sets $\mathcal{D}_1$ and $\mathcal{D}_2$. $\mathcal{D}_1$ is the training set to be used for fitting the Gaussian and the mixture of two Gaussians whose fits are compared by RIFT and $\mathcal{D}_2$ is the test set to be used for performing the test.

Depth function for sets in $\mathcal{A}$, $d : \mathcal{A} \longrightarrow \{0, 1, 2, \ldots\}$ s.t. $d(T_0) = 0$.

while $\mathcal{A} \neq \phi$ **do**

1. Pick $T_j \in \mathcal{A}$, s.t. $d(T_j) = \min_i d(T_i)$.
2. Use $T_{j\mathcal{D}_1}$ to fit a single truncated Gaussian $\widehat{p}_1$ and a mixture of two truncated Gaussians $\widehat{p}_2$.
3. Use $T_{j\mathcal{D}_2}$ along with $\widehat{p}_1$ and $\widehat{p}_2$ to perform RIFT at level $\alpha/2^{2d(T_j)+1}$.

if *reject* RIFT **then**

Split T_j into T_{i+1} and T_{i+2} according to $\widehat{p}_2$;

if $|T_{(i+1)\mathcal{D}}| > 2(4d+1)$ *and* $|T_{(i+2)\mathcal{D}}| > 2(4d+1)$ **then**

Remove T_j from $\mathcal{A}$, add T_{i+1} and T_{i+2} to $\mathcal{A}$ and set $i = i + 2$ and $d(T_{i+1}) = d(T_{i+2}) = d(T_j) + 1$;

else

Remove T_j from $\mathcal{A}$ and add T_j from $\mathcal{B}$;

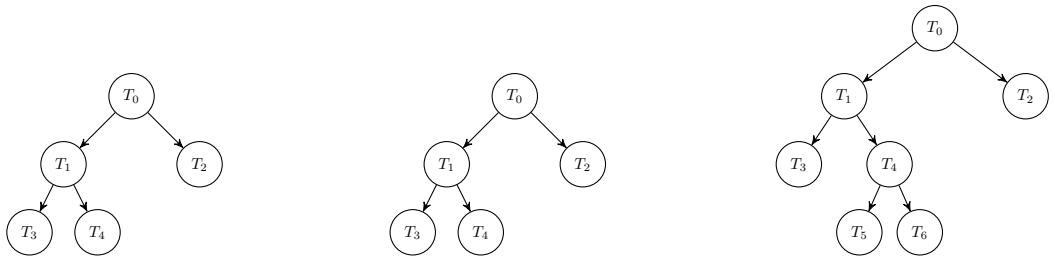
end

else

Remove T_j from $\mathcal{A}$ and add T_j from $\mathcal{B}$;

end

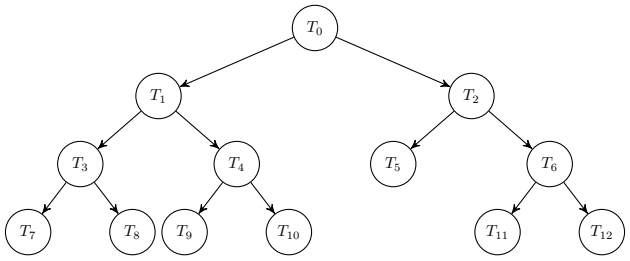
end



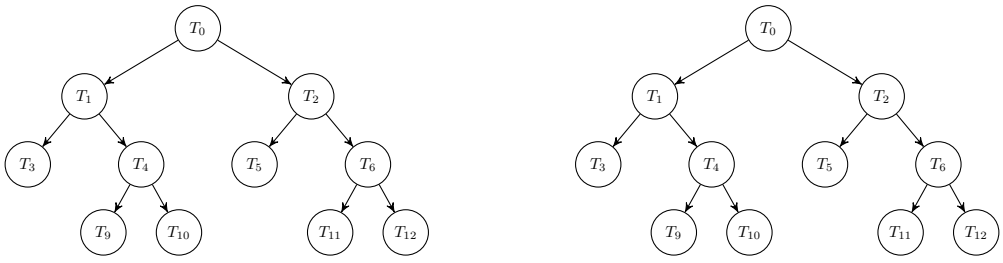
(a) Testing for the split of T_2. **(b)** RIFT at T_2 fails to reject H_0. Testing for the split of T_4. **(c)** RIFT at T_4 rejects H_0.

Figure 5.1: Example of intermediate steps for the top-down hierarchical RIFT procedure.

To demonstrate a splitting step of the top-down clustering process as shown in Algorithm 1, let us say we pick $T_j = T_2$ as shown in Figure 5.1, where $\mathcal{A} = \{T_2, T_3, T_4\}$ and $\mathcal{B} = \phi$. We estimate the truncated Gaussian ($\widehat{p_1}$) and the truncated mixture of two Gaussians ($\widehat{p_2}$) using $T_{2\mathcal{D}_1}$ and then use $T_{2\mathcal{D}_1}$ to perform RIFT at level $\alpha/2^3$. As shown in Figure 5.1, if we fail to reject RIFT, then $\mathcal{A} = \{T_3, T_4\}$ and $\mathcal{B} = \{T_2\}$. The next T_j under consideration is either T_3 or T_4. Suppose we consider $T_j = T_4$ and repeat all the steps as in the case of T_2 to perform RIFT at level $\alpha/2^5$. If in this case RIFT rejects the null, then we split T_4 into T_5 and T_6. Therefore now $\mathcal{A} = \{T_3, T_5, T_6\}$ and $\mathcal{B} = \{T_2\}$. We continue this way until $\mathcal{A}$ is completely empty and then the final clustering is given by the nodes in $\mathcal{B}$.



(a) Testing for the split of T_3 into T_7 and T_8.



(b) RIFT at T_3 fails to reject H_0. Testing for the split of T_6 into T_{11} and T_{12}. **(c)** RIFT at T_6 rejects H_0.

Figure 5.2: Example of intermediate steps for the bottom-up hierarchical RIFT procedure.

Algorithm 2: Bottom-up Hierarchical RIFT

Result: Leaf nodes of a binary tree that give the hierarchical clustering at significance level α ($\mathcal{B}_2$).

Set of "to be split leaf nodes" $= \mathcal{A}_1 = \{\mathbb{R}^d\}$.

Set of "not to be split leaf nodes" $= \mathcal{B}_1 = \phi$, the null set.

Set of "parent nodes" $= \mathcal{C} = \phi$, the null set.

Partition of $\mathbb{R}^d = \mathcal{P} = \mathcal{A}_1 \cup \mathcal{B}_1$.

Initializing node labels: $i = 0, T_0 = \mathbb{R}^d$.

Split the data $\mathcal{D}$ into two sets $\mathcal{D}_1$ and $\mathcal{D}_2$ as in Algorithm 1.

Depth function for sets, $d : (\mathcal{A}_1 \cup \mathcal{B}_1 \cup \mathcal{C}) \longrightarrow \{0, 1, 2, \ldots\}$ s.t. $d(T_0) = 0$.

Parent function, $p : \{1, 2, \ldots\} \longrightarrow \{0, 1, 2, \ldots\}$.

Building Stage:

while $\mathcal{A}_1 \neq \phi$ **do**

  Pick $T_j \in \mathcal{A}_1$, s.t. $d(T_j) = \min_i d(T_i)$;

  Use $T_{j\mathcal{D}_1}$ to fit a single truncated Gaussian $\widehat{p}_1$ and a mixture of two truncated Gaussians $\widehat{p}_2$;

  Split T_j into T_{i+1} and T_{i+2} according to $\widehat{p}_2$;

  if $|T_{(i+1)\mathcal{D}}| > 2(4d+1)$ *and* $|T_{(i+2)\mathcal{D}}| > 2(4d+1)$ **then**

    Remove T_j from $\mathcal{A}_1$, add T_{i+1} and T_{i+2} to $\mathcal{A}_1$ and set $i = i+2$, $p(i+1) = p(i+2) = j$ and $d(T_{i+1}) = d(T_{i+2}) = d(T_j) + 1$;

  else

    Remove T_j from $\mathcal{A}_1$ and add T_j from $\mathcal{B}_1$;

  end

end

Trimming Stage:

Set of "to be merged leaf nodes" $= \mathcal{A}_2 = \mathcal{B}_1$.

Set of "not to be merged leaf nodes" $= \mathcal{B}_2 = \phi$, the null set.

Set of "parent nodes" $= \mathcal{C}$.

while $\mathcal{A}_2 \neq \phi$ **do**

  Pick $T_j \in \mathcal{A}_2$, s.t. $d(T_j) = \max_i d(T_i)$ and j is even. Then $T_{j-1} \in \mathcal{A}_2$ and $T_{p(j)} \in \mathcal{C}$ hold;

  Use $T_{p(j)\mathcal{D}_2}$ to perform RIFT at level $\alpha/2^{2d(T_j)-1}$;

  if *reject* RIFT **then**

    Remove T_j and T_{j-1} from $\mathcal{A}_2$ and add them to $\mathcal{B}_2$;

  else

    Remove T_j and T_{j-1} from $\mathcal{A}_2$, remove $T_{p(j)}$ from $\mathcal{C}$ and add $T_{p(j)}$ to $\mathcal{A}_2$;

  end

end

The bottom-up version as shown in Algorithm 2 has two stages. A building stage and a trimming stage. In the building stage we build a binary tree such that each of the leaf nodes has at least $2(4d+1)$ data points. That is, it has enough data to estimate a mixture of two Gaussians with different diagonal covariance matrices and to perform a test. In the trimming stage, we start with the leaf nodes, and test whether their parent should have been split using RIFT. If we reject the test, we keep the leaf nodes and if we fail to reject we trim off the corresponding leaf nodes.

To demonstrate a trimming step of the bottom-down clustering process, let us say we first build the binary tree given in Figure 5.2(a). Then $\mathcal{A}_2 = \{T_5, T_7, T_8, T_9, T_{10}, T_{11}, T_{12}\}$, $\mathcal{B}_2 = \phi$ and $\mathcal{C} = \{T_0, T_1, T_2, T_3, T_4, T_6\}$. Suppose we pick $T_j = T_8$ as shown in Figure 5.2(a), then $T_{j-1} = T_7 \in \mathcal{A}_2$ and $T_{p(j)} = T_3 \in \mathcal{C}$. We estimate the truncated Gaussian ($\widehat{p_1}$) and the truncated mixture of two Gaussians ($\widehat{p_2}$) using $T_{3\mathcal{D}_1}$. We use $T_{3\mathcal{D}_1}$ to perform RIFT at level $\alpha/2^5$. As shown in Figure 5.2(b), if we fail to reject RIFT, then we trim the children of T_3 and now $\mathcal{A}_2 = \{T_3, T_5, T_9, T_{10}, T_{11}, T_{12}\}$, $\mathcal{B}_2 = \phi$ and $\mathcal{C} = \{T_0, T_1, T_2, T_4, T_6\}$.. The next T_j under consideration is either T_{10} or T_{12}. Suppose we consider $T_j = T_{12}$ and repeat all the steps as in the case of $T_j = T_8$ to perform RIFT at level $\alpha/2^5$. If in this case RIFT rejects the null, then we keep the split of T_6 into T_{11} and T_{12}. Therefore now $\mathcal{A}_2 = \{T_3, T_5, T_9, T_{10}\}$, $\mathcal{B}_2 = \{T_{11}, T_{12}\}$ and $\mathcal{C} = \{T_0, T_1, T_2, T_4, T_6\}$.. We continue this way until $\mathcal{A}_2$ is completely empty and then the final clustering is given by the nodes in $\mathcal{B}_2$.

5.2 A Sequential Approach

RIFT can also be used in a sequential model selection framework. Using $\mathcal{D}_1$ we fit a mixture of k Gaussians for $k = 1, 2, \ldots, K_n$ where K_n can be chosen to be quite large, for example, $K_n = \sqrt{n}$. Now, using $\mathcal{D}_2$, we choose k by testing a series of hypotheses. For $j = 1, 2, \ldots,$ we test the null that $\widehat{p}_j$ fits better than any $\widehat{p}_s$ for $s > j$. Formally, we test

$$H_{0j} := K(p, \widehat{p}_j) - K(p, \widehat{p}_s) \leq 0 \quad \text{for all } s > j$$

versus

$$H_{1j} := K(p, \widehat{p}_j) - K(p, \widehat{p}_s) > 0 \quad \text{for some } s > j.$$

We reject H_{0j} if

$$\max_{j < s \leq K_n} \frac{\sqrt{n}\, \widehat{\Gamma}_{js}}{\widehat{\tau}_{js}} > z_{\alpha/m_j} \tag{5.1}$$

where $m_j = K_n - j$, $\widehat{\Gamma}_{js} = \frac{1}{n}\sum_{i \in \mathcal{D}_2} R_i$, $R_i = \log\left(\widehat{p}_s(X_i)/\widehat{p}_j(X_i)\right)$ and $\widehat{\tau}^2_{js} = \frac{1}{n}\sum_{i \in \mathcal{D}_2}(R_i - \overline{R})^2$.

Let $\widehat{k}$ be the first value of j for which H_{0j} is not rejected. We then use $\widehat{p}_{\widehat{k}}$ to define the clusters. Notice that, unlike procedures like AIC or BIC, this method provides a valid, asymptotic, type I error control.

Lemma 5.2.1. *Under H_{0j},*

$$\limsup_{n\to\infty} \mathbb{P}(\text{rejecting } H_{0j}) \leq \alpha. \tag{5.2}$$

This follows from the results in Section 3. Of course, Γ can be replaced with the ℓ_2 version or the median version.

Chapter 6

Experimental Performance of the Tests

6.1 Simulations

In this section we compare SigClust and the RIFT variants we proposed through a variety of simulations. In Section 6.1.1 we investigate the asymptotic normality of the RIFT statistic defined in (4.3) under the null. In Section 6.1.2 we compare the power of various tests for detecting and splitting a mixture of two Gaussians. Finally, in Sections 6.1.3 and 6.1.4 we study hierarchical clustering using the RIFT statistic and evaluate model selection using the sequential RIFT procedure.

6.1.1 Asymptotic Normality of the RIFT Test Statistic

In this section we check if the distribution of the RIFT test statistic is indeed Normal as claimed in Theorem 4.1. We explore four simulated data sets and use Q-Q plots to check for Normality. For the four examples, we generate data from the following distributions:

1. $0.5N(\mu, \mathbf{I}_d) + 0.5N(-\mu, \mathbf{I}_d)$, with $d = 2$, $n = 1000$ and $\mu = (2, 0)$.
2. A mixture of two uniform distributions over rectangles given by, 0.5 Unif($[-2, -1] \times [0, 1]$) + 0.5 Unif($[2, 3] \times [0, 1]$), with $d = 2$ and $n = 1000$.
3. $0.5N(\mu, \mathbf{I}_d) + 0.5N(-\mu, \mathbf{I}_d)$, with $d = 1000$, $n = 1000$ and $\mu = (10, 0, \ldots, 0)$.
4. A single Gaussian distribution, $N(\mathbf{0}, \Sigma)$, where $\Sigma_{11} = 100$ and $\Sigma_{jj} = 1$ for $j = 2, \ldots, d$.

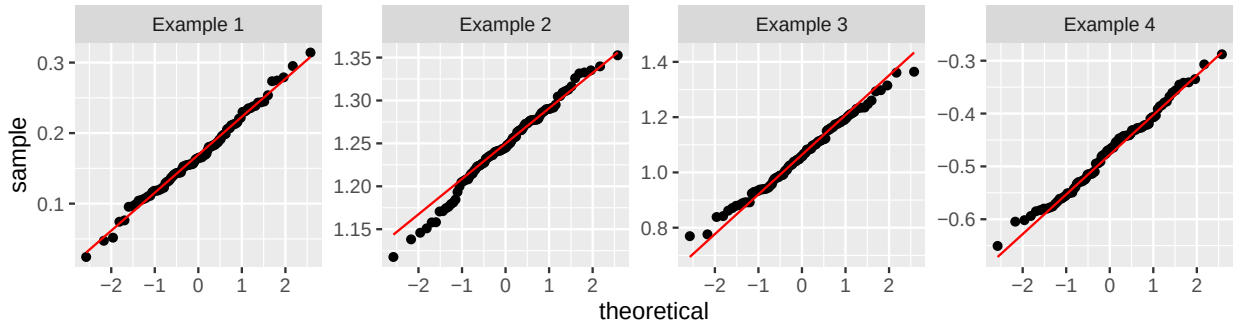


Figure 6.1: Q-Q plots to check Normality of the RIFT test statistic.

The test statistic, $\widehat{\Gamma}$ defined in (4.3) is computed for 100 simulations in each of these cases. Figure 6.1 provides the Q-Q plots of the test statistic. We notice that all of them are close to Normal, confirming the result in Theorem 4.1.

6.1.2 Comparing the Different Tests for Mixtures of Two Gaussians

We first consider data generated from a collection of mixture of two Gaussians, $0.5N(\mu, \mathbf{I}_d) + 0.5N(-\mu, \mathbf{I}_d)$, where $\mu = (a, 0, \ldots, 0)$, with varying distances (varying a) between their means. We compare the power of the RIFTs, SigClust, Mardia's Kurtosis Test and two versions of Zhou's Nearest Neighbour tests in detecting the two clusters. Mardia's Kurtosis Test and the two versions of Zhou's Nearest Neighbour tests are described in detail in Appendix 4.5. Specifically, we compare the number of times the tests correctly reject the null hypothesis that the data is generated from a single (Gaussian) cluster.

First, we compare the effect of varying the number of observations (n) for the different tests. We run 100 simulations where we generate observations from a mixture of two 2D Gaussian distributions given by, $0.5N(\mu, \mathbf{I}_d) + 0.5N(-\mu, \mathbf{I}_d)$, with $d = 2$ and $\mu = (2, 0)$. Figure 6.2 gives the proportion of tests that reject the null hypothesis that the underlying distribution has just one cluster at level $\alpha = 0.05$. We see that M-RIFT and SigClust have comparable power, and that they have higher power than the other tests. We also notice that Mardia's Kurtosis test and RIFT have comparable power, but they do not perform as well as SigClust or M-RIFT.

Next we vary the value of a and see how increasing or decreasing the distance between the two distributions changes the ability of the tests to reject. We fix $n = 1000$. Figure 6.2 compares the proportion of times the tests detect the two distributions at $\alpha = 0.05$, as we vary the distance between them. Notice that Mardia's Kurtosis test and both the RIFTs perform better than SigClust in this case. In particular, they detect the two clusters for smaller values of a when compared to SigClust. Also notice that SigClust does not detect the presence of the two clusters at all when the distance between the two clusters is ≤ 1.5. For the rest of our simulations, we consider comparisons between the RIFTs and SigClust.

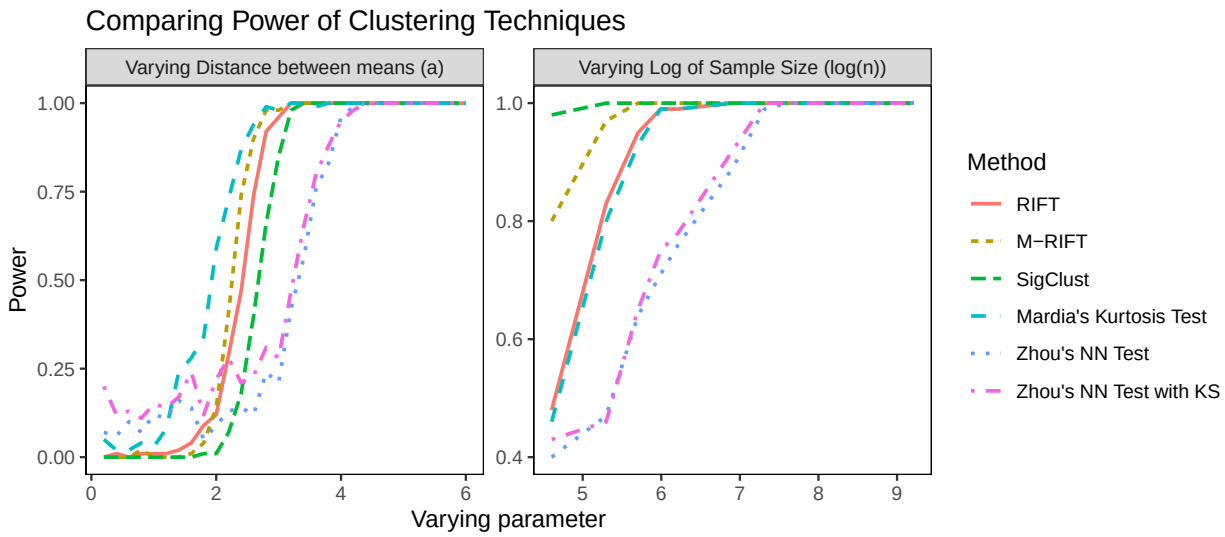


Figure 6.2: Comparing the power of the tests with increasing distance between the two mixture distributions (increasing a) and varying the total number of observations, n in terms of $\log(n)$.

Signal in One Direction

We compare the power of our test with SigClust while checking whether the tests control the type-I error at $\alpha = 0.05$. We consider a mixture of two normal distributions, $0.5N(0, \Sigma)+0.5N(\mu, \Sigma)$, where $\mu = (a, 0, \ldots, 0)$ with $a = 0, 10, 20$ and $\Sigma = \mathbf{I}_d$. The sample size is $n = 500$, we use 450 points to estimate the Gaussian mixture parameters and 50 points to test the hypothesis. The dimension is $d = 1000$. Notice that when $a = 0$, the distribution reduces to a single Gaussian distribution and as we take larger a, the signal strength grows. The empirical distributions of p-values, after 30 realizations of the experiment, for RIFT, Median RIFT (M-RIFT) and SigClust are shown in Figure 6.3. We notice that the SigClust has very good power for both $a = 10$ and $a = 20$, whereas the RIFTs catch up for $a = 20$.

Signal in All Directions

Now we consider data with signal in all directions and compare the tests at $\alpha = 0.05$. We consider a mixture of two normal distributions, $0.5N(0, \Sigma) + 0.5N(\mu, \Sigma)$, where $\mu = (a, a, \ldots, a)$ with $a = 0, 0.5, 0.7$ and Σ is a diagonal matrix with $\Sigma_{11} = 100$ and $\Sigma_{jj} = 1$ for $j = 2, \ldots, d$. We consider a high-dimensional setting where the sample size is chosen to be $n = 100$ and the dimension is $d = 1000$. Figure 6.4 shows the p-values generated by each of the tests. In this case, we see that all the tests perform similarly well. SigClust performs only slightly better than the RIFTs.

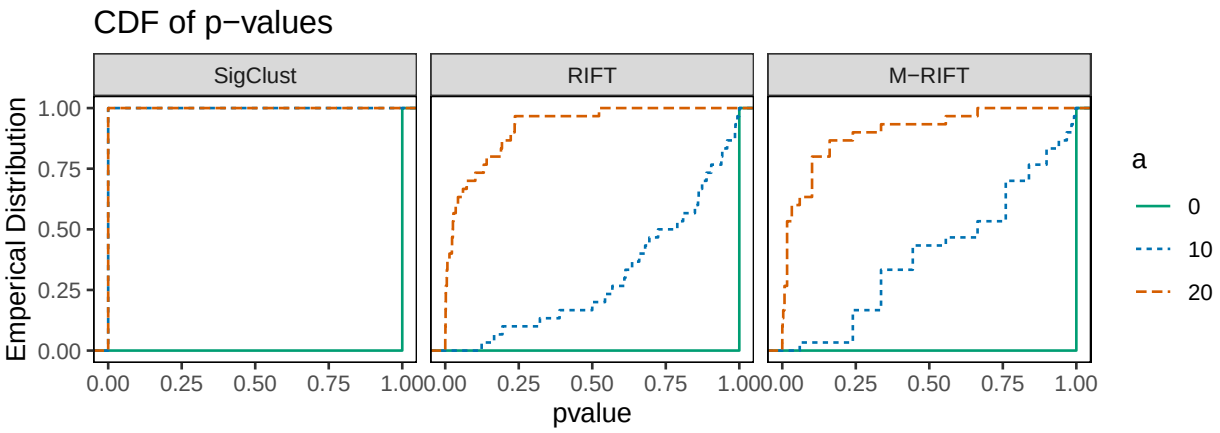


Figure 6.3: Comparing the empirical distribution of the p-values when signal is exactly in one direction.

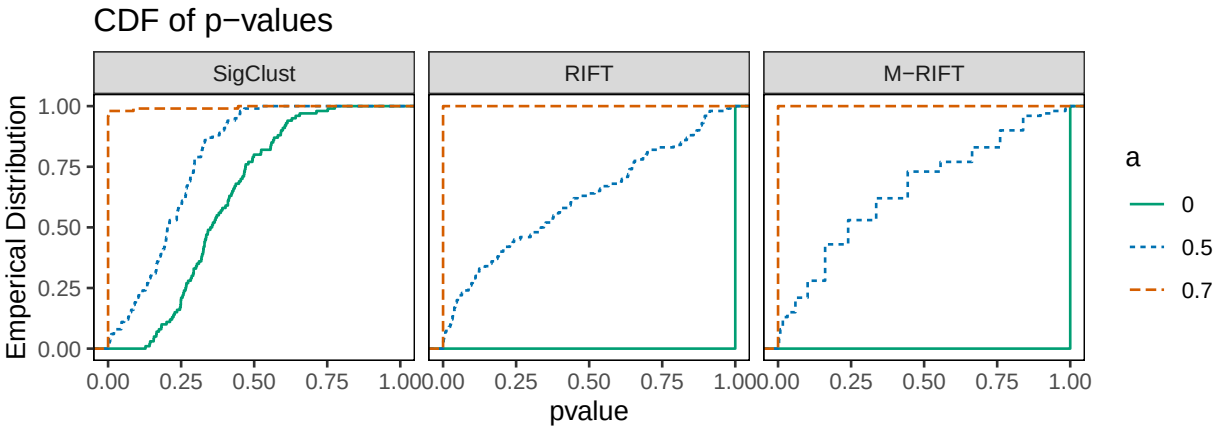


Figure 6.4: Comparing the empirical distribution of the p-values when signal is in all directions.

Example where SigClust Fails

Finally, we compare the power of the RIFTs with SigClust and Mardia's Kurtosis test in detecting the signal in one direction if the variability in another direction is very high, at $\alpha = 0.05$. We consider a mixture of two normal distributions, $0.5N(0, \Sigma) + 0.5N(\mu, \Sigma)$, where $\mu = (a, 0, \ldots, 0)$ with $a = 0, 10, 20$ and Σ is a diagonal matrix with $\Sigma_{jj} = 400$ for $j = 2$ and $\Sigma_{jj} = 1$ for $j \neq 2$. That is, we are trying to detect the signal in the first dimension while the variability in the second dimension is very high. The sample size is $n = 100$ and dimension is $d = 5$.

The empirical distributions of the p-values are shown in Figure 6.5. We notice that SigClust has almost no power in detecting the signal in one direction when there is high variability in any other direction, whereas both the RIFTs have high power while controlling the type-I error. Mardia's Kurtosis test also has higher power than SigClust but has lower power than the RIFTs.

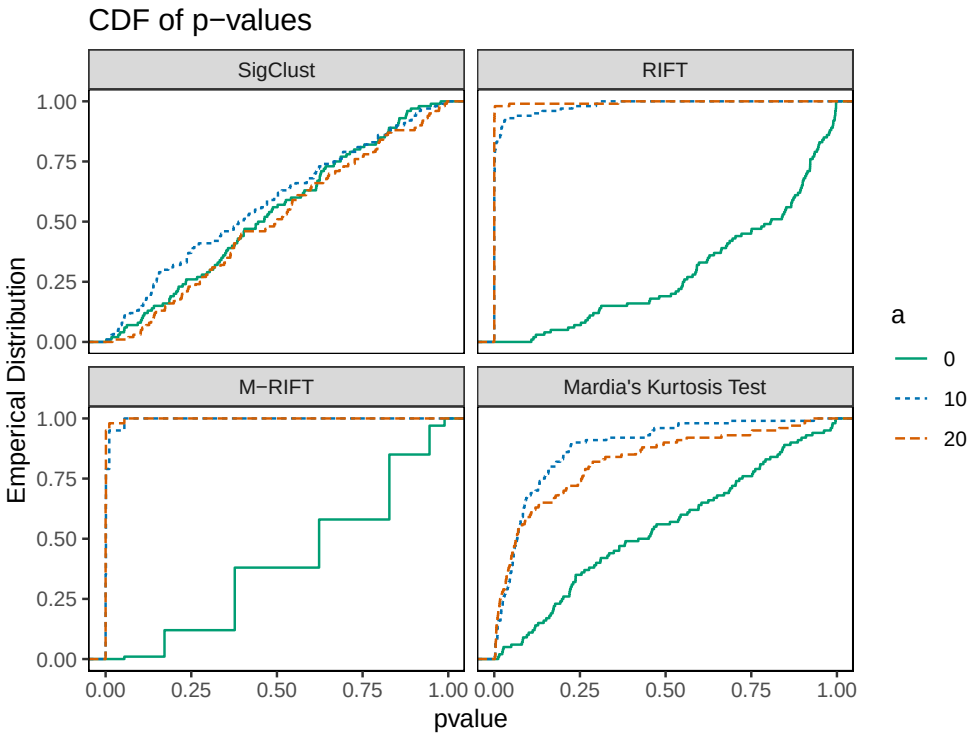


Figure 6.5: Comparing the power of the tests with higher signal in one direction and high variability in another.

6.1.3 Hierarchical Clustering Example: Four Cluster Setting (K = 4)

In this section, we compare the tests in a hierarchical setting. We compare the RIFTs, SigClust and truncated SigClust, where for the SigClusts the clustering is performed using k-means clustering with $k = 2$ and mixture of Gaussians is used in the case of the RIFTs. We consider the alternative setting in which observations are drawn from a mixture of four clusters, each of which is a Gaussian distribution with covariance matrix $\Sigma = I_d$. Our motive is to study how the different tests behave at each split in a hierarchical setting.

We compare the methods for two arrangements of the four Gaussian components. In the first setting, the four components are placed at the vertices of a square with side length δ and in the second setting, the four components are placed at the vertices of a regular tetrahedron with side length δ. 50 samples were drawn from each of the Gaussian components for 100 simulations. We control the overall type I error for all the methods at $\alpha = 0.05$.

For each of the simulations, we use the four tests RIFT, M-RIFT, SigClust and truncated SigClust in the hierarchical setting and record the number of clusters given by each. Table 6.1 gives the simulation results for some values of d and δ. We notice that M-RIFT performs better than the other tests in all the experiments. We also notice that the top-down hierarchical algorithms tend to give more clusters than the bottom-up hierarchical algorithms. In the case of RIFT and M-RIFT, we notice that the top-down algorithms identify

Table 6.1: Measuring performance of hierarchical algorithms on a mixture of 4 Gaussians using the no. of simulations (out of 100) that give a particular number of significant clusters.

Method	Algorithm type	Parameters			Number of clusters					
		d	δ	arr.	1	2	3	**4**	5	≥ 6
RIFT	Top-down	2	6	square	1	3	22	**74**	0	0
M-RIFT					0	0	4	**96**	0	0
SigClust					16	0	0	**43**	13	28
Trunc. SigClust					16	0	0	**46**	14	24
RIFT	Bottom-up	2	6	square	10	8	29	**53**	0	0
M-RIFT					0	0	10	**90**	0	0
SigClust					77	0	0	**20**	3	0
Trunc. SigClust					57	0	1	**30**	12	0
RIFT	Top-down	3	4	tetrahedral	1	5	27	**67**	0	0
M-RIFT					0	0	5	**95**	0	0
SigClust					86	0	1	**5**	1	7
Trunc. SigClust					82	2	0	**7**	2	7
RIFT	Bottom-up	3	4	tetrahedral	9	13	40	**38**	0	0
M-RIFT					0	1	24	**75**	0	0
SigClust					58	0	24	**8**	10	0
Trunc. SigClust					50	0	23	**12**	15	0
RIFT	Top-down	3	5	tetrahedral	0	0	9	**91**	0	0
M-RIFT					0	0	0	**100**	0	0
SigClust					71	0	0	**7**	4	18
Trunc. SigClust					72	2	0	**8**	5	13
RIFT	Bottom-up	3	5	tetrahedral	0	0	27	**73**	0	0
M-RIFT					0	0	1	**99**	0	0
SigClust					54	0	29	**7**	10	0
Trunc. SigClust					48	0	26	**11**	15	0

Table 6.2: Comparing the different algorithms for selecting the ideal number of clusters when samples are generated from a mixture of four Gaussian distributions. $n = 400$ and $K_n = \sqrt{n} = 20$. The table gives the number of simulations that identify the particular number of significant clusters over 100 replications.

Method	Parameters			Number of clusters					
	d	δ	arr.	1	2	3	**4**	5	≥ 6
S-RIFT (KL)	10	6	Tetrahedral	0	0	32	**68**	0	0
S-RIFT (ℓ_2)				60	22	11	**7**	0	0
AIC				0	0	46	**54**	0	0
BIC				1	41	58	**0**	0	0
S-RIFT (KL)	10	10	Tetrahedral	0	0	5	**93**	2	0
S-RIFT (ℓ_2)				55	16	25	**4**	0	0
AIC				0	0	7	**93**	0	0
BIC				0	0	99	**1**	0	0
S-RIFT (KL)	20	80	Tetrahedral	0	4	86	**10**	0	0
S-RIFT (ℓ_2)				94	5	1	**0**	0	0
AIC				0	7	93	**0**	0	0
BIC				1	99	0	**0**	0	0

four as the correct number of clusters more often than the bottom-up algorithms. In general, M-RIFT and RIFT identify four as the number of significant clusters present, more often than SigClust or truncated SigClust.

6.1.4 Sequential RIFT

Now we compare the proposed sequential model selection approach (Sequential RIFT or S-RIFT) to AIC and BIC. We use two versions of the model selection approach - one using the Kullback-Leibler distance and one using the ℓ_2 distance between the estimated and the true densities. Using two simulated experiments, we compare these methods to using AIC and BIC.

We reconsider the four cluster example used in the hierarchical clustering setting where the four components are placed at the vertices of a regular tetrahedron with side length δ. 100 samples are drawn from each of the four Gaussian components ($n = 400$) for 100 simulations. For each simulation, we use S-RIFT with the two different distances - Kullback-Leibler distance and ℓ_2 distance with $K_n = \sqrt{n} = 20$ and record the number of clusters given by them. We also record the number of clusters that give the minimum AIC and BIC for each simulation. Table 6.2 gives the results of the simulations when the overall type I error is controlled at $\alpha = 0.05$.. We notice that S-RIFT using Kullback-Leibler distance out-performs all the other methods. AIC performs very similar to it for $d = 10$ and $\delta = 10$, but we notice that for $d = 20$ and $\delta = 80$, S-RIFT using Kullback-Leibler distance is the only one that detects the four clusters for some simulations.

To further explore the properties of Sequential RIFT, we also study a simulation with 10 clusters. We generate n data points from 10 Gaussian components with means given by:

$$\mu_1 = (\mathbf{a}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}), \qquad \mu_6 = (-\mathbf{a}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}),$$
$$\mu_2 = (\mathbf{0}, \mathbf{a}, \mathbf{0}, \mathbf{0}, \mathbf{0}), \qquad \mu_7 = (\mathbf{0}, -\mathbf{a}, \mathbf{0}, \mathbf{0}, \mathbf{0}),$$
$$\mu_3 = (\mathbf{0}, \mathbf{0}, \mathbf{a}, \mathbf{0}, \mathbf{0}), \qquad \mu_8 = (\mathbf{0}, \mathbf{0}, -\mathbf{a}, \mathbf{0}, \mathbf{0}),$$
$$\mu_4 = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{a}, \mathbf{0}), \qquad \mu_9 = (\mathbf{0}, \mathbf{0}, \mathbf{0}, -\mathbf{a}, \mathbf{0}),$$
$$\mu_5 = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{a}), \qquad \mu_6 = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, -\mathbf{a}),$$

where $\mathbf{a} = (a, a, \ldots, a)$ and $\mathbf{0} = (0, 0, \ldots, 0)$ are vectors of length $p = d/5$. Each Gaussian component has mean 0 and variance σ^2. We generate $n/10$ data points from each of the Gaussians.

We consider dimensions $d = 30$, so $p = 6$ and consider two values of $n = 1000, 1500$. We vary the distance between the means by considering two values of $a = 200, 500$ and consider three variances $\sigma^2 = 0.001, 0.04, 0.16$. For each of the three variances, we simulate 100 samples and record the number of clusters given by S-RIFT, AIC and BIC. In each case, the largest number of possible clusters is taken to be $K_n = \sqrt{n}$ and the overall type I error is controlled at $\alpha = 0.05$.

The estimates of the number of clusters given by S-RIFT, AIC and BIC are recorded in Table 6.3. We notice that in every case S-RIFT using Kullback-Leibler distance outperforms all the other methods. AIC performs the next best. We notice that both S-RIFT using ℓ_2 loss and BIC tend to under estimate the number of clusters.

6.1.5 Summary of the Simulations

For two clusters which are separated in just one of the dimensions, if the variance in the other dimensions isn't too large, SigClust out-performs all the other methods for small sample sizes. RIFT and Mardia's Kurtosis Test show comparable results. But when the distance between the clusters is small, or when the variance in some other dimension is much larger than the separation, SigClust loses power completely and RIFT and Mardia's Kurtosis Test out-perform SigClust. We also observe that as the dimension increases, RIFT has lower power than the SigClust.

For the simulated examples that have more than two clusters, hierarchical clustering using RIFT detects the true number of clusters much better than hierarchical clustering using SigClust. Finally, we notice that using S-RIFT to detect the correct number of clusters is better than minimizing the AIC or BIC. We also see that the version using the Kullback-Leibler distance out-performs the one using ℓ_2 distance, which tends to under-estimate the number of clusters.

Table 6.3: Comparing the different algorithms for selecting the ideal number of clusters when samples are generated from a mixture of 10 Gaussian distributions. The entries of the table give the numbers of simulations (out of a total of 100) for which a certain estimate of the number of clusters is obtained.

Method	Parameters			Number of clusters					
	n	a	σ^2	≤ 5	6	7	8	9	**10**
S-RIFT (KL)	1000	200	0.001	0	1	0	45	54	**0**
S-RIFT (ℓ_2)				100	0	0	0	0	**0**
AIC				0	1	1	52	46	**0**
BIC				13	3	46	38	0	**0**
S-RIFT (KL)	1000	200	0.04	0	2	7	71	20	**0**
S-RIFT (ℓ_2)				100	0	0	0	0	**0**
AIC				2	0	7	84	7	**0**
BIC				100	0	0	0	0	**0**
S-RIFT (KL)	1000	200	0.16	1	2	21	65	11	**0**
S-RIFT (ℓ_2)				100	0	0	0	0	**0**
AIC				3	0	22	75	0	**0**
BIC				100	0	0	0	0	**0**
S-RIFT (KL)	1500	200	0.001	0	0	0	0	22	**78**
S-RIFT (ℓ_2)				96	3	1	0	0	**0**
AIC				0	0	0	0	67	**33**
BIC				0	0	0	0	100	**0**
S-RIFT (KL)	1500	200	0.04	0	0	0	0	72	**28**
S-RIFT (ℓ_2)				93	2	0	5	0	**0**
AIC				0	0	0	0	100	**0**
BIC				0	0	0	77	23	**0**
S-RIFT (KL)	1500	200	0.16	0	0	0	0	92	**8**
S-RIFT (ℓ_2)				95	2	3	0	0	**0**
AIC				0	0	0	0	100	**0**
BIC				0	0	0	100	0	**0**
S-RIFT (KL)	1500	500	0.001	0	0	0	0	8	**92**
S-RIFT (ℓ_2)				96	3	1	0	0	**0**
AIC				0	0	0	0	31	**69**
BIC				0	0	0	0	100	**0**
S-RIFT (KL)	1500	500	0.04	0	0	0	0	45	**55**
S-RIFT (ℓ_2)				94	1	0	5	0	**0**
AIC				0	0	0	0	95	**5**
BIC				0	0	0	4	96	**0**
S-RIFT (KL)	1500	500	0.16	0	0	0	0	73	**27**
S-RIFT (ℓ_2)				95	2	3	0	0	**0**
AIC				0	0	0	0	97	**3**
BIC				0	0	0	55	45	**0**

Table 6.4: Clusterings given by RIFT and SigClust for the multi-cancer gene expression data set.

True Classes	RIFTs Classes		
	HNSC	LUSC	LUAD
HNSC	79	21	0
LUSC	7	70	23
LUAD	0	1	99

True Classes	SigClust Classes		
	HNSC	LUSC	LUAD
HNSC	90	10	0
LUSC	4	74	22
LUAD	0	1	99

6.2 Application to Gene Expression Data

To further compare the power of the RIFTs to the power of the SigClusts in the hierarchical setting, we apply the approach to a cancer gene expression data set. We consider a data set consisting of three different cancer types - head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). Since we have samples from three distinctively different cancers, we expect the methods to be able to detect the presence of three different clusters. We compare the clusterings given by hierarchical RIFT and M-RIFT with hierarchical SigClust at level $\alpha = 0.05$.

We combine data on 100 tumor samples from each of HNSC, LUSC and LUAD to create a data set of 300 samples, similar to Kimes et al. (2017). The data is obtained from The Cancer Genome Atlas (TCGA) project (Network et al., 2012, 2014) whose RNA sequence data v2 is available at `https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2`. We used the R package TCGA2STAT (Wan et al., 2015) to download the TCGA data into a format that can be directly used for our statistical analysis.

There are a total of 20,501 genes of which we use the 500 genes that have the highest median absolute deviation (MAD) about the median. To scale the data appropriately, we consider a log-transformation of the data. In order to do so, first we replace all expression values that are zero with the smallest non-zero expression value for all genes over the data and then take a log-transformation.

SigClust was implemented with 1000 simulations at every node. The top-down and the bottom-up versions of both RIFT and M-RIFT correctly give 3 clusters. The top-down version of SigClust gives 9 clusters and the bottom-up version gives 5 clusters. All the algorithms first create a split between LUAD and the other two cancers and then the next split separates HNSC and LUSC. Table 6.4 gives the clusterings given by the first two splits for the RIFTs and SigClust. Note that even though SigClust gives better clusters, it splits all the clusters further into smaller clusters.

Hence, similar to the simulations with multiple clusters in Section 6.1.3, in this case also hierarchical clustering using RIFT detects the true number of clusters much better than hierarchical clustering using SigClust.

Part III

Inference for Anomaly Detection: Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests

Chapter 7

Introduction to Model-Independent Detection of New Physics Signals

Statistical and machine learning tools have been extensively used over the past few decades to answer fundamental questions (Bhat, 2011) such as: What are the basic building blocks of our universe? What are the fundamental forces of nature? Is there a greater underlying symmetry in our universe?

To answer these fundamental questions, one needs to experimentally test the predictions of the Standard Model, which describes our current understanding of fundamental particles and how they interact with each other. These tests can lead to discoveries of new particles and the development of new theories that can better describe our universe. For example, the recent empirical confirmation of the Higgs boson was an essential step towards its inclusion in the Standard Model (Aad et al., 2012; Chatrchyan et al., 2012).

In experiments conducted within large particle accelerators, e.g., the Large Hadron Collider (LHC), the searches for new physics signals have traditionally been conducted using fully supervised model-dependent data analysis methods. These searches are generally structured as a likelihood ratio test based on a model assumption for the specific new signal that is being searched for (Williams, 2010; Cowan et al., 2011; ATLAS Collaboration and CMS Collaboration, 2011). Supervised, multivariate classification algorithms such as neural networks and boosted decision trees have demonstrated an excellent performance in increasing the signal-to-background ratio in searches. Therefore, they have been successfully used to separate the signal events from the background events. Additionally, the classifier output is used to perform the likelihood ratio tests for the detection of the signal (Aad et al., 2012; Chatrchyan et al., 2012).

In general, in this approach, the training signal samples for the classifier are generated using a Monte Carlo (MC) event generator based on conjectured physics models. Hence the classifier relies very heavily on these simulations. There are multiple disadvantages to this approach. Firstly, this approach cannot be used

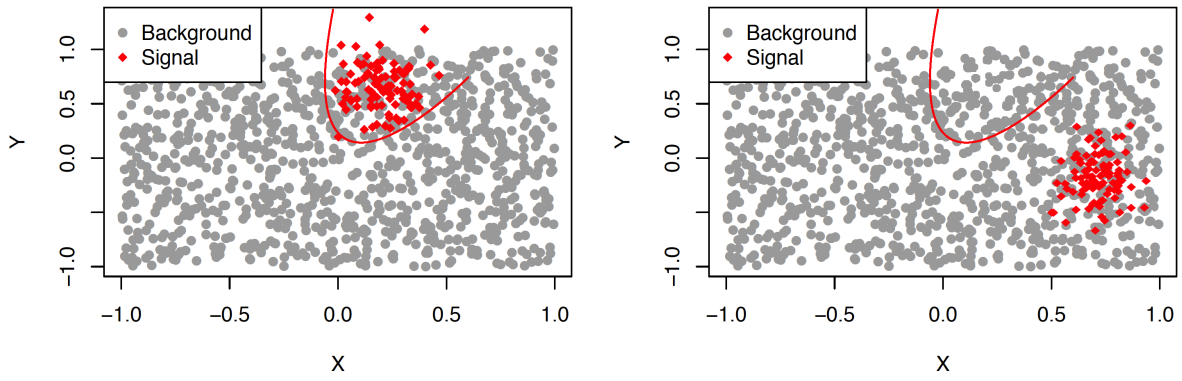


Figure 7.1: Decision boundary using a supervised classifier to separate the signal (red) from the background (grey). (a) The boundary of the classifier when trained on signal generated from the assumed signal model. (b) When there is a systematic error in the training signal data, the test completely misses the actual signal data.

to search for signals that we are not specifically looking for or have not considered yet. Secondly, systematic errors can be influential on supervised classifiers and so any error or imprecision in the signal model will adversely affect the method. Figure 7.1 illustrates the problem more clearly. If a classifier is trained on training signal data as shown in Figure 7.1(a), it gives the classification boundary as shown. But what if the signal data actually looks like Figure 7.1(b)? Then the classifier ends up misclassifying the signal as background. So an algorithm trained on a wrong signal model might completely miss the actual signal.

In contrast, in this thesis, we propose tests to search for new physics signals in a model-independent fashion, without assuming any model for the signal. Specifically, we search in the experimental data (as collected from particle detectors) for any signal that deviates from the background process (as explained by the Standard Model). The proposed methods are based on the assumption that there exists an accurate representation of the background, i.e., a sample of particle collisions containing no signal events. In most cases, the background is simulated using MC simulations, though in some cases it might be possible to additionally use real measurements.

Since we do not use a signal sample instead of having three collections of data, we only have two. The first data set is a labelled sample generated from the background process and the second data set is an unlabelled experimental sample as observed from the particle detectors. The experimental sample is assumed to be drawn from an unknown distribution, which is a mixture of the background distribution and possibly a signal distribution. We then use a semi-supervised approach that trains a classifier to differentiate the background data from the experimental data.

We use the trained classifier to propose two different tests to detect the presence of signal in the experimental data. The first test is based on a likelihood ratio test statistic that is estimated using the

classifier output. The second test is based on the performance of the classifier measured using the area under the curve test statistic. Both the tests are based on the argument that in the absence of signal events in the experimental data, a classifier should not be able to differentiate the experimental data from the background data.

We further propose using active subspace methods (Constantine, 2015) to identify the characteristics of variables and their dependencies on each other that separate the background data from the experimental data, affecting the outcome of the classifier. This can be used to characterize the signal region in the experimental data.

The advantage of model-independent tests is that they can detect any discrepancy between the background data and the experimental data independent of the distribution of the signal events. In the case that a signal is found in the experimental data, the signal should be investigated further in order to decide if it results from (a) an inaccurate background MC generator, or (b) a particle detector defect or a lack of understanding of the detector, or (c) a previously unknown physics process.

Due to their advantages, model-independent approaches have been used for new physics searches at the Tevatron (Aaltonen et al., 2009; Bertram et al., 2012), HERA (Aktas et al., 2004), and the LHC (CMS Collaboration, 2017; Aaboud et al., 2019). These methods typically compare a large set of binned distributions to the prediction from the background Monte Carlo simulation, in search for bins in the experimental data that exhibit a deviation larger than some predefined threshold. For example, Aaboud et al. (2019) employed by the ATLAS Collaboration uses a (quasi-)model-independent method that uses some generic features of the potential new physics signals. These approaches have two problems: (a) they do not consider the dependency structures between the variables in the data and (b) they might miss certain signals that do not show a localized excess in one of the studied distributions.

Casa and Menardi (2018), on the other hand, use a semi-supervised nonparametric clustering algorithm under the assumption that in high energy physics, a new particle manifests itself as a significant peak emerging from the background process. They use nonparametric modal clustering to search for a signal that is expected to emerge as a bump in the background distribution. This method suffers from the second problem as mentioned above, i.e., it might miss certain signals that do not show a localized excess.

Model-independent semi-supervised searches were also proposed by Kuusela et al. (2012) and Vatanen et al. (2012) who use multivariate Gaussian mixture models to estimate the densities of the background and the experimental data. They first model the background using a multivariate Gaussian mixture model. They then fit a mixture of this background model and a number of additional Gaussians to the experimental data. The test of deviation of the experimental data from the background is performed by testing for the significance of the additional Gaussian components which quantify the anomalous contribution. The drawback of this method is that Gaussian mixture models are very difficult to fit in the high-dimensional

setting. Additionally, since the signal strength is typically very low, the quality of the fit influences the power of the test in detecting the signal.

As mentioned earlier, classification algorithms have demonstrated an excellent performance in detecting signals in the model-dependent approaches. This motivates us to use classifiers in order to find the deviations of the experimental data from the background. The trained multivariate classifiers inherently model the dependency structures between the variables in the data. This approach is better than the mixture modelling methods because classifiers tend to work better in high-dimensional spaces. The proposed methods make no assumptions about the signal at all. By keeping the methods free of any signal model assumptions, we are more likely to detect any kind of unpredictable new signal as well as be unaffected by inaccuracies in the MC signal modelling.

7.1 Organization of Part III

In the following chapter, Chapter 8, we first introduce the problem setup mathematically. We then describe comparable supervised methods in Section 8.2. The proposed model-independent semi-supervised methods are introduced in Section 8.3. We introduce both tests based on the likelihood ratio test statistic and the area under the curve (AUC) test statistic. In Section 8.5 we describe active subspace methods to understand the subspace affecting the classifier the most, leading to an understanding of the signal region. Finally in Chapter 9, we demonstrate the performance of the proposed methods and compare them to the supervised approaches as well as nearest neighbor two-sample tests introduced in Schilling (1986) and Henze (1988).

Chapter 8

Anomaly Detection Algorithms

Let the data be denoted as:

$$\text{Experimental:} \quad W_1, \ldots, W_N \tag{8.1}$$

$$\text{Background:} \quad X_1, \ldots, X_m \tag{8.2}$$

$$\text{Signal:} \quad Y_1, \ldots, Y_n \tag{8.3}$$

where the experimental data $W_1, \ldots, W_N$ is generated from an inhomogeneous Poisson point process with intensity function

$$\nu(w) = b(w) + \mu s(w) \tag{8.4}$$

where $b(\cdot)$ is the intensity function of the background process, $s(\cdot)$ is the intensity function of the signal process and $\mu \geq 0$ is the signal strength modifier. Let $B = \int b(w)dw$ be the expected background event rate and $S = \int s(w)dw$ be the expected signal event rate. Define $p_b(w) = b(w)/B$ and $p_s(w) = s(w)/S$. Then the background data $X_1, \ldots, X_m$ and the signal $Y_1, \ldots, Y_n$ are auxiliary samples generated from

$$Y_1, \ldots, Y_n \sim p_s$$

$$X_1, \ldots, X_m \sim p_b.$$

The goal is to test $H_0 : \mu = 0$.

The unbinned likelihood for the experimental data is

$$\mathcal{L}(\mu) = e^{-(B+\mu S)} \prod_i (Bp_b(W_i) + \mu S p_s(W_i)). \tag{8.5}$$

We then have that

$$\frac{\mathcal{L}(\mu)}{\mathcal{L}(0)} = e^{-\mu S} \prod_i \left[1 + \frac{\mu S}{B}\psi(W_i)\right] \tag{8.6}$$

where $\psi(w) = p_s(w)/p_b(w)$.

Remark. Even though p_b and p_s are functions of the high-dimensional vector w, if ψ were known then the ratio (8.6) is one-dimensional in the sense that $\psi(w) \in \mathbb{R}$.

The binned likelihood is

$$\mathcal{L}(\mu) = \prod_j \frac{(b_j + \mu s_j)^{n_j}}{n_j!} e^{-(b_j + \mu s_j)} \tag{8.7}$$

where n_j is the count in bin j, $b_j = \int_{\Omega_j} b(w)dw$, $s_j = \int_{\Omega_j} s(w)dw$ and Ω_j denotes bin j. Then

$$\frac{\mathcal{L}(\mu)}{\mathcal{L}(0)} = \prod_j \left(1 + \frac{\mu s_j}{b_j}\right)^{n_j} e^{-\mu s_j} = \prod_j (1 + \mu\psi_j)^{n_j} e^{-\mu s_j} \tag{8.8}$$

were $\psi_j = s_j/b_j$.

For the rest of this part of the thesis, we assume that we know the total number of events N in the experimental data and condition on N. Then the likelihood, conditional on the total number of events N is

$$\mathcal{L}_{\text{mix}}(\lambda) = \prod_i [(1-\lambda)p_b(W_i) + \lambda p_s(W_i)] \tag{8.9}$$

where

$$\lambda = \frac{\mu S}{B + \mu S}.$$

Since, testing $\mu = 0$ is equivalent to testing $\lambda = 0$, the goal is to test $H_0 : \lambda = 0$. We have that

$$\frac{\mathcal{L}_{\text{mix}}(\lambda)}{\mathcal{L}_{\text{mix}}(0)} = \prod_i [(1-\lambda) + \lambda\psi(W_i)] \tag{8.10}$$

only depends on ψ and not on S and B.

Remark. This connects the Poisson model to the mixture density model.

8.1 Idealized Case

Suppose that the functions b and s are known. Then we can test H_0 using the usual LRT, namely, $T = -2\log \mathcal{L}_{\text{mix}}(0)/\mathcal{L}_{\text{mix}}(\widehat{\lambda})$.

The null distribution can be obtained exactly by simulating under H_0. Of course we can test $H_0 : \lambda = \lambda_0$ for any λ_0. Inverting this test gives a confidence interval for λ.

Remark. An alternative to the LRT is the score test

$$\sum_i \frac{p_s(W_i)}{p_b(W_i)} - N$$

which is asymptotically Normal with mean 0 under H_0. This has the advantage that there is no need to estimate λ for the test.

8.2 Model Dependent (Supervised) Case

In this case we assume that additionally, we observe auxiliary samples

$$Y_1, \ldots, Y_n \sim p_s$$
$$X_1, \ldots, X_m \sim p_b.$$

The strategy is: use a classifier to estimate $\psi = p_s/p_b$ then use the test based on (8.10). Note that, in practice, p_s and p_b depend on nuisance parameters θ.

First, we combine samples

$$Z_1, \ldots, Z_{n+m} = X_1, \ldots, X_m, Y_1, \ldots, Y_n$$

and we define $S_i = 1$ if Z_i is from the signal distribution and $S_i = 0$ otherwise. We train a classifier $h(z)$ that separates the signal from the background.

$$h(z) = \widehat{\mathbb{P}}\left(S = 1 | Z = z\right). \tag{8.11}$$

Now,

$$\begin{aligned}
\mathbb{P}\left(S = 1 | Z = z\right) &= \frac{\mathbb{P}\left(Z = z | S = 1\right)\mathbb{P}\left(S = 1\right)}{\mathbb{P}\left(Z = z | S = 1\right)\mathbb{P}\left(S = 1\right) + \mathbb{P}\left(Z = z | S = 0\right)\mathbb{P}\left(S = 0\right)} \\
&= \frac{np_s(z)}{np_s(z) + mp_b(z)} \\
&= \frac{n\psi(z)}{n\psi(z) + m}.
\end{aligned}$$

Therefore, an estimate of $\psi = p_s/p_b$ is given by

$$\widehat{\psi}(z) = \frac{mh(z)}{n(1 - h(z))}. \tag{8.12}$$

We can plug this into equation (8.10) and find the maximum likelihood estimate of λ as

$$\widehat{\lambda} = \arg\max_{\lambda} \sum_i \log\left((1-\lambda) + \lambda\widehat{\psi}(W_i)\right), \tag{8.13}$$

which can be computed by Newton's method.

Then as $n \to \infty$,

$$-2\log\Lambda = -2\log\left(\frac{\mathcal{L}_{\text{mix}}(0)}{\mathcal{L}_{\text{mix}}(\widehat{\lambda})}\right) = 2\sum_i \log\left((1-\widehat{\lambda}) + \widehat{\lambda}\widehat{\psi}(W_i)\right) \overset{d}{\rightsquigarrow} \frac{1}{2}\delta_0 + \frac{1}{2}\chi_1^2, \tag{8.14}$$

where δ_0 is a degenerate distribution at 0. We can either use the asymptotic distribution to test the null $H_0 : \lambda = 0$ vs $H_1 : \lambda > 0$, or we can simulate additionally from the background model under the null hypothesis to get the empirical distribution of the test statistic. Furthermore, since under the null distribution, the background data and the experimental data have the same distribution, we can even permute the additionally simulated background and the experimental and compute the test statistic on the permuted data to get the empirical distribution of the test statistic under the null. For the experiments in Section 9.2, since we have only limited background data, we use a subset of the data for training the classifier and another subset in order to perform the permutation test. Since the background testing set is limited in size and under the null, the experimental data has the same distribution as the background, for the bootstrap method we sample with replacement from a mixture of both test experimental and background data.

8.2.1 Score Statistic

Similar to the idealized case, an alternative to the LRT statistic is the score test statistic

$$\sum_i \frac{p_s(W_i)}{p_b(W_i)} - N,$$

where we can again estimate $\psi = ps/p_b$ by $\widehat{\psi}$. So an alternate statistic to (8.14) is given by:

$$S = \frac{1}{N}\sum_{i=1}^{N} \widehat{\psi}(W_i). \tag{8.15}$$

This has the advantage that there is no need to estimate λ for the test and hence the test is not sensitive to the estimation process of λ. Similar to the LRT, we can find the null distribution by generating additional simulations from the background model or by permuting the additional background samples with the experimental and then finding the empirical distribution of the test statistic.

8.3 Model Independent (Semi-Supervised) Case

In this case, we assume that we don't have access to (or don't completely trust) the signal training sample $Y_1, \ldots, Y_n \sim p_s$. So we only observe auxiliary samples

$$X_1, \ldots, X_m \sim p_b.$$

So we have

$$\begin{aligned} X_1, \ldots, X_m &\sim p_b \\ W_1, \ldots, W_N &\sim q = (1-\lambda)p_b + \lambda p_s. \end{aligned}$$

We want to test $H_0 : \lambda = 0$ which is equivalent to $H_0 : p_b = q$.

One strategy is to use a classifier like before, but this time we estimate $\widetilde{\psi} = q/p_b$, where $q = (1-\lambda)p_b + \lambda p_s$. We then use the test based on (8.10). A second strategy is to use the area under the curve statistic (AUC) to evaluate the performance of the classifier. We discuss both the strategies below.

8.3.1 Test based on likelihood ratio test statistic

The difference between this case and the model dependent case is that, in this case instead of training a classifier to differentiate between signal and background events, we train a classifier to differentiate between experimental and background events. We propose three different methods that use the likelihood ratio test statistic to find test $H_0 : p_b = q$.

Method 8.3.1. LRT Permutation Method — this is a slow method.

1. Train a classifier $\widetilde{h}$ to differentiate between $X_1, \ldots, X_m$ and $W_1, \ldots, W_N$.

2. Compute the classifier based LRT statistic T based on Equation (8.10) using arguments similar to Equation (8.12) as

$$T = \log\left(\frac{1-\pi}{\pi}\right) + \frac{1}{N}\sum_i \log\left(\frac{\widetilde{h}(W_i)}{1-\widetilde{h}(W_i)}\right) \tag{8.16}$$

   where $\pi = N/(m+N)$. This is an estimate of the Neyman-Pearson test $\sum_i \log q(W_i)/p(W_i)$. Note that the sum is only over W_i.

3. Get the p-value by permutating the labels of $\{X_1, \ldots, X_m, W_1, \ldots, W_N\}$, re-training a classifier each time and finding the corresponding LRT statistic T.

Method 8.3.2. LRT Asymptotic Method — this is a fast method.

1. Break $X_1, \ldots, X_m$ into two groups $\mathcal{X}_1$ and $\mathcal{X}_2$ of sizes m_1 and m_2 respectively.

2. Break $W_1, \ldots, W_N$ into two groups $\mathcal{W}_1$ and $\mathcal{W}_2$ of sizes N_1 and N_2 respectively.

3. Construct the classifier $\widetilde{h}$ from $\mathcal{X}_1$ and $\mathcal{W}_1$.

4. Evaluate T as defined in Equation (8.16) on $\mathcal{W}_2$ only as

$$T = \log\left(\frac{1-\pi}{\pi}\right) + \frac{1}{N_2}\sum_{W_i \in \mathcal{W}_2} \log\left(\frac{\widetilde{h}(W_i)}{1-\widetilde{h}(W_i)}\right) \tag{8.17}$$

   where $\pi = N_1/(m_1 + N_1)$.

5. Then conditional on $\mathcal{X}_1$ and $\mathcal{W}_1$, under H_0, $\mathcal{X}_2$ and $\mathcal{W}_2$ should have the same distribution. Hence, T should have the same distribution as T_0 defined as

$$T_0 = \log\left(\frac{1-\pi}{\pi}\right) + \frac{1}{m_2}\sum_{X_i \in \mathcal{X}_2} \log\left(\frac{\widetilde{h}(X_i)}{1-\widetilde{h}(X_i)}\right).$$

   Then conditional on $\mathcal{X}_1$ and $\mathcal{W}_1$, under H_0,

$$\frac{\sqrt{N_2}(T - T_0)}{\sqrt{2}\sigma_{0T}} \overset{d}{\rightsquigarrow} N(0,1),$$

   where

$$\sigma_{0T}^2 = \mathrm{Var}_0\left(\log\left(\frac{\widetilde{h}(X)}{1-\widetilde{h}(X)}\right)\middle|\, \mathcal{X}_1, \mathcal{W}_1\right)$$

   and Var_0 is variance under p_b. This can be estimated by the variance of data in $\mathcal{X}_2$.

Method 8.3.3. LRT Bootstrap Method — this is faster than permutation method but slower than the asymptotic method.

1. Repeat steps 1–4 from Method 8.3.2 with $N_2 = m_2$.

2. Estimate the null distribution of T by drawing bootstrap samples (sample with replacement) from $\mathcal{X}_2 \cup \mathcal{W}_2$ of size $m_2 = N_2$ and computing T repeatedly. Then get the p-value.

Method 8.3.4. LRT Faster Permutation Method — speed similar to the bootstrap method.

1. Repeat steps 1–4 from Method 8.3.2.

2. Estimate the null distribution of T by drawing N_2 permuted samples from $\mathcal{X}_2 \cup \mathcal{W}_2$ and computing T repeatedly. Then get the p-value.

8.3.2 Test based on area under the curve (AUC) statistic

We have noticed that under the null $H_0 : \lambda = 0$, the experimental distribution $q = p_b$. So a classifier trained to differentiate the experimental data from the background data should have an AUC of 0.5. That is, an AUC that is significantly greater than 0.5 is an evidence of $q \neq p_b$. Therefore, we argue that testing $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$ is equivalent to testing $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$.

Similar to using the likelihood ratio test statistic, this test can also be performed in various ways. The first method uses in-sample AUC and the remaining three methods (similar to the methods for LRT) use out-of-sample AUC as the test statistic.

Method 8.3.5. AUC Permutation Method — this is a slow method.

1. Train a classifier $\widetilde{h}$ to differentiate between $X_1, \ldots, X_m$ and $W_1, \ldots, W_N$.

2. The AUC (θ) is then defined as

$$\theta = \mathbb{P}\left(\widetilde{h}(W) > \widetilde{h}(X)\right)$$

   which can be estimated by the AUC test statistic

$$\widehat{\theta} = \frac{1}{mN} \sum_{i=1}^{m} \sum_{j=1}^{N} \mathbb{I}\left\{\widetilde{h}(W_j) > \widetilde{h}(X_i)\right\}. \tag{8.18}$$

3. Get the p-value by permutating the labels of $\{X_1, \ldots, X_m, W_1, \ldots, W_N\}$, re-training a classifier each time and finding the corresponding AUC statistic $\widehat{\theta}$ given by (8.18).

Method 8.3.6. AUC Asymptotic Method — this is a fast method.

1. Perform steps 1–3 from Method 8.3.2 to get a classifier $\widetilde{h}$ based on $\mathcal{X}_1$ and $\mathcal{W}_1$.

2. Evaluate $\widehat{\theta}$ as defined in Equation (8.18) using $\mathcal{X}_2$ and $\mathcal{W}_2$ as

$$\widehat{\theta} = \frac{1}{m_2 N_2} \sum_{X_i \in \mathcal{X}_2} \sum_{W_j \in \mathcal{W}_2} \mathbb{I}\left\{\widetilde{h}(W_j) > \widetilde{h}(X_i)\right\}. \tag{8.19}$$

3. Then the expectation and variance of the estimate $\widehat{\theta}$ is given by Newcombe (2006) as

$$\begin{aligned}\mathbb{E}\left[\widehat{\theta}\right] &= \frac{1}{m_2 N_2} \sum_{X_i \in \mathcal{X}_2} \sum_{W_j \in \mathcal{W}_2} \mathbb{E}\left[\mathbb{I}\left\{\widetilde{h}(W_j) > \widetilde{h}(X_i)\right\}\right] \\ &= \mathbb{P}\left(\widetilde{h}(W) > \widetilde{h}(X)\right) = \theta,\end{aligned}$$

$$\begin{aligned}V\left(\widehat{\theta}\right) &\approx \frac{\theta(1-\theta)}{m_2 N_2}\left[1 + (N^* - 1)\left(\frac{1-\theta}{2-\theta} + \frac{\theta}{1+\theta}\right)\right] \\ &= \frac{\theta(1-\theta)}{m_2 N_2}\left[2N^* - 1 + \frac{3N^* - 3}{(2-\theta)(1+\theta)}\right],\end{aligned}$$

where $N^* = (m_2 + N_2)/2$. This can be estimated by,

$$\widehat{V\left(\widehat{\theta}\right)} = \frac{\widehat{\theta}(1-\widehat{\theta})}{m_2 N_2}\left[2N^* - 1 + \frac{3N^* - 3}{(2-\widehat{\theta})(1+\widehat{\theta})}\right] \tag{8.20}$$

Then under H_0,

$$\frac{\widehat{\theta} - 0.5}{\sqrt{\widehat{V\left(\widehat{\theta}\right)}}} \approx N(0,1), \tag{8.21}$$

which can be used to find the p-value.

The bootstrap method and the faster permutation method for the AUC are very similar to the LRT versions. We briefly detail them below.

Method 8.3.7. AUC Bootstrap Method — this is faster than permutation method but slower than the asymptotic method.

1. Repeat steps 1–2 from Method 8.3.6.
2. Draw m_2 bootstrapped X's and N_2 bootstrapped W's (sample with replacement) from $\mathcal{X}_2 \cup \mathcal{W}_2$ and compute $\widehat{\theta}$ using (8.19).
3. Estimate the null distribution of $\widehat{\theta}$ by computing $\widehat{\theta}$ using step 3 repeatedly. Then get the p-value.

Method 8.3.8. AUC Faster Permutation Method — speed similar to the bootstrap method.

1. Repeat steps 1–2 from Method 8.3.6.
2. Estimate the null distribution of $\widehat{\theta}$ by drawing m_2 X's and N_2 W's after permuting samples from $\mathcal{X}_2 \cup \mathcal{W}_2$ and computing $\widehat{\theta}$ repeatedly using (8.19). Then get the p-value.

8.3.3 Finding an estimate of the signal strength λ

In the supervised case, we estimate the λ using its MLE during the test itself. Whereas we don't directly estimate λ in the semi-supervised case. Recollect that in the semi-supervised case we combine the data:

$$(\widetilde{Z}_1, \ldots, \widetilde{Z}_{m+N}) = (X_1, \ldots, X_m, W_1, \ldots, W_N).$$

We define $\pi = N/(m+N)$,

$$W_i = \begin{cases} 0 & i \leq m \\ 1 & i > m. \end{cases}$$

and find the classifier $\widetilde{h}(x) = \widehat{P}(W = 1|Z = z)$, which is an estimate of $h_0(x) = P(W = 1|Z = z)$. Then

$$h_0(x) = P(W = 1|Z = z) = \frac{q(z)\pi}{q(z)\pi + p_b(z)(1-\pi)}.$$

Hence

$$\widetilde{\psi} = \frac{q}{p_b} = \frac{(1-\pi)h_0}{\pi(1-h_0)}.$$

First note that

$$\frac{(1-\pi)h_0}{\pi(1-h_0)} = \frac{q}{p_b} = \frac{(1-\lambda)p_b + \lambda p_s}{p_b} = (1-\lambda) + \lambda\widetilde{\psi}. \tag{8.22}$$

It seems that λ and $\psi = p_s/p_b$ are not identifiable.

But over any non-signal region S^c, we expect that $\psi = 0$. Then

$$\frac{(1-\pi)h_0(u)}{\pi(1-h_0(u))} = 1 - \lambda$$

for all $u \in S^c$. So

$$\lambda = 1 - \frac{(1-\pi)h(u)}{\pi(1-h(u))}$$

for all $u \in S^c$. An estimate of λ is then

$$\widehat{\lambda} = 1 - \left(\frac{1-\pi}{\pi}\right)\frac{1}{k}\sum_j \frac{\widetilde{h}(Z_j)}{1-\widetilde{h}(Z_j)}$$

where the sum is over all $Z_j \in \widehat{S}^c$ and k is the number of observations in $\widehat{S}^c$.

8.4 Combining the Two Methods (Best of Both Worlds)

We can actually combine the model dependent and model independent methods by using a bigger mixture. We replace (8.1) with

$$p_W(w) = \lambda_0 p_0(w) + \lambda_1 p_1(w) + \lambda_2 p_2(w) \tag{8.23}$$

where p_0 is the background, p_1 is the signal from the model (using the sample $Y_1, \ldots, Y_n$) and p_2 represents any unknown signal not captured by the model signal p_1.

8.5 Interpreting the Classifier Using Active Subspace Methods

To discover the signal region, first we need to understand what influences the classifier, and then interpret it. We use active subspace ideas presented in Constantine (2015) to detect the variables that influence the classifier the most and identify the subspace that could potentially lead us to the signal region.

Given the classifier $\widetilde{h}(\cdot)$ introduced in sections 8.3.1 and 8.3.2, define

$$C = \mathbb{E}\left[\left(\nabla_{\mathbf{z}}\widetilde{h} - \mathbb{E}\left[\nabla_{\mathbf{z}}\widetilde{h}\right]\right)\left(\nabla_{\mathbf{z}}\widetilde{h} - \mathbb{E}\left[\nabla_{\mathbf{z}}\widetilde{h}\right]\right)^T\right],$$

where $\nabla_{\mathbf{z}}$ denotes the usual d-dimensional gradient operator. Then C has a real eigen-value decomposition,

$$C = M\Lambda M^T, \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d), \quad \lambda_1 \geq \ldots \geq \lambda_d \geq 0, \tag{8.24}$$

M has columns $\{\mathbf{m}_1, \ldots, \mathbf{m}_d\}$, the normalized eigenvectors of C. The $\mathbf{m}_1$ corresponding to λ_1 best captures the change in $\widetilde{h}$, followed by $\mathbf{m}_2$ and so on. Therefore the eigen vectors corresponding to the leading eigen values $\lambda_1, \lambda_2, \ldots$, give an idea about the directions along which the classifier output changes the most.

Towards this end, we propose the following algorithm for the experimental data

$$W_1, \ldots, W_N \sim q = (1-\lambda)p_b + \lambda p_s.$$

Method 8.5.1. Active Subspace Method – finding the subspace that captures the most variation in the classifier $\widetilde{h}(\cdot)$.

1. Estimate $\nabla_{\mathbf{z}}\widetilde{h}(W_j)$ by using a local linear smoother that estimates $\widetilde{h}(\cdot)$ at $W_1, \ldots, W_N$. Let's call the estimate $\nabla_{\mathbf{z}}h_j = \widehat{\nabla_{\mathbf{z}}\widetilde{h}(W_j)}$.

2. Estimate C using

$$\widehat{C} = \frac{1}{N}\sum_{j=1}^{N}\left(\nabla_{\mathbf{z}}h_j - \overline{\nabla_{\mathbf{z}}h_j}\right)\left(\nabla_{\mathbf{z}}h_j - \overline{\nabla_{\mathbf{z}}h_j}\right)^T, \tag{8.25}$$

where $\overline{\nabla_{\mathbf{z}} h_j} = \sum_{j=1}^{N} \nabla_{\mathbf{z}} h_j / N$.

3. Find the eigen-value decomposition of $\widehat{C}$ as

$$\widehat{C} = \widehat{M}\widehat{\Lambda}\widehat{M}^T,$$

which gives the estimates $\widehat{M}$ and $\widehat{\Lambda}$ of M and Λ as defined in (8.24) respectively.

4. Then $\overline{\nabla_{\mathbf{z}} h_j} = \sum_{j=1}^{N} \nabla_{\mathbf{z}} h_j / N$, followed by $\mathbf{m}_1, \mathbf{m}_2, \ldots$ best capture the change in $\widetilde{h}(\cdot)$.

We use projections onto these vectors to identify the subspace that most influences the semi-supervised classifier. We also use sparse PCA instead of PCA to find sparser active vectors.

Chapter 9

Experiments: Search for the Higgs Boson

We demonstrate the performance of the proposed anomaly detection classifier tests on the Higgs boson machine learning challenge hosted by Kaggle at https://www.kaggle.com/c/higgs-boson. The challenge consists of simulated data provided by the ATLAS experiment at CERN to optimize the analysis of the Higgs boson.

The goal here is to demonstrate the performance of the proposed tests in identifying the presence of the Higgs boson particle and their applicability to the search of new physics signals in experimental particle physics. The presence of such a *signal* generally reveals itself as a tiny significant excess of certain type of collision events in a particle detector that are unexplainable by known *background* processes. So the goal is to detect and extract these minute signals from a very large set of background events.

9.1 Data Description

The Higgs boson has many different ways through which it can decay in an experiment and produce other particles. The challenge particularly focusses on the events where it decays into two tau particles (Adam-Bourdarios et al., 2014). The data provided for the challenge consists of events labelled as background and signal where the signal class is comprised of events in which the Higgs boson decays into two taus. The events are simulated using the official ATLAS full detector simulator. The simulator yields simulated events with properties that mimic the statistical properties of the real events of the signal type as well as several important backgrounds. The signal consists of events in which Higgs bosons (with fixed mass 125 GeV) were produced. The background events are generated by other known processes which can produce events

that mimic the signal. Our objective is to show that semi-supervised classifier tests are able to identify such signals without any prior knowledge.

The challenge provides a training set, a validation set and a test set. We use the training set to demonstrate the performance of the classifier tests. The training set has 250K observations, where each observation is a simulated collision event. There are $d = 30$ features whose individual details can be found in the Appendix B of Adam-Bourdarios et al. (2014). Here we give some insight into the most important characteristics of the features.

The features prefixed with `PRI` (for PRImitives) are "raw" quantities as measured by the detector. Those prefixed with `DER` (for DERived) are quantities computed from the primitive features. The missing values in the data are structurally absent as some events have no jet (`PRI_jet_num`= 0) and hence no such a thing as a "leading jet". Thus the associated primitive quantities are all missing (Adam-Bourdarios et al., 2015). Since the derived quantities are functions of the primitives, we use just the primitive variables ($d = 16$) for our analysis. Also to avoid any missing values in the data, we only consider events that have two jets (`PRI_jet_num`= 2) which results in 50,379 events, 24,645 background events and 25,734 signal events.

Among the primitive features, five of them provide the azimuth angle ϕ of the particles generated in the event (variables ending with `_phi`). These features are rotation invariant in the sense that the event doesn't change if all of them are rotated together with any angle. Hence to interpret these variables more easily using the active subspace methods, we remove the invariance of the azimuth angle variables by rotating all the ϕ's and setting the azimuth angle of the leading jet at 0 (`PRI_leading_phi`= 0).

Additionally, we take logarithmic transformations of the variables that give the transverse momentum of the particles produced (variables ending with `_pt`), the missing transverse energy (`PRI_met`) and the total transverse energy in the detector (`PRI_met_sumet`).

Exploratory data analysis of data as well as details and justifications for the transformations considered above, can be found in Appendix B. In the following sections, we explore the power of the classifier tests described in Section 8.3 to detect the signal from the background and then use the active subspace methods introduced in Section 8.5 to explore the signal region.

9.2 Anomaly Detection Using the Classifier Tests

We compare the power of the methods introduced in Chapter 8 to nearest neighbor two-sample tests as introduced in Schilling (1986) and Henze (1988). We consider the asymptotic version of the test and a permutation version of the test to compare with the classifier tests. We compare the power of the tests in detecting the signal, by varying the signal strength from $\lambda = 0.2$ to $\lambda = 0.01$. We also make sure the tests have the right error control under the null case ($\lambda = 0$).

For the methods introduced in Chapter 8, we consider m =12,322, n =7,322, N =12,323, where the amount of signal events in the experimental data is varied according to the λ as $\lfloor N\lambda \rfloor$. Similarly, the number of background events in the experimental data is given by $\lceil N(1-\lambda) \rceil$. The data-splitting is done by taking m_1 = 7,322, m_2 =5,000, N_1 =7,323, N_2 =5,000. We also take the splits such that in $\mathcal{W}_1$ and $\mathcal{W}_2$ there are $\lfloor N_1\lambda \rfloor$ and $\lfloor N_2\lambda \rfloor$ signal events respectively.

Additionally, for the model-dependent methods, since we have only finitely many samples from the background available and not the background generator itself, we use $\mathcal{X}_1$ that has m_1 = 7,322 background samples along with n =7,322 signal samples for training the supervised classifier. Then we bootstrap and permute from $\mathcal{X}_2$ with m_2 =5,000 background samples to find the empirical distributions of the test statistics. In real life, since the MC generator is known, we should be able to generate more training background samples.

For each of the bootstrap and permutation methods we consider 1,000 bootstrap and permutation cycles respectively. The tests are run on 100 random samplings of the data and the number of times each of the tests rejects the null that there is no signal, is given in Table 9.1. Permutation 1 indicates the faster permutation method in Section 8.3, that uses out-of sample test statistics for testing and Permutation 2 indicates the slower permutation method that uses in-sample test statistics for testing and re-trains the classifier in every permutation cycle.

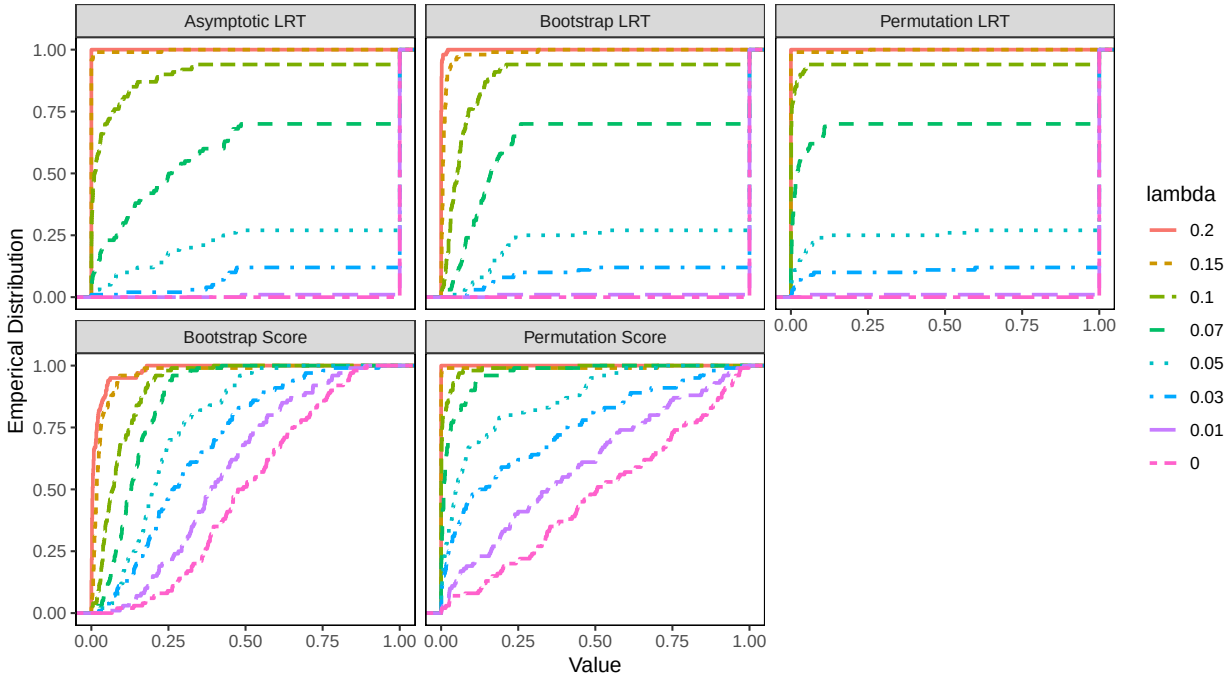


Figure 9.1: Supervised Methods

Table 9.1: Power of detecting the signal for each model in 100 random samplings of the Higgs boson data. We consider 1000 iterations for the bootstrap and permutation methods.

			Signal Strength (λ)							
Model	Method	k	0.2	0.15	0.1	0.07	0.05	0.03	0.01	0
Supervised LRT	Asymptotic		100	99	70	22	5	1	0	0
	Bootstrap		100	96	46	10	1	0	0	0
	Permutation		100	99	93	59	19	8	1	0
Supervised Score	Bootstrap		90	83	37	10	2	2	0	0
	Permutation		100	99	94	80	51	35	13	7
Semi-Supervised LRT	Asymptotic		100	99	63	16	20	8	5	7
	Bootstrap		100	93	33	7	6	2	0	1
	Permutation 1		100	99	60	17	19	9	5	8
	Permutation 2		53	11	1	4	2	1	2	6
Semi-Supervised AUC	Asymptotic		100	96	63	17	17	7	6	8
	Bootstrap		100	97	62	16	16	7	5	8
	Permutation 1		100	97	62	18	16	7	6	8
	Permutation 2		100	100	74	38	23	9	4	6
k-NN Two-Sample Test	Asymptotic	15	46	30	12	4	7	3	6	2
		20	51	30	13	3	7	2	7	2
		25	54	29	10	6	9	4	7	3
		30	55	30	11	5	8	3	4	2
	Permutation	15	94	63	21	14	12	6	9	8
		20	98	66	26	13	12	8	9	6
		25	98	70	27	12	10	6	8	9
		30	98	74	33	10	10	5	8	5

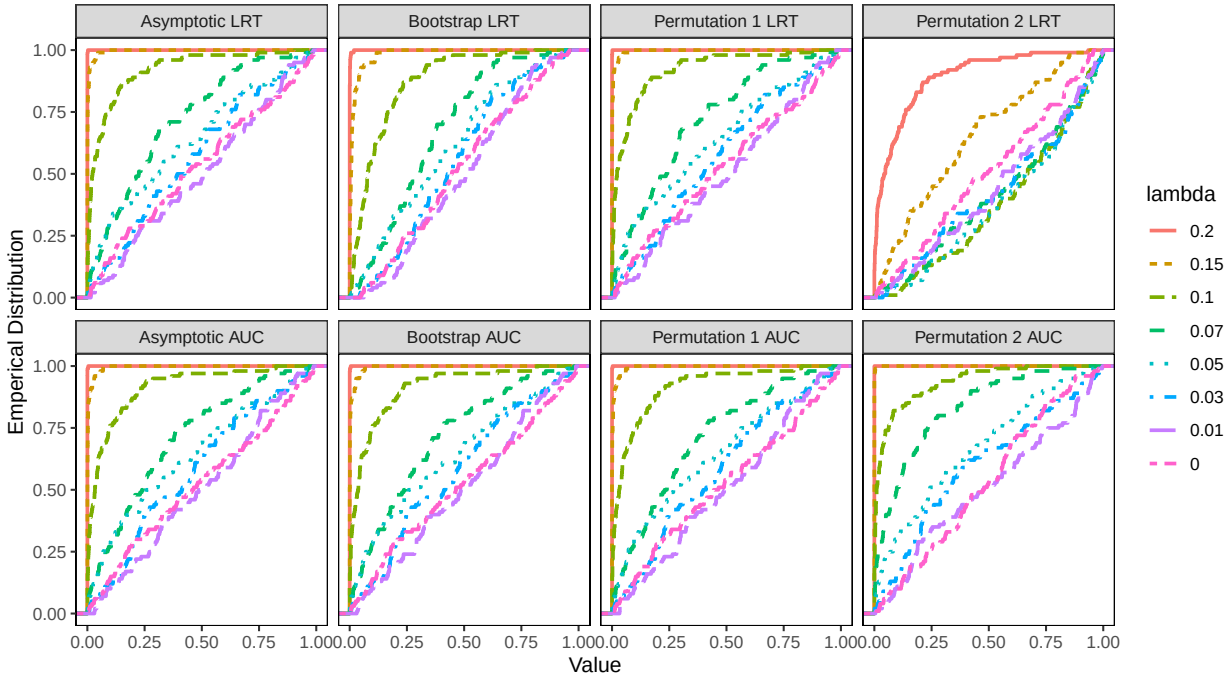


Figure 9.2: Semi-Supervised Methods

As shown in Table 9.1, among the supervised methods, the permuted score test supersedes all the other supervised methods and additionally has the right type I error control (Figure 9.1). The supervised tests that use the LRT statistic seem to be over conservative. This appears to be caused by $\widehat{\lambda}_{\text{MLE}}$, the maximum likelihood estimator of λ, being zero most of the time for very small λ values.

Among the semi-supervised approaches, the AUC slower permutation method (permutation 2) has comparable power to the supervised methods and even performs better than some of them as shown in Table 9.1. For the asymptotic and the faster permutation methods, using LRT gives similar performance to AUC for the semi-supervised approaches. The permuted two-sample nearest neighbor tests, don't perform as well as the semi-supervised AUC methods even though they out-perform their asymptotic versions by a large margin.

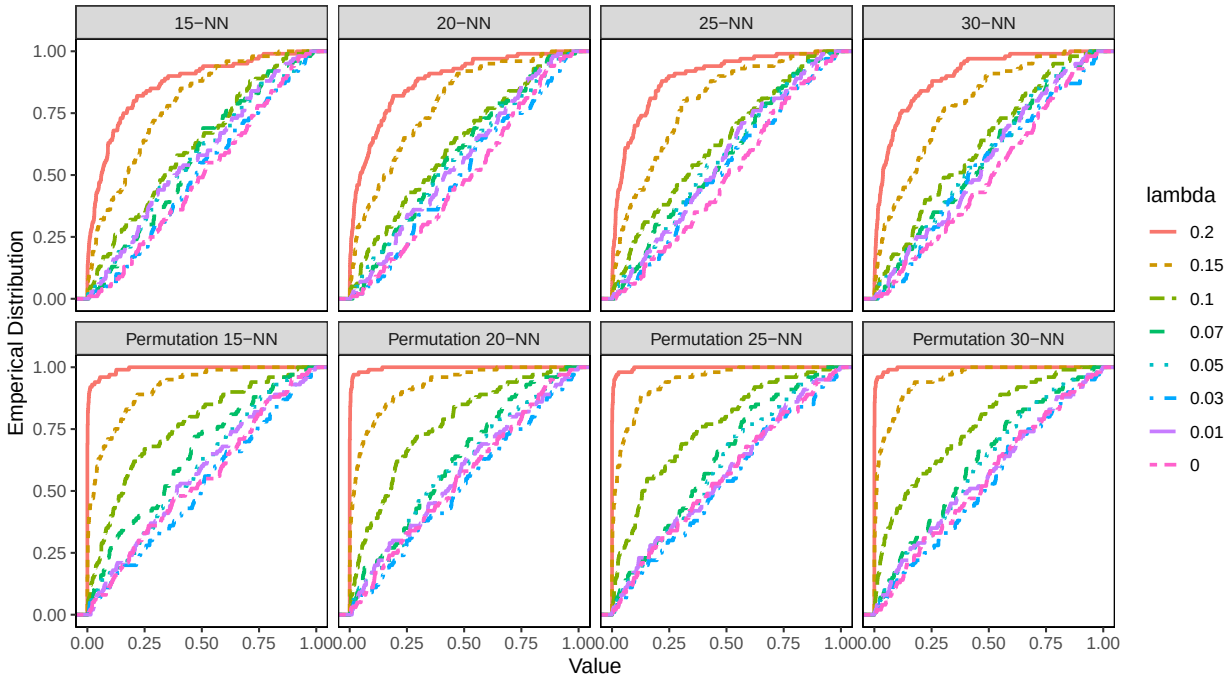


Figure 9.3: Nearest Neighbor Two-Sample Test Methods

Figures 9.2 and 9.3 show the empirical distributions of the p-values for the different λ values. All the tests appear to have correct type I error control ($\lambda = 0$ case). We also notice that none of the tests have any power to detect signals that are less than 5% of the experimental data ($\lambda = 0.05$). The permutation 2 method using LRT appears to have an especially low power for all λ values since it uses an in-sample version of the test statistic.

In conclusion, we see that the semi-supervised AUC methods generally out-perform nearest neighbor two-sample tests and LRT methods and additionally give comparable performance to the supervised methods in detecting the signal in the experimental data.

9.3 Application of Active Subspace Methods

We demonstrate the application of the active subspace methods, for a single random simulation (one of the 100 simulations explored in Section 9.2) that detects the signal at significance level α, when $\lambda = 0.1$, i.e., 10% of the experimental data is from the signal sample. We consider $\lambda = 0.1$, since the random forests demonstrate some power in detecting the signal (Table 9.1) and it's more realistic than higher λ values.

As described in Section 9.2, for the semi-supervised methods, we consider a training set of $m_1 = 7{,}322$ background events and $N_1 = 7{,}323$ experimental events, which contains $\lfloor N_1\lambda \rfloor = 732$ signal events. We test for the presence of signal using a test set of $m_2 = 5{,}000$ background events and $N_1 = 5{,}000$ experimental events, which contains $\lfloor N_2\lambda \rfloor = 500$ signal events. We train a random forest classifier on the training data to differentiate between the background and the experimental events.

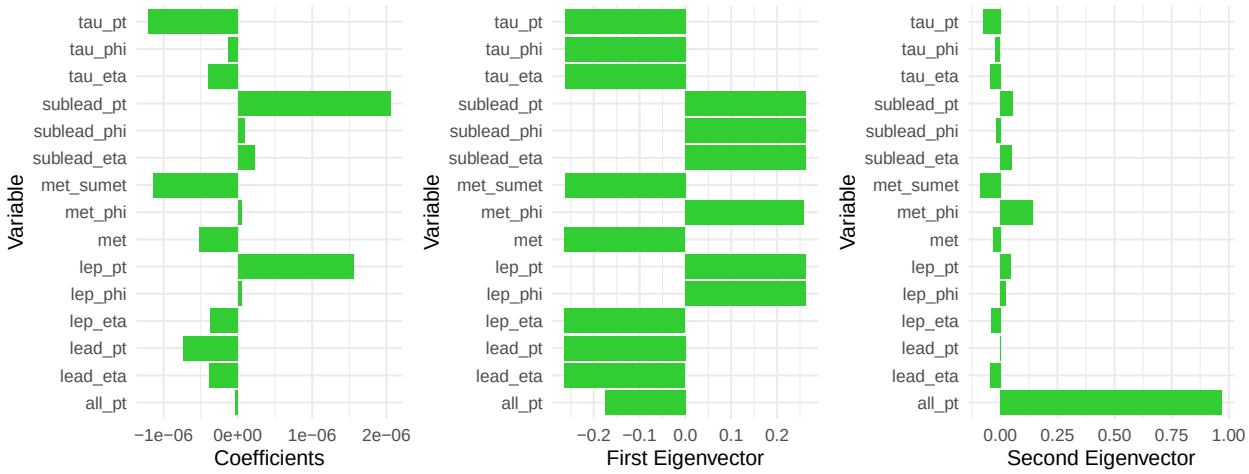


Figure 9.4: Active Subspace Variables for $\lambda = 0.1$.

We then use the Method 8.5.1 presented in Section 8.5 to find the active subspace. The first step of the algorithm requires us to choose a smoothing parameter as well as a linear smoother. We choose a Gaussian kernel smoother as the linear smoother and the smoothing parameter selection is described in Appendix B.1. Figures 9.4 and 9.5 give us the active variables using PCA and sparse PCA respectively.

The first variable is the same in both Figures 9.4 and 9.5, since the first variable is the mean gradients projection as defined in Method 8.5.1 which gives the direction of change of the classifier. We see that the transverse momentums of the sub-leading jet (`sublead_pt`) and the lepton (`lep_pt`) positively affect the classifier, which implies that signal events display higher momentums of the sub-leading jet and the lepton as compared to background events. Additionally signal events demonstrate low hadronic tau transverse momentum (`tau_pt`) and total transverse energy (`met_sumet`) in the detector.

The first eigenvector gives the first principal component of the gradients, which demonstrates the relationship between the variables that causes the most variability in the gradients of the classifier. The

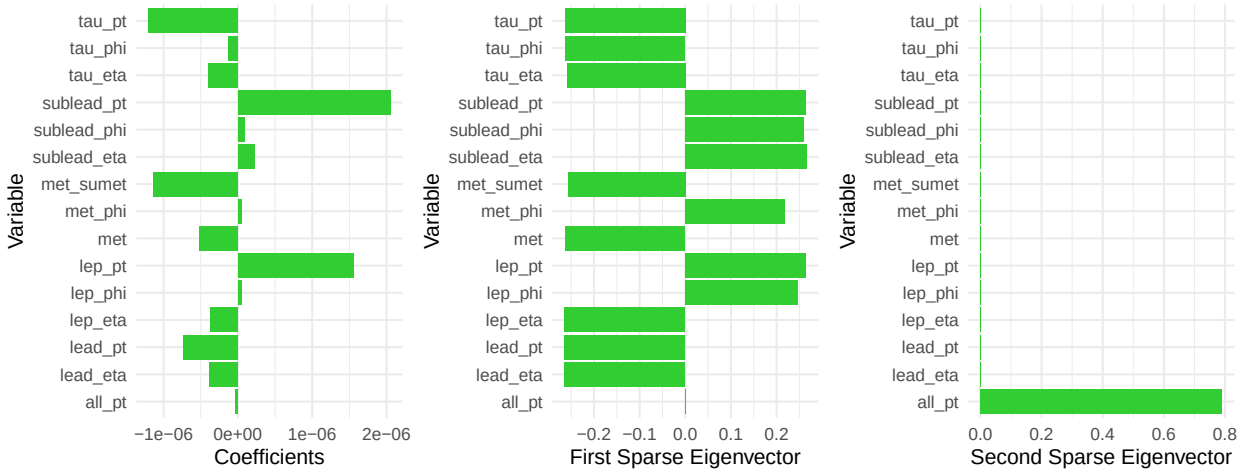


Figure 9.5: Sparse Active Subspace Variables for $\lambda = 0.11$.

second eigen vector, i.e. the second principal component of the gradients, indicates that increasing scalar sum of the transverse momentum of all the jets (`all_pt`) in the event leads to variability in the classifier, and hence helps in differentiating the signal from the background.

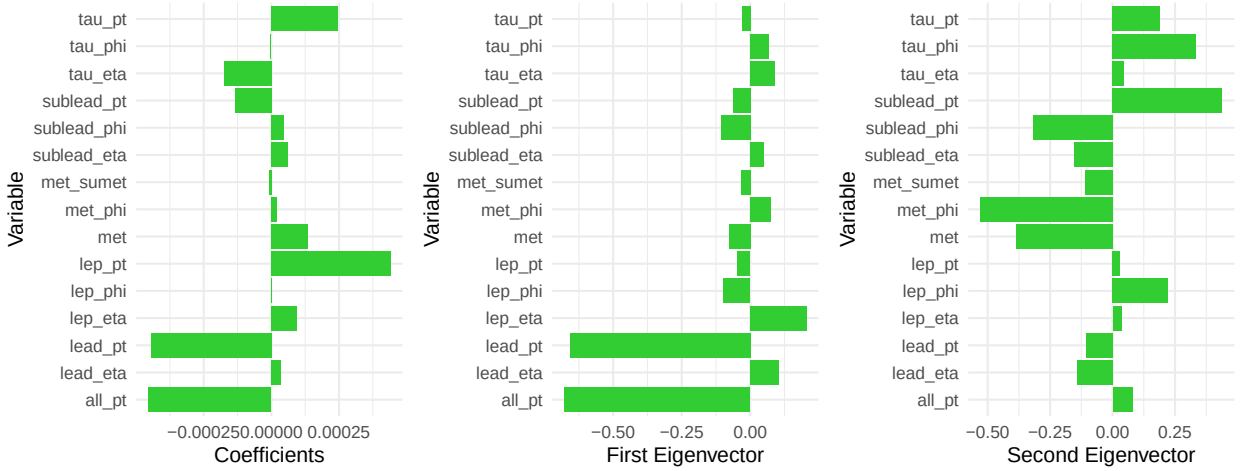


Figure 9.6: Active Subspace Variables for $\lambda = 1.5$.

We further increase λ to $\lambda = 1.5$ to explore the results of the active subspace methods in a case where the signal is more easily distinguishable by the random forest classifier. Figures 9.6 and 9.7 give us the active variables using PCA and sparse PCA respectively for data in which $\lambda = 1.5$. In this case, sparse PCA gives much cleaner results as compared to normal PCA. We note that the mean gradient projection in this case indicates that low scalar sum of the transverse momentum of all the jets (`all_pt`) and low transverse momentum of the leading jet (`lead_pt`) seems to indicate the presence of a signal event. Furthermore, high transverse momentum of the hadronic tau (`tau_pt`) also seems to indicate the presence of a signal event.

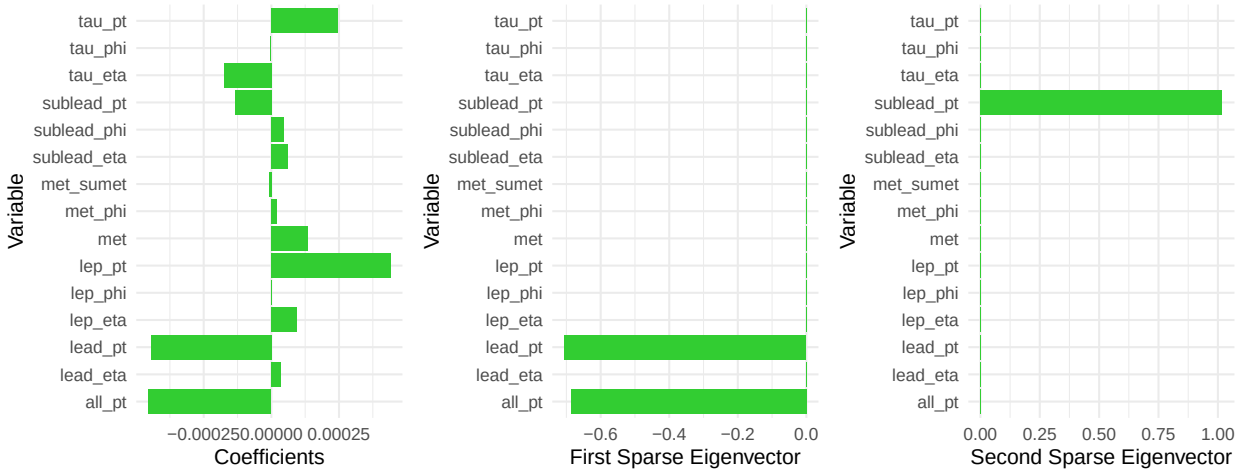


Figure 9.7: Sparse Active Subspace Variables for $\lambda = 1.5$

These are supported by the $\lambda = 0.1$ case as well, since some of the insights gained in the previous case indicated that (`lead_pt`) influences the classifier and that high transverse momentum of the lepton (`lep_pt`) supports the occurrence of a signal event. The first and second sparse eigen vectors for $\lambda = 1.5$ also seem to indicate that the when `lead_pt` and `all_pt` change in the same direction, together it causes variation in the classifier. Hence both of them increasing together might be an indication of a signal event. Further the value of `sublead_pt` influences the classifier as well, which is again supported by the analysis of the data for $\lambda = 0.1$.

From both of these simulations we conclude that, the active subspace methods propose an algorithm to interpret the classifier that is able to detect the signal in the experimental data. The $\lambda = 0.1$ and $\lambda = 0.15$ simulations seem to imply that classifier is most influenced by `lep_pt`, `all_pt` , `lead_pt` and `lsubead_pt`. Higher sparse principal components for $\lambda = 1.5$, also identify the transverse momentum of the hadron tau (`tau_pt`) and the relationship between the phi angles between the leading jet and the missing transverse energy (`met_phi`) as important for detecting the signal events.

Part IV

Conclusion

Conclusion

10.1 Summary

The main contributions presented in this thesis are:

1. A new clustering algorithm for high-dimensional data that is equipped with a significance guarantee. The algorithm uses a new test, RIFT, that is based on the idea of relative fit and does not require the user to provide a predetermined number of clusters. Additionally, we do the following:
   (a) Use the test to derive significant clusters in both a hierarchical and sequential manner.
   (b) Study the limiting distribution and power of a similar test called SigClust (Liu et al., 2008), which introduces a test based on the k-means objective.
2. Model-independent anomaly detection tests using semi-supervised classifiers, that can detect the presence of signal events hidden in background events in high energy particle physics data sets. Additionally, we do the following:
   (a) Propose active subspace methods to identify the subspace affecting the classifier most strongly, leading to an understanding of the signal region.
   (b) Compare the proposed tests with model-dependent supervised methods as well as nearest neighbor two-sample tests on a data set related to the search for the Higgs boson.

In our work on inference for clustering, we presented an analysis of the SigClust procedure of Liu et al. (2008) in certain examples when the dimension d was held fixed. On the other hand, increasing dimension was considered in the work of Liu et al. (2008), but only under restrictive conditions. A more thorough understanding of the power of hypothesis testing based approaches when d increases is warranted.

We subsequently presented a different hypothesis testing based approach for clustering with mixtures of Gaussians based on *relative fit*. By testing the relative fit of different mixtures based on data splitting we get a simple test statistic with a Normal limiting distribution. As with any method, there are cases where the method works well but there are also cases where it fails. The main advantage of our approach is that it

uses a test with a simple limiting distribution and the test does not rely on the assumption that the model is correct.

In our work on inference for anomaly detection, we presented multiple model-independent methods that search for signal without assuming any signal model. By not assuming any signal model, we retain the ability to detect unknown and unexpected signals. We used a semi-supervised classifier to distinguish the experimental data from the background data and used the performance of the classifier to perform a test to detect a significant difference between the two data sets. We demonstrated the use of two statistics, the likelihood ratio test statistic (LRT) and the area under the curve statistic (AUC).

We compared the power of the methods to detect the Higgs boson at different signal strengths and showed that a version of the proposed AUC methods has comparable power to the model-independent methods. So even when the signal model is correctly assumed by the model-dependent methods, the proposed model-independent methods appear to still have power to detect the presence of the signal. However, when the signal model is incorrectly assumed, model-dependent methods might totally miss the signal, whereas the proposed model-independent ones should still be able to detect them. In particular, the proposed methods demonstrate the ability to find new particles without any a priori knowledge of their properties.

10.2 Vision and Future Work

Moving forward, we aim to increase the usability of the significant clustering methods as well as extend the use of interpretable model-independent methods in the high energy physics domain. The following are the ways in which we plan to extend the work presented in this thesis:

High-dimensional Clustering. Clustering in high-dimensions is still a difficult task despite the huge amount of research available on it. As discussed in the thesis, high dimensional Gaussians are difficult to fit, which affects the performance of RIFT. One possible solution that is worth pursuing, is to perform clustering after dimension reduction, for example, by using random projections. Another possible solution is to explore better ways of fitting a mixture of Gaussians to high-dimensional data. Another direction of research is to look at data generated from a distribution with k clusters and analyze whether the hierarchical clustering algorithm proposed in Chapter 5 of the thesis, detects these k clusters.

Semi-Supervised Anomaly Detection in Particle Physics. In this thesis, we explore a test that considers a model for the background data and a model for the experimental data that is a mixture of the background model and an anomalous component. There is literature on estimating the anomalous component using non-parametric methods under some moment and symmetry conditions (Bordes et al., 2006; Casa and Menardi, 2018). There is also research that uses variational autoencoders to do anomaly detection (Hajer et al., 2018; Cerri et al., 2019). A future direction of work is to evaluate how these methods compare to the

proposed tests using classifiers. Another direction is to explore the interpretability of model-independent methods to understand or quantify properties of the detected anomalies.

Relative Fit Methods. In the thesis, we propose a test, RIFT, that is based on the relative fit of two density estimators. But the problem of selecting a density estimator that best fits the data, from two density estimates, is surprisingly difficult to solve. An important consideration is choosing the distance measure used for comparing the fits to the true density. For RIFT, we used the Kullback-Leibler distance to compare the fits of two Gaussian mixture models. Some other alternatives are the ℓ_1 distance (Devroye et al., 1997; Devroye and Lugosi, 2012) and the Hellinger distance. An interesting direction of research is to compare the performance of the different distance measures in detecting the difference in fits.

Interdisciplinary Collaborations. In many fields, clustering is used to uncover real groupings inherent in the data. If the data is split into more clusters than the real groupings in the data, the resulting clusters could be arbitrary and consequently potentially misleading. For example in the medical sciences, especially in bioinformatics, scientists look for actual groupings in the patients for prognosis as well as to develop treatments for the different groups. These examples demonstrate the need for clustering methods with significance guarantees, that identify the real groupings inherent in the data. We intend to study how these approaches could be applied and adapted to these new settings.

Bibliography

Aaboud, M., Aad, G., Abbott, B., Abdinov, O., Abeloos, B., Abidi, S. H., AbouZeid, O., Abraham, N., Abramowicz, H., Abreu, H., et al. (2019). A strategy for a general search for new phenomena using data-derived signal regions and its application within the atlas experiment. *The European Physical Journal C*, 79(2):120. 6, 53

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S. A., Abdelalim, A. A., Abdinov, O., Aben, R., Abi, B., Abolins, M., et al. (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. 51

Aaltonen, T., Adelman, J., Akimoto, T., Albrow, M. G., Gonzalez, B. A., Amerio, S., Amidei, D., Anastassov, A., Annovi, A., Antos, J., et al. (2009). Global search for new physics with $2.0 fb^{-1}$ at cdf. *Physical Review D*, 79(1):011101. 53

Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kegl, B., and Rousseau, D. (2014). Learning to discover: the higgs boson machine learning challenge. *URL http://higgsml. lal. in2p3. fr/documentation*, page 9. 67, 68, 137

Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kégl, B., and Rousseau, D. (2015). The higgs boson machine learning challenge. In *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, pages 19–55. 68

Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26. 3

Aktas, A., Andreev, V., Anthonis, T., Asmone, A., Babaev, A., Backovic, S., Bähr, J., Baranov, P., Barrelet, E., Bartel, W., et al. (2004). A general search for new phenomena in ep scattering at hera. *Physics Letters B*, 602(1-2):14–30. 53

ATLAS Collaboration and CMS Collaboration (2011). LHC Higgs Combination Group, Procedure for the LHC Higgs boson search combination in Summer 2011. Technical report, CMS-NOTE-2011-005. 5, 51

Balakrishnan, S., Wainwright, M. J., Yu, B., et al. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120. 25

Bertram, I., Fox, H., Ross, A., Williams, M., Ratoff, P., et al. (2012). Model independent search for new phenomena in pp (bar) collisions at $\sqrt{s} = 1.96$ tev. *Physical Review D*, 85(9). 53

Bhat, P. C. (2011). Multivariate analysis methods in particle physics. *Annual Review of Nuclear and Particle Science*, 61:281–309. 51

Bickel, P. J. and Breiman, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *The Annals of Probability*, pages 185–214. 31

Bock, H. H. (1985). On some significance tests in cluster analysis. *Journal of Classification*, 2(1):77–108. 12, 16, 17, 19, 89, 90, 101, 104

Bordes, L., Delmas, C., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian journal of statistics*, 33(4):733–752. 78

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press. 94, 115, 117

Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A., Savage, M. I., Osborne, C. K., Hilsenbeck, S. G., Chang, J. C., et al. (2015). Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, 21(7):1688–1698. 4

Casa, A. and Menardi, G. (2018). Nonparametric semisupervised classification for signal detection in high energy physics. *arXiv preprint arXiv:1809.02977*. 6, 53, 78

Cerri, O., Nguyen, T. Q., Pierini, M., Spiropulu, M., and Vlimant, J.-R. (2019). Variational autoencoders for new physics mining at the large hadron collider. *Journal of High Energy Physics*, 2019(5):36. 78

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58. 5

Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the l 2 distance between competing models. *Journal of the American Statistical Association*, 99(466):488–498. 13

Chatrchyan, S., Khachatryan, V., Sirunyan, A. M., Tumasyan, A., Adam, W., Aguilo, E., Bergauer, T., Dragicevic, M., Erö, J., Fabjan, C., et al. (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. 51

Chen, J. (2017). On finite mixture models. *Statistical Theory and Related Fields*, 1(1):15–27. 11, 13, 25

Chen, J., Li, P., et al. (2009). Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, 37(5A):2523–2542. 13

Chen, J., Li, P., and Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107(499):1096–1105. 13

CMS Collaboration (2017). MUSiC, a model unspecific search for new physics, in pp collisions at $\sqrt{s} = 8$ TeV. *CMS Physics Analysis Summary CMS-PAS-EXO-14/016*. 53

Constantine, P. G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*, volume 2. SIAM. 6, 53, 64

Cowan, G., Cranmer, K., Gross, E., and Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2):1554. 5, 51

Dacunha-Castelle, D., Gassiat, E., et al. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes. *The Annals of Statistics*, 27(4):1178–1209. 11, 13, 25

Devroye, L. and Lugosi, G. (2012). *Combinatorial methods in density estimation*. Springer Science & Business Media. 79

Devroye, L., Lugosi, G., et al. (1997). Nonasymptotic universal smoothing factors, kernel complexity and yatracos classes. *The Annals of Statistics*, 25(6):2626–2637. 79

Engelman, L. and Hartigan, J. A. (1969). Percentage points of a test for clusters. *Journal of the American Statistical Association*, 64(328):1647–1648. 12

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. 13

Garcia-Escudero, L., Gordaliza, A., Matran, C., and Mayo-Iscar, A. (2009). A robust maximal f-ratio statistic to detect clusters structure. *Communications in Statistics-Theory and Methods*, 38(5):682–694. 12

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38:897–906. 11, 13, 25

Ghosh, J. K. and Sen, P. K. (1984). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. *Berkeley Conference In Honor of Jerzy Neyman and Jack Kiefer*. 11, 13, 29

Gu, J., Koenker, R., and Volgushev, S. (2017). Testing for homogeneity in mixture models. *Econometric Theory*, pages 1–46. 11, 13, 25

Hajer, J., Li, Y.-Y., Liu, T., and Wang, H. (2018). Novelty detection meets collider physics. *arXiv preprint arXiv:1807.10261*. 78

Hartigan, J. (1978). Asymptotic distributions for clustering criteria. *The Annals of Statistics*, pages 117–131. 96

Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Barkeley Conference in Honor of J. Neyman and J. Kiefer*, volume 2, pages 807–810. 13

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley. 12

Hennessy, B. T., Gonzalez-Angulo, A.-M., Stemke-Hale, K., Gilcrease, M. Z., Krishnamurthy, S., Lee, J.-S., Fridlyand, J., Sahin, A., Agarwal, R., Joy, C., et al. (2009). Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer research*, 69(10):4116–4124. 4

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783. 8, 54, 68

Huang, H., Liu, Y., Yuan, M., and Marron, J. (2015). Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, 24(4):975–993. 13

Kimes, P. K., Liu, Y., Neil Hayes, D., and Marron, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics*, 73(3):811–821. 4, 8, 11, 12, 16, 48

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. In *Journal of Physics: Conference Series*, volume 368, page 012032. IOP Publishing. 5, 6, 53

Lee, K. L. (1979). Multivariate tests for clusters. *Journal of the American Statistical Association*, 74(367):708–714. 12

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105(491):1084–1092. 13

Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, 123(1):61–81. 13

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. (2008). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293. 4, 7, 8, 11, 12, 13, 15, 16, 77

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605. 140

Maitra, R., Melnykov, V., and Lahiri, S. N. (2012). Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, 107(497):378–392. 12

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530. 30

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 115–128. 30

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons. 11, 12, 13

McLachlan, G. J. and Rathnayake, S. (2014). On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355. 11, 13

McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C., and Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469. 12

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179. 12

Network, C. G. A. R. et al. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519. 48

Network, C. G. A. R. et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543. 48

Newcombe, R. G. (2006). Confidence intervals for an effect size measure based on the mann–whitney statistic. part 2: asymptotic methods and evaluation. *Statistics in Medicine*, 25(4):559–573. 62

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160. 4

Pollard, D. (1982). A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926. 16, 19, 89, 91, 103, 104

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast cancer research*, 12(5):R68. 3, 4

Qiu, D. (2010). A comparative study of the k-means algorithm and the normal mixture model for clustering: Bivariate homoscedastic case. *Journal of Statistical Planning and Inference*, 140(7):1701 – 1711. 103, 104

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(2):405–408. 120

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65. 12

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806. 8, 54, 68

Suzuki, R. and Shimodaira, H. (2006). Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542. 12

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528. 12

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423. 12

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. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. 5, 6, 53

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110. 3

Vogt, M. and Schmid, M. (2017). Clustering with statistical error control. *arXiv preprint arXiv:1702.02643*. 8, 12

Wan, Y.-W., Allen, G. I., and Liu, Z. (2015). Tcga2stat: simple tcga data access for integrated statistical analysis in r. *Bioinformatics*, 32(6):952–954. 48

Williams, M. (2010). How good are your fits? unbinned multivariate goodness-of-fit tests in high energy physics. *Journal of Instrumentation*, 5(09):P09004. 5, 51

Zhou, S. and Jammalamadaka, S. R. (1993). Goodness of fit in multidimensions based on nearest neighbour distances. *Journal of Nonparametric Statistics*, 2(3):271–284. 31

Appendix

Appendix A

Proofs of Theorems, Lemmas and Corollaries in Part I

A.1 Proofs of Results Presented Under the Null Hypothesis of SigClust

In this Appendix, we prove Theorem 3.1 and all the results required to prove it. We first note that the regularity conditions ((ii), (iii) and (iv)) of Pollard (1982) and hence of Corollary 6.5 in Bock (1985) are satisfied by a $N(0, \Sigma)$ distribution. Furthermore, the 2-means solution is unique, under the conditions on Σ. Additionally, Lemma A.0.1 in Appendix A.1.2 shows that (v) holds. Thus it follows from Pollard's result that

$$\sqrt{n}(\mathbf{b_n} - \boldsymbol{\mu}) \rightsquigarrow N(0, G_0^{-1} V G_0^{-1}),$$

where $\boldsymbol{\mu}$ is the vector that minimizes the population within cluster sum of squares for the 2-means clustering, V is the $kd \times kd$ diagonal matrix with

$$V_i = 4\mathbb{E}\left[(X - \mu_i)(X - \mu_i)^T \mathbb{I}_{A_i}\right] \tag{A.1}$$

as its ith diagonal block and G_0 is analogously defined to the G as defined in equation (3.6) for the alternative. So, G_0 is a matrix made up of $d \times d$ matrixes of the form,

$$(G_0)_{ij} = \begin{cases} 2\mathbb{P}(A_i)\mathbf{I}_d - 2r_{ij}^{-1} \int_{M_{ij}} f(x)(x - \mu_i)(x - \mu_i)^T \, d\sigma(x) & \text{for} \quad i = j \\ -2r_{ij}^{-1} \int_{M_{ij}} f(x)(x - \mu_i)(x - \mu_j)^T \, d\sigma(x) & \text{for} \quad i \neq j, \end{cases} \tag{A.2}$$

for $i, j \in \{1, 2\}$ where $r_{ij} = \|\mu_i - \mu_j\|$, $f(\cdot)$ is the corresponding density function and $\sigma(\cdot)$ is the $(d-1)$ dimensional Lebesgue measure. Here A_i denotes the set of points in $\mathbb{R}^d$ closer to μ_i than to any other μ_j, M_{ij} denotes the face common to A_i and A_j and $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. We show that G_0 is positive definite in Lemma A.0.1.

Now using Corollary 6.5 in Bock (1985), Lemma 3.0.1 follows immediately. In the next section, Appendix A.1.1, we prove Claims (3.2) and (3.3) which together with Lemma 3.0.1 give Theorem 3.1.

A.1.1 Proof of Claims (3.2) and (3.3)

Proof of Claim (3.2): The vector $\boldsymbol{\mu}$ that minimizes the population within cluster sum of squares for the 2-means clustering has components given by

$$\mu_1 = \left(-\sigma_1\sqrt{\frac{2}{\pi}}, 0, \ldots, 0\right)^T, \quad \text{and}$$
$$\mu_2 = \left(\sigma_1\sqrt{\frac{2}{\pi}}, 0, \ldots, 0\right)^T.$$

The corresponding (optimal) population clusters are

$$A_1 = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 \leq 0\} \quad \text{and,}$$
$$A_2 = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 \geq 0\}.$$

Thus, it follows that,

$$\begin{aligned}
W(\boldsymbol{\mu}) &= \mathbb{E}\left[\|X - \mu_1\|^2 \mathbb{I}_{\{X_1<0\}}\right] + \mathbb{E}\left[\|X - \mu_2\|^2 \mathbb{I}_{\{X_1>0\}}\right] \\
&= 2\mathbb{E}\left[\|X - \mu_1\|^2 \mathbb{I}_{\{X_1<0\}}\right] \\
&= 2\left(\mathbb{E}\left[(X_1 - \mu_{11})^2 \mathbb{I}_{\{X_1<0\}}\right] + \sum_{i=2}^{d} \mathbb{E}\left[X_i^2 \mathbb{I}_{\{X_1<0\}}\right]\right) \\
&= 2\left(\frac{\sigma_1^2}{2}\left(1 - \frac{2}{\pi}\right) + \sum_{i=2}^{d} \frac{\sigma_i^2}{2}\right) \\
&= \sum_{i=1}^{d} \sigma_i^2 - \frac{2\sigma_1^2}{\pi},
\end{aligned}$$

which yields Claim (3.2).

Proof of Claim (3.3)**:** In a similar fashion we can compute τ^2. Observe that,

$$\begin{aligned}
\tau^2 + [W(\boldsymbol{\mu})]^2 &= \frac{1}{2}\left\{\mathbb{E}\left[\|X-\mu_1\|^4 \middle| X_1 < 0\right] + \mathbb{E}\left[\|X-\mu_2\|^4 \middle| X_1 > 0\right]\right\} \\
&= \mathbb{E}\left[\|X-\mu_1\|^4 \middle| X_1 < 0\right] \\
&= \mathbb{E}\left[\left((X_1-\mu_{11})^2 + \sum_{i=2}^{d} X_i^2\right)^2 \middle| X_1 < 0\right] \\
&= \mathbb{E}\left[(X_1-\mu_{11})^4 \middle| X_1 < 0\right] + 2\sum_{i=2}^{d} \mathbb{E}\left[(X_1-\mu_{11})^2 \middle| X_1 < 0\right] \mathbb{E}\left[X_i^2\right] \\
&\quad + 2\sum_{i=2}^{d}\sum_{j=2, j\neq i}^{d} \mathbb{E}\left[X_i^2\right]\mathbb{E}\left[X_j^2\right] + \sum_{i=2}^{d} \mathbb{E}\left[X_i^4\right] \\
&= \sigma_1^4\left(3 - \frac{4}{\pi} - \frac{12}{\pi^2}\right) + 2\sum_{i=2}^{d}\sigma_1^2\sigma_i^2\left(1-\frac{2}{\pi}\right) + 2\sum_{i=2}^{d}\sum_{j=2,j\neq i}^{d}\sigma_i^2\sigma_j^2 + 3\sum_{i=2}^{d}\sigma_i^4.
\end{aligned}$$

Plugging in the value of $W(\boldsymbol{\mu})$ we have,

$$\begin{aligned}
\tau^2 &= \sigma_1^4\left(2 - \frac{16}{\pi^2}\right) + 2\sum_{i=2}^{d}\sigma_i^4 \\
&= 2\sum_{i=1}^{d}\sigma_i^4 - \frac{16\sigma_1^4}{\pi^2},
\end{aligned}$$

which is precisely Claim (3.3). □

A.1.2 Proof of G_0 Being Positive Definite

In order to use the result in Pollard (1982) to prove Lemma 3.0.1 we need to verify that condition (v) holds. The vector $\boldsymbol{\mu}$ that minimizes the population within cluster sum of squares for the 2-means clustering is given above along with the two optimum population clusters. We additionally have that $M_{12} = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 = 0\}$, $\mathbb{P}(A_1) = \mathbb{P}(A_2) = 0.5$ and $r_{12} = 2\sigma_1\sqrt{\frac{2}{\pi}}$. The form of V and G_0 can then be given by:

Lemma A.0.1. *If $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ follows $N(0, \Sigma)$, that is, $\mathbb{P}$ is the distribution of $N(0, \Sigma)$, where Σ has diagonal elements $\sigma_1^2 > \sigma_2^2 \geq \sigma_3^2 \geq \ldots \geq \sigma_d^2 > 0$, then for $i, j \in \{1, 2\}$, $i \neq j$,*

$$V_i = \begin{pmatrix} \frac{\sigma_1^2}{2}\left(1-\frac{2}{\pi}\right) & 0 & \ldots & 0 \\ 0 & \frac{\sigma_2^2}{2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{\sigma_d^2}{2} \end{pmatrix}$$

and the matrix G_0 as defined in equation (A.2) is positive definite.

Proof of Lemma A.0.1. The different blocks of the variance matrix are given by,

$$V_1 = V_2 = \mathbb{E}\left[(X-\mu_1)(X-\mu_1)^T \mathbb{I}_{\{X_1<0\}}\right]$$
$$= \begin{pmatrix} \mathbb{E}\left[(X_1-\mu_{11})^2\mathbb{I}_{\{X_1<0\}}\right] & \mathbb{E}\left[(X_1-\mu_{11})X_2\mathbb{I}_{\{X_1<0\}}\right] & \dots & \mathbb{E}\left[(X_1-\mu_{11})X_d\mathbb{I}_{\{X_1<0\}}\right] \\ \mathbb{E}\left[(X_1-\mu_{11})X_2\mathbb{I}_{\{X_1<0\}}\right] & \mathbb{E}\left[X_2^2\mathbb{I}_{\{X_1<0\}}\right] & \dots & \mathbb{E}\left[X_2X_d\mathbb{I}_{\{X_1<0\}}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[(X_1-\mu_{11})X_d\mathbb{I}_{\{X_1<0\}}\right] & \mathbb{E}\left[X_2X_d\mathbb{I}_{\{X_1<0\}}\right] & \dots & \mathbb{E}\left[X_d^2\mathbb{I}_{\{X_1<0\}}\right] \end{pmatrix}$$

$$\begin{aligned} \mathbb{E}\left[(X_1-\mu_{11})^2\mathbb{I}_{\{X_1<0\}}\right] &= \mathbb{E}\left[(X_1^2 - 2\mu_{11}X_1 + \mu_{11}^2)\mathbb{I}_{\{X_1<0\}}\right] \\ &= \frac{1}{2}\mathbb{E}[X_1^2] - 2\mu_{11}\left(\frac{1}{2}\mu_{11}\right) + \frac{1}{2}\mu_{11}^2 \\ &= \frac{1}{2}\sigma_1^2 - \frac{1}{2}\left(\frac{2}{\pi}\sigma_1^2\right) \\ &= \frac{\sigma_1^2}{2}\left(1-\frac{2}{\pi}\right) \end{aligned}$$

For $j \neq 1$,

$$\mathbb{E}\left[(X_1-\mu_{11})X_j\mathbb{I}_{\{X_1<0\}}\right] = \mathbb{E}\left[(X_1-\mu_{11})\mathbb{I}_{\{X_1<0\}}\right]\mathbb{E}[X_j] = 0.$$

Let $i \neq j$, and $i, j \in \{2, \dots, d\}$,

$$\mathbb{E}\left[X_iX_d\mathbb{I}_{\{X_1<0\}}\right] = \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_d\right]\mathbb{E}\left[\mathbb{I}_{\{X_1<0\}}\right] = 0.$$

For $j \neq 1$,

$$\mathbb{E}\left[X_j^2\mathbb{I}_{\{X_1<0\}}\right] = \frac{1}{2}\mathbb{E}\left[X_j^2\right] = \frac{\sigma_j^2}{2}.$$

Therefore,

$$V_1 = V_2 = \begin{pmatrix} \frac{\sigma_1^2}{2}\left(1-\frac{2}{\pi}\right) & 0 & \dots & 0 \\ 0 & \frac{\sigma_2^2}{2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\sigma_d^2}{2} \end{pmatrix}.$$

Now for $x \in M_{12}$,

$$(x-\mu_1)(x-\mu_1)^T = \begin{pmatrix} (x_1-\mu_{11})^2 & (x_1-\mu_{11})(x_2-\mu_{12}) & \dots & (x_1-\mu_{11})(x_d-\mu_{1d}) \\ (x_1-\mu_{11})(x_2-\mu_{12}) & (x_2-\mu_{12})^2 & \dots & (x_2-\mu_{12})(x_d-\mu_{1d}) \\ \vdots & \vdots & \ddots & \vdots \\ (x_1-\mu_{11})(x_d-\mu_{1d}) & (x_2-\mu_{12})(x_d-\mu_{1d}) & \dots & (x_d-\mu_{1d})^2 \end{pmatrix}$$

$$= \begin{pmatrix} \mu_{11}^2 & -\mu_{11}x_2 & \dots & -\mu_{11}x_d \\ -\mu_{11}x_2 & x_2^2 & \dots & x_2x_d \\ \vdots & \vdots & \ddots & \vdots \\ -\mu_{11}x_d & x_2x_d & \dots & x_d^2 \end{pmatrix},$$

$$(x-\mu_1)(x-\mu_2)^T = \begin{pmatrix} (x_1-\mu_{11})(x_1-\mu_{21}) & (x_1-\mu_{11})(x_2-\mu_{22}) & \dots & (x_1-\mu_{11})(x_d-\mu_{2d}) \\ (x_2-\mu_{12})(x_1-\mu_{21}) & (x_2-\mu_{12})(x_2-\mu_{22}) & \dots & (x_2-\mu_{12})(x_d-\mu_{2d}) \\ \vdots & \vdots & \ddots & \vdots \\ (x_d-\mu_{1d})(x_1-\mu_{21}) & (x_d-\mu_{1d})(x_2-\mu_{22}) & \dots & (x_d-\mu_{1d})(x_d-\mu_{2d}) \end{pmatrix}$$

$$= \begin{pmatrix} -\mu_{11}^2 & -\mu_{11}x_2 & \dots & -\mu_{11}x_d \\ -\mu_{21}x_2 & x_2^2 & \dots & x_2x_d \\ \vdots & \vdots & \ddots & \vdots \\ -\mu_{21}x_d & x_2x_d & \dots & x_d^2 \end{pmatrix}.$$

Also, note that $\mathbb{I}_{M_{12}} = \mathbb{I}_{\{X \in M_{12}\}} = \mathbb{I}_{\{X_1=0\}}$. Therefore,

$$\mu_{11}^2 \int_{M_{12}} f(x)\ d\sigma(x) = \frac{2\sigma_1^2}{\pi} \frac{1}{\sqrt{2\pi}\sigma_1} = \sqrt{\frac{2}{\pi^3}}\sigma_1.$$

For $2 \le j \le d$,

$$\begin{aligned} \mu_{11} \int_{M_{12}} x_j\ f(x)\ d\sigma(x) &= \mu_{11}\ \mathbb{E}[X_j]\ \frac{1}{\sqrt{2\pi}\sigma_1} = 0, \\ \mu_{21} \int_{M_{12}} x_j\ f(x)\ d\sigma(x) &= \mu_{21}\ \mathbb{E}[X_j]\ \frac{1}{\sqrt{2\pi}\sigma_1} = 0, \\ \int_{M_{12}} x_j^2\ f(x)\ d\sigma(x) &= \mathbb{E}[X_j^2]\ \frac{1}{\sqrt{2\pi}\sigma_1} = \frac{\sigma_j^2}{\sqrt{2\pi}\sigma_1}. \end{aligned}$$

Let $i \neq j$, and $i, j \in \{2, \dots, d\}$,

$$\int_{M_{12}} x_i\ x_j\ f(x)\ d\sigma(x) = \mathbb{E}[X_i]\ \mathbb{E}[X_j]\ \frac{1}{\sqrt{2\pi}\sigma_1} = 0.$$

Then the matrix G_0 can be derived as,

$$(G_0)_{22} = (G_0)_{11} = \mathbf{I}_d - \frac{1}{\sigma_1}\sqrt{\frac{\pi}{2}} \begin{pmatrix} \sqrt{\frac{2}{\pi^3}}\sigma_1 & 0 & \dots & 0 \\ 0 & \frac{\sigma_2^2}{\sqrt{2\pi}\sigma_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\sigma_d^2}{\sqrt{2\pi}\sigma_1} \end{pmatrix} = \begin{pmatrix} 1-\frac{1}{\pi} & 0 & \dots & 0 \\ 0 & 1-\frac{\sigma_2^2}{2\sigma_1^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1-\frac{\sigma_d^2}{2\sigma_1^2} \end{pmatrix},$$

$$(G_0)_{21} = (G_0)_{12} = -\frac{1}{\sigma_1}\sqrt{\frac{\pi}{2}} \begin{pmatrix} -\sqrt{\frac{2}{\pi^3}}\sigma_1 & 0 & \dots & 0 \\ 0 & \frac{\sigma_2^2}{\sqrt{2\pi}\sigma_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\sigma_d^2}{\sqrt{2\pi}\sigma_1} \end{pmatrix} = \begin{pmatrix} \frac{1}{\pi} & 0 & \dots & 0 \\ 0 & -\frac{\sigma_2^2}{2\sigma_1^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -\frac{\sigma_d^2}{2\sigma_1^2} \end{pmatrix}.$$

Using the result from Boyd and Vandenberghe (2004), we have that the symmetric matrix G_0 is positive definite if and only if $(G_0)_{11}$ and $G_0/(G_0)_{11}$ (the Schur complement of $(G_0)_{11}$ in G_0) are both positive definite. $(G_0)_{11}$ is a diagonal matrix with strictly positive entries on its diagonal since $\sigma_1^2 > \sigma_j^2$ for $j \neq 1$. Therefore, $(G_0)_{11}$ is trivially a positive definite matrix. To show $G_0/(G_0)_{11}$ is also positive definite first we simplify it.

$$\begin{aligned} G_0/(G_0)_{11} &= (G_0)_{22} - (G_0)_{21}\left[(G_0)_{11}\right]^{-1}(G_0)_{12} \\ &= \begin{pmatrix} 1-\frac{1}{\pi} & 0 & \dots & 0 \\ 0 & 1-\frac{\sigma_2^2}{2\sigma_1^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1-\frac{\sigma_d^2}{2\sigma_1^2} \end{pmatrix} - \begin{pmatrix} \frac{1}{\pi^2} & \frac{\pi}{\pi-1} & 0 & \dots & 0 \\ 0 & \frac{\sigma_2^4}{4\sigma_1^4} & \frac{2\sigma_1^2}{2\sigma_1^2-\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{\sigma_d^4}{4\sigma_1^4} \; \frac{2\sigma_1^2}{2\sigma_1^2-\sigma_d^2} \end{pmatrix} \\ &= \begin{pmatrix} 1-\frac{1}{\pi-1} & 0 & \dots & 0 \\ 0 & \frac{2(\sigma_1^2-\sigma_2^2)}{2\sigma_1^2-\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{2(\sigma_1^2-\sigma_d^2)}{2\sigma_1^2-\sigma_d^2} \end{pmatrix}, \end{aligned}$$

which is again a diagonal matrix with strictly positive entries on its diagonal since $\sigma_1^2 > \sigma_j^2$ for $j \neq 1$. Therefore, $G_0/(G_0)_{11}$ is also a positive definite matrix, which implies G_0 itself is a positive definite matrix.

□

A.1.3 Limiting Distribution Under the Null When $\sigma_1^2 = \sigma_2^2$

We will generally focus on the non-spherical case since it yields tractable limiting distributions. But here we briefly mention what happens when the null distribution is spherical. In this case, the limiting distribution is quite complicated, For simplicity, we only consider the special case $d = 2$. To find the distribution of the test statistic, we first find the distribution of the between-cluster sum of squares, where the between-cluster sum of squares for a partition given by centers $\mathbf{a} = (a_1, \ldots, a_k)$ and the set of corresponding convex polyhedrons $A_1, \ldots, A_k$ is defined as,

$$B_n(a) = \frac{1}{n}\sum_{j=1}^{k} n_j \|a_j - \overline{X}\|^2, \quad n_j = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A_j\}}.$$

When 2-means clustering is applied to two dimensional data, the two partitions can also be uniquely identified using the separating line dividing them. The line containing the optimal centers is perpendicular to this line. Consider the line joining the centers and the point where it meets the separating line, say p. This line can uniquely be identified by the angle the line makes with the x-axis, β, and its distance from the origin, c. Therefore, instead of defining between-cluster sum of squares as a function of the centers of the partition, we can also define it as a function of β, c and p denoted by $B_n(\beta, c, p)$. Therefore corresponding to the two centers of the optimal partition $\mathbf{b_n} = (b_{n1}, b_{n2})$, we can also find the optimal hyperplane for the data, denoted by (β_n, c_n, p_n).

We perform a 2-means clustering on the data which finds the optimal partition of the data in order to minimize the within-cluster sum of squares $W_n(\mathbf{b_n})$ and maximize the between-cluster sum of squares denoted by $B_n(\beta_n, c_n, p_n)$. Then,

$$B_n(\beta_n, c_n, p_n) = \max_{\beta} \max_{c} \max_{p} B_n(\beta, c, p).$$

We also define $B_n(\beta) = \max_c \max_p B_n(\beta, c, p)$.

Theorem A.1. *If $X_1, \ldots, X_n \sim N(0, \Sigma)$, $X_i \in \mathbb{R}^2$, where Σ has diagonal elements $\sigma_1^2 = \sigma_2^2 = 1$, then $\sqrt{n}(B_n(\beta_n, c_n, p_n) - 2/\pi)$ is asymptotically distributed as the maximum of a Gaussian process $Z(\beta)$ on the circle $0 \le \beta < 2\pi$, where $Z(\beta)$ has mean 0 and the covariance between $Z(\beta)$ and $Z(\phi)$ is given by*

$$\left\{ \frac{8}{\pi^2} \left(\sin\alpha + (\pi - \alpha)\cos\alpha - 2 \right) \right\}, \quad \alpha = |\beta - \phi| \le \pi.$$

Note that $B_n(\beta_n, c_n, p_n) = \max_\beta \max_c \max_p B_n(\beta, c, p) = \max_\beta B_n(\beta)$. Let $Z(\beta) = \sqrt{n}\left(B_n(\beta) - \frac{2}{\pi}\right)$, then $\sqrt{n}\left(B_n(\beta_n, c_n) - \frac{2}{\pi}\right) = \max_\beta Z(\beta)$. Then the proof of the theorem follows directly from Lemmas A.1.1 and A.1.2 stated and proved below. □

Lemma A.1.1. *If* $X_1, \ldots, X_n \sim N(0, \Sigma)$, $X_i \in \mathbb{R}^2$, *where* Σ *has diagonal elements* $\sigma_1^2 = \sigma_2^2 = 1$, *then* $\forall\ 0 \leq \beta < 2\pi$,

$$\sqrt{n}\left(B_n(\beta) - \frac{2}{\pi}\right) \rightsquigarrow N\left(0, \frac{8}{\pi}\left(1 - \frac{2}{\pi}\right)\right) \quad as \quad n \to \infty.$$

Proof of Lemma A.1.1. As the bivariate circular normal is invariant to the angle β, without loss of generality we can assume $\beta = 0$. Then the optimal centers of the partition b_{n1}, b_{n2} lie on a line parallel to the x-axis. Now if we condition on c_n, then the line containing the centers is deterministic and hence the between-cluster sum of squares after performing 2-means clustering on the data is same as the between-cluster sum of squares after projecting the data onto the line joining the centers. So $B_n(\beta)$ is the same as between-cluster sum of squares for $Y_1, \ldots, Y_n$ where Y_i has the same distribution as X_{i1}. Now $Y_i's$ are univariate with $Y_i \sim N(0, 1)$.

Hartigan (1978) showed that for univariate normal data, $Y_1, \ldots Y_n \sim N(0, 1)$, on performing 2-means clustering the asymptotic distribution of between-cluster sum of squares, $B_n(\mathbf{b_n})$ can be given as

$$\sqrt{n}\left(B_n(\mathbf{b_n}) - \frac{2}{\pi}\right) \rightsquigarrow N\left(0, \frac{8}{\pi}\left(1 - \frac{2}{\pi}\right)\right) \quad \text{as} \quad n \to \infty,$$

where $\mathbf{b_n}$ is the vector of cluster centers for the optimal partition. Therefore,

$$\sqrt{n}\left(B_n(\beta) - \frac{2}{\pi}\right)\Big| c_n \rightsquigarrow N\left(0, \frac{8}{\pi}\left(1 - \frac{2}{\pi}\right)\right) \quad \text{as} \quad n \to \infty,$$

which does not depend on c_n. Hence,

$$\sqrt{n}\left(B_n(\beta) - \frac{2}{\pi}\right) \rightsquigarrow N\left(0, \frac{8}{\pi}\left(1 - \frac{2}{\pi}\right)\right) \quad \text{as} \quad n \to \infty. \ \square$$

Lemma A.1.2. *The asymptotic covariance between* $\sqrt{n}B_n(\beta)$ *and* $\sqrt{n}B_n(\phi)$ *is given by,*

$$\lim_{n\to\infty} n\ Cov(B_n(\beta), B_n(\phi)) = \frac{16}{\pi^2}\left(\sin\alpha + \left(\frac{\pi}{2} - \alpha\right)\cos\alpha - 1\right), \quad \alpha = |\beta - \phi| \leq \pi.$$

Proof of Lemma A.1.2. Hartigan (1978) provides a Taylor's expansion of $B_n(\beta)$ about the population between-cluster sum of squares $B_n(\mu) = \frac{2}{\pi}$ as,

$$B_n(\beta) = \frac{2}{\pi} + \frac{1}{2}\left(\|b_{n1}(\beta) - b_{n2}(\beta)\|_2 - \|\mu_1(\beta) - \mu_2(\beta)\|_2\right)\|\mu_1(\beta) - \mu_2(\beta)\|_2 + o_p(n^{-1}),$$

where $(b_{n1}(\beta), b_{n2}(\beta))$ are the centers for the optimal partition of the data corresponding to $B_n(\beta)$ and (μ_1, μ_2) are the centers for the optimal partition of the entire population. For $\beta = 0$, the optimal centers are $\mu_1(0) = (-\sqrt{2/\pi}, 0)$ and $\mu_2(0) = (\sqrt{2/\pi}, 0)$ and hence $\|\mu_1(0) - \mu_2(0)\|_2 = 2\sqrt{\frac{2}{\pi}}$. Also as the density of

$N(0, \mathbf{I}_d)$ is rotationally invariant, $\|\mu_1(\beta) - \mu_2(\beta)\|_2 = 2\sqrt{\frac{2}{\pi}}$ for any β. Therefore,

$$B_n(\beta) = \frac{2}{\pi} + \left(\|b_{n1}(\beta) - b_{n2}(\beta)\|_2 - 2\sqrt{\frac{2}{\pi}} \right) \sqrt{\frac{2}{\pi}} + o_p(n^{-1}).$$

Due to the rotational invariance of the bivariate circular normal, $\mathrm{Cov}(B_n(\beta), B_n(\phi)) = \mathrm{Cov}(B_n(0), B_n(\alpha))$, where $\alpha = |\beta - \phi| \leq \pi$. Therefore it is enough to consider $\mathrm{Cov}(B_n(0), B_n(\alpha))$ where,

$$\lim_{n\to\infty} n \ \mathrm{Cov}(B_n(0), B_n(\alpha)) = \lim_{n\to\infty} \frac{2}{\pi} \ \mathrm{Cov}\left(\sqrt{n} \ \|b_{n1}(0) - b_{n2}(0)\|_2, \sqrt{n} \ \|b_{n1}(\alpha) - b_{n2}(\alpha)\|_2\right).$$

To find $\|b_{n1}(\alpha) - b_{n2}(\alpha)\|_2$, if we rotate the axes by an angle of $-\alpha$, any point (X_{i1}, X_{i2}) is now given by

$$(X_{i1}\cos\alpha + X_{i2}\sin\alpha, X_{i2}\cos\alpha - X_{i1}\sin\alpha).$$

For easier notation let us define $Z_i := X_{i1}\cos\alpha + X_{i2}\sin\alpha$ for $i = 1, 2, \ldots, n$. Let us also define $p_n := \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{X_{i1}>0\}}$ and $p_n' := \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{Z_i>0\}}$. Then,

$$\|b_{n1}(0) - b_{n2}(0)\|_2 = \frac{\sum_{i=1}^n X_{i1}\mathbb{I}_{\{X_{i1}>0\}}}{np_n} - \frac{\sum_{i=1}^n X_{i1}\mathbb{I}_{\{X_{i1}<0\}}}{n(1-p_n)},$$

$$\text{and} \quad \|b_{n1}(\alpha) - b_{n2}(\alpha)\|_2 = \frac{\sum_{i=1}^n Z_i\mathbb{I}_{\{Z_i>0\}}}{np_n'} - \frac{\sum_{i=1}^n Z_i\mathbb{I}_{\{Z_i<0\}}}{n(1-p_n')}.$$

Using the Law of Total Covariance,

$$\begin{aligned}
&\mathrm{Cov}\left(\|b_{n1}(0) - b_{n2}(0)\|_2, \|b_{n1}(\alpha) - b_{n2}(\alpha)\|_2\right) \\
&\quad = \mathbb{E}\left[\mathrm{Cov}\left(\|b_{n1}(0) - b_{n2}(0)\|_2, \|b_{n1}(\alpha) - b_{n2}(\alpha)\|_2 \,|\, p_n, p_n'\right)\right] \\
&\qquad + \mathrm{Cov}\left(\mathbb{E}\left[\|b_{n1}(0) - b_{n2}(0)\|_2 \,|\, p_n\right], \mathbb{E}\left[\|b_{n1}(\alpha) - b_{n2}(\alpha)\|_2 \,|\, p_n'\right]\right) \\
&\quad = I + II \quad \text{(say)}.
\end{aligned}$$

The second term (II) can be easily simplified as

$$\begin{aligned}
&\mathrm{Cov}\left(\mathbb{E}\left[\frac{\sum_{i=1}^n X_{i1}\mathbb{I}_{\{X_{i1}>0\}}}{np_n} - \frac{\sum_{i=1}^n X_{i1}\mathbb{I}_{\{X_{i1}<0\}}}{n(1-p_n)} \,\middle|\, p_n\right], \mathbb{E}\left[\frac{\sum_{i=1}^n Z_i\mathbb{I}_{\{Z_i>0\}}}{np_n'} - \frac{\sum_{i=1}^n Z_i\mathbb{I}_{\{Z_i<0\}}}{n(1-p_n')} \,\middle|\, p_n'\right]\right) \\
&\quad = \mathrm{Cov}\left(\mathbb{E}[X_{i1}\mathbb{I}_{\{X_{i1}>0\}}] - \mathbb{E}[X_{i1}\mathbb{I}_{\{X_{i1}<0\}}], \mathbb{E}[Z_i\mathbb{I}_{\{Z_i>0\}}] - \mathbb{E}[Z_i\mathbb{I}_{\{Z_i<0\}}]\right) \\
&\quad = \mathrm{Cov}\left(2\frac{1}{\sqrt{2\pi}}, 2\frac{1}{\sqrt{2\pi}}\right) = 0.
\end{aligned}$$

The first term (I) becomes,

$$\begin{aligned}
I &= \mathbb{E}\left[\mathrm{Cov}\left(\frac{\sum_{i=1}^n X_{i1}\mathbb{I}_{\{X_{i1}>0\}}}{np_n} - \frac{\sum_{i=1}^n X_{i1}\mathbb{I}_{\{X_{i1}<0\}}}{n(1-p_n)}, \frac{\sum_{i=1}^n Z_i\mathbb{I}_{\{Z_i>0\}}}{np'_n} - \frac{\sum_{i=1}^n Z_i\mathbb{I}_{\{Z_i<0\}}}{n(1-p'_n)}\,\middle|\, p_n, p'_n\right)\right] \\
&= \mathbb{E}\left[\mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1>0\right)\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}>0, Z_i>0\}}}{n^2 p_n p'_n}\right] \\
&\qquad - \mathbb{E}\left[\mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1<0\right)\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}>0, Z_i<0\}}}{n^2 p_n (1-p'_n)}\right] \\
&\qquad - \mathbb{E}\left[\mathrm{Cov}\left(X_{11}, Z_1 | X_{11}<0, Z_1>0\right)\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}<0, Z_i>0\}}}{n^2 (1-p_n) p'_n}\right] \\
&\qquad + \mathbb{E}\left[\mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1>0\right)\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}<0, Z_i<0\}}}{n^2 (1-p_n)(1-p'_n)}\right].
\end{aligned}$$

Due to the symmetry of the Gaussian distribution about the origin,

$$\mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1>0\right) = \mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1>0\right),$$

$$\text{and} \quad \mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1<0\right) = \mathrm{Cov}\left(X_{11}, Z_1 | X_{11}<0, Z_1>0\right).$$

Hence the first term becomes,

$$\begin{aligned}
I = \mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1>0\right)\mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}>0, Z_i>0\}}}{n^2 p_n p'_n} + \frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}<0, Z_i<0\}}}{n^2 (1-p_n)(1-p'_n)}\right] \\
- \mathrm{Cov}\left(X_{11}, Z_1 | X_{11}>0, Z_1<0\right)\mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}>0, Z_i<0\}}}{n^2 p_n (1-p'_n)} + \frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}<0, Z_i>0\}}}{n^2 (1-p_n) p'_n}\right]
\end{aligned}$$

As $np_n \sim \mathrm{Bin}(n, 0.5)$ and $np'_n \sim \mathrm{Bin}(n, 0.5)$, $1/p_n \xrightarrow{p} 2$ and $1/p'_n \xrightarrow{p} 2$. Hence by Slutsky's theorem and weak law of large numbers,

$$\begin{aligned}
\lim_{n\to\infty} n\,\mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}>0, Z_i>0\}}}{n^2 p_n p'_n} + \frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}<0, Z_i<0\}}}{n^2 (1-p_n)(1-p'_n)}\right] &= 4\left(\mathbb{P}(X_{11}>0, Z_1>0) + \mathbb{P}(X_{11}<0, Z_1<0)\right) \\
&= 8\mathbb{P}(X_{11}>0, Z_1>0).
\end{aligned}$$

Similarly using Slutsky's theorem and weak law of large numbers,

$$\lim_{n\to\infty} n\,\mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}>0, Z_i<0\}}}{n^2 p_n (1-p'_n)} + \frac{\sum_{i=1}^n \mathbb{I}_{\{X_{i1}<0, Z_i>0\}}}{n^2 (1-p_n) p'_n}\right] = 8\mathbb{P}(X_{11}>0, Z_1<0).$$

On the other hand, we can find the covariances as,

$$\begin{aligned}
\text{Cov}\left(X_{11}, Z_1 | X_{11} > 0, Z_1 > 0\right) &= \mathbb{E}\left[X_{11} Z_1 | X_{11} > 0, Z_1 > 0\right] - \mathbb{E}[X_{11} | X_{11} > 0]\mathbb{E}[Z_1 | Z_1 > 0] \\
&= \frac{\mathbb{E}\left[X_{11} Z_1 \mathbb{I}_{\{X_{11}>0, Z_1>0\}}\right]}{\mathbb{P}(X_{11} > 0, Z_1 > 0)} - \mathbb{E}[X_{11} | X_{11} > 0]\mathbb{E}[Z_1 | Z_1 > 0], \\
\text{Cov}\left(X_{11}, Z_1 | X_{11} > 0, Z_1 < 0\right) &= \frac{\mathbb{E}\left[X_{11} Z_1 \mathbb{I}_{\{X_{11}>0, Z_1<0\}}\right]}{\mathbb{P}(X_{11} > 0, Z_1 < 0)} - \mathbb{E}[X_{11} | X_{11} > 0]\mathbb{E}[Z_1 | Z_1 < 0].
\end{aligned}$$

As $X_{11}, X_{12} \sim N(0,1)$, it implies $Z_1 = X_{11}\cos\alpha + X_{12}\sin\alpha \sim N(0,1)$. Hence,

$$\mathbb{E}[X_{11} | X_{11} > 0] = \mathbb{E}[Z_1 | Z_1 > 0] = \sqrt{\frac{2}{\pi}},$$

$$\text{and} \quad \mathbb{E}[Z_1 | Z_1 < 0] = -\sqrt{\frac{2}{\pi}}.$$

To find the first expectation, we define $R = \sqrt{X_{11}^2 + X_{12}^2}$ and $\beta = \tan^{-1}\left(X_{12}/X_{11}\right)$ such that $X_{11} = R\cos\beta$ and $X_{12} = R\sin\beta$. Then the Jacobian, $\partial(x_1, x_2)/\partial(r, \beta)$ can be given by,

$$\frac{\partial(x_1, x_2)}{\partial(r, \beta)} = \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \beta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \beta} \end{vmatrix} = \begin{vmatrix} \cos\beta & -r\sin\beta \\ \sin\beta & r\cos\beta \end{vmatrix} = r.$$

Also $x_1 > 0$ can be written as $\beta \in [-\pi/2, \pi/2]$ and assuming $0 < \alpha < \pi/2$, $x_1\cos\alpha + x_2\sin\alpha = r\cos(\beta - \alpha)$ and $x_1\cos\alpha + x_2\sin\alpha > 0$ can be written as $\beta - \alpha \in [-\pi/2, \pi/2]$ or $\beta \in [\alpha - (\pi/2), \alpha + (\pi/2)]$. $x_1\cos\alpha + x_2\sin\alpha < 0$ can be written as $\beta - \alpha \in [-3\pi/2, -\pi/2]$ or $\beta \in [\alpha - 3\pi/2, \alpha - \pi/2]$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[X_{11} Z_1 \mathbb{I}_{\{X_{11}>0, Z_1>0\}}\right] &= \mathbb{E}\left[X_{11}(X_{11}\cos\alpha + X_{12}\sin\alpha)\mathbb{I}_{\{X_{11}>0\}}\mathbb{I}_{\{X_{11}\cos\alpha + X_{12}\sin\alpha > 0\}}\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1(x_1\cos\alpha + x_2\sin\alpha)\mathbb{I}_{\{x_1>0\}}\mathbb{I}_{\{x_1\cos\alpha + x_2\sin\alpha > 0\}}\ \frac{1}{2\pi} e^{-\left(\frac{x_1^2}{2} + \frac{x_2^2}{2}\right)} dx_1\ dx_2 \\
&= \int_0^{\infty}\int_{\alpha - \frac{\pi}{2}}^{\frac{\pi}{2}} r\cos\beta\ r\cos(\beta - \alpha)\ \frac{1}{2\pi} e^{-\frac{r^2}{2}} r\ d\beta\ dr \\
&= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} r^3\ \frac{1}{\sqrt{2\pi}}\ e^{-\frac{r^2}{2}}\ dr\ \int_{\alpha - \frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2}\left(\cos(2\beta - \alpha) + \cos\alpha\right)\ d\beta \\
&= \frac{1}{\sqrt{2\pi}}\ \sqrt{\frac{2}{\pi}}\ \int_{\alpha - \frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2}\left(\cos(2\beta - \alpha) + \cos\alpha\right)\ d\beta \\
&= \frac{1}{2\pi}\left(\frac{\sin(2\beta - \alpha)}{2} + \beta\cos\alpha\right)\Bigg|_{\alpha - \frac{\pi}{2}}^{\frac{\pi}{2}} \\
&= \frac{1}{2\pi}\left(\frac{\sin(\pi - \alpha) - \sin(\alpha - \pi)}{2} + (\pi - \alpha)\cos\alpha\right) \\
&= \frac{1}{2\pi}\left(\sin\alpha + (\pi - \alpha)\cos\alpha\right)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[X_{11}Z_1\mathbb{I}_{\{X_{11}>0,Z_1<0\}}\right] &= \mathbb{E}\left[X_{11}(X_{11}\cos\alpha + X_{12}\sin\alpha)\mathbb{I}_{\{X_{11}>0\}}\mathbb{I}_{\{X_{11}\cos\alpha+X_{12}\sin\alpha<0\}}\right] \\
&= \frac{1}{2\pi}\int_{-\frac{\pi}{2}}^{\alpha-\frac{\pi}{2}} \cos(2\beta-\alpha)+\cos\alpha\ d\beta \\
&= \frac{1}{2\pi}\left(\frac{\sin(2\beta-\alpha)}{2}+\beta\cos\alpha\right)\Bigg|_{-\frac{\pi}{2}}^{\alpha-\frac{\pi}{2}} \\
&= \frac{1}{2\pi}\left(\frac{\sin(\alpha-\pi)-\sin(-\alpha-\pi)}{2}+\alpha\cos\alpha\right) \\
&= \frac{1}{2\pi}\left(\frac{\sin(\pi+\alpha)-\sin(\pi-\alpha)}{2}+\alpha\cos\alpha\right) \\
&= \frac{1}{2\pi}\left(\alpha\cos\alpha-\sin\alpha\right)
\end{aligned}$$

Plugging in all the derivations we get,

$$\begin{aligned}
\lim_{n\to\infty} nI &= \left(\frac{\frac{1}{2\pi}\left(\sin\alpha+(\pi-\alpha)\cos\alpha\right)}{\mathbb{P}(X_{11}>0,Z_1>0)}-\frac{2}{\pi}\right)8\mathbb{P}(X_{11}>0,Z_1>0) \\
&\qquad -\left(\frac{\frac{1}{2\pi}\left(\alpha\cos\alpha-\sin\alpha\right)}{\mathbb{P}(X_{11}>0,Z_1<0)}+\frac{2}{\pi}\right)8\mathbb{P}(X_{11}>0,Z_1<0) \\
&= \frac{4}{\pi}(2\sin\alpha+(\pi-2\alpha)\cos\alpha)-\frac{16}{\pi}\left(\mathbb{P}(X_{11}>0,Z_1>0)+\mathbb{P}(X_{11}>0,Z_1<0)\right) \\
&= \frac{8}{\pi}\left(\sin\alpha+\left(\frac{\pi}{2}-\alpha\right)\cos\alpha\right)-\frac{8}{\pi}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty} n\ \mathrm{Cov}\left(B_n(0),B_n(\alpha)\right) &= \frac{2}{\pi}\left(\frac{8}{\pi}\left(\sin\alpha+\left(\frac{\pi}{2}-\alpha\right)\cos\alpha\right)-\frac{8}{\pi}\right) \\
&= \frac{16}{\pi^2}\left(\sin\alpha+\left(\frac{\pi}{2}-\alpha\right)\cos\alpha-1\right). \ \square
\end{aligned}$$

We also note that setting $\alpha=0$, gives us $\lim_{n\to\infty} n\ \mathrm{V}(B_n(\beta)) = \frac{8}{\pi}\left(1-\frac{2}{\pi}\right)$. $\square$

A.2 Proof of Results Presented Under the Alternate Hypothesis of SigClust

In Section 3.2, we study the geometry of $k-$means under the alternative. Recall, under the alternative of SigClust we suppose that, we observe n samples from:

$$X \sim \frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D) \tag{A.3}$$

where $\theta_1 = (a/2, 0, \ldots, 0) \in \mathbb{R}^d$ and $a > 0$. Furthermore, D is a diagonal matrix with elements $\Sigma_{jj} = \sigma_j^2$, such that $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$. Throughout this Appendix and the next, we denote,

$$\begin{aligned} u &:= \frac{a}{2\sigma_1}, \\ \kappa &:= \left[\frac{a}{2}\mathbb{P}(|Z| \leq u) + \sqrt{\frac{2}{\pi}}\sigma_1 \exp(-u^2/2)\right], \\ \widetilde{\sigma}^2 &= \left[\sum_{i=1}^{d} \sigma_i^2\right] + \frac{a^2}{4}. \end{aligned}$$

In this Appendix, in Section A.2.1 we first prove Theorem 3.2, which is a result analogous to Theorem 6.4 (b) of Bock (1985, pp. 101) for symmetric 2-means clustering, that gives the limiting distribution of the within sum of squares under the alternative. This Theorem assumes two things: first, the existence of a unique minimizer of the within sum of squares and second, the positive definiteness of the matrix G defined in equation (3.6).

We prove Lemma 3.3.1 that shows the positive definiteness of G in Appendix A.2.3. In Section A.2.2, we prove Theorem 3.3 that gives the optimal population split which results in the minimum within sum of squares under the alternative. The idea behind the proof is that when condition (3.7) is true, the population-level optimal 2-means solution is unique and is given by:

$$\boldsymbol{\mu}^* = \left(\begin{bmatrix} \kappa \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} -\kappa \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right), \tag{A.4}$$

and when condition (3.9) is true then the population-level optimal 2-means solution is unique and is given by:

$$\boldsymbol{\mu}^* = \left(\begin{bmatrix} 0 \\ \sqrt{\frac{2}{\pi}}\sigma_2 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\sqrt{\frac{2}{\pi}}\sigma_2 \\ \vdots \\ 0 \end{bmatrix}, \right). \tag{A.5}$$

Now the reason that we have $\sqrt{\frac{2}{\pi}}\sigma_2$ in equation (A.5) is because $E[X_2|X_2 > 0] = \sqrt{\frac{2}{\pi}}\sigma_2$ and similarly we have κ in equation (A.4) because $E[X_1|X_1 > 0] = \kappa$, which is given by the following lemma:

Lemma A.1.3. *Suppose that*

$$Y \sim \frac{1}{2}N(-a/2, \sigma^2) + \frac{1}{2}N(a/2, \sigma^2),$$

and that $Z \sim N(0, 1)$*, then we have that,*

$$\mathbb{E}[Y|Y > 0] = \frac{a}{2}\mathbb{P}(|Z| \leq u) + \sqrt{\frac{2}{\pi}}\sigma_1 \exp(-u^2/2) = \kappa.$$

Additionally, in order to find the within sum of squares for a particular split, we first introduce two lemmas that give the resulting within sum of squares $W(b)$ corresponding to particular forms of separating hyperplanes. More specifically, the following lemmas give the within sum of squares $W(b)$ corresponding to any separating hyperplane $\mathcal{H}(b)$ where b is of the form $\{b \in \mathbb{R}^d : b_1 \geq 0,\ \sum_{i=1}^d b_i^2 = 1\}$. Recollect that,

$$W(b) = E\left[\|X - E[X|b^T X > 0]\|^2 | b^T X > 0\right].$$

Lemma A.1.4. *For a separating hyperplane* $\mathcal{H}(b) = \{y \in \mathbb{R}^d : b^T y = 0\}$ *when*

$$b \in \{b \in \mathbb{R}^d : b_1 \geq 0,\ b_1^2 + b_2^2 = 1,\ b_j = 0\ \forall\ j \geq 3\},$$

the corresponding within sum of squares $W(b)$ *is given by:*

$$W(b) = \sum_{j=1}^d \sigma_j^2 + \frac{a^2}{4} - \left[\left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 - \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right),$$

where $\phi(\cdot)$ *and* $\Phi(\cdot)$ *are respectively the density function and the distribution function of standard normal distribution.*

Lemma A.1.5. *For any fixed $i \geq 2$ and a separating hyperplane $\mathcal{H}(b) = \{y \in \mathbb{R}^d : b^T y = 0\}$, when*

$$b \in \{b \in \mathbb{R}^d : b_i = 1 \quad \textit{and} \quad b_j = 0 \ \forall \ j \neq i\},$$

the corresponding within sum of squares $W(b)$ is given by:

$$W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_i^2.$$

Furthermore, to prove Theorem 3.3 we extend projection arguments made in Qiu (2010) to the $d-$dimensional scenario. So we provide a lemma that gives the projection of a cluster center onto the separating hyperplane.

Lemma A.1.6. *If $Y \sim N(\theta_1, D)$, where $\theta_1 = (a/2, 0, \ldots, 0) \in \mathbb{R}^d$ and D is a diagonal matrix. Then the i^{th} coordinate of the projection of $E[Y|b^T Y > 0]$ onto the separating hyperplane $\mathcal{H}(b)$ when $\sum_{i=1}^{d} b_i^2 = 1$, is given by:*

$$\mathcal{P}_i = \frac{a}{2}\mathbb{I}\{i = 1\} - \frac{ab_1 b_i}{2} + \frac{b_i Var(Y_i) - b_i \left(b^T D b\right)}{b^T D b} \left(E\left[b^T Y \middle| b^T Y > 0\right] - \frac{ab_1}{2}\right).$$

Finally, in order to compare the resulting within sum of squares from different separating hyperplanes we need the following lower bound on κ^2:

Lemma A.1.7.

$$\kappa^2 - \frac{2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4}\right) \geq \begin{cases} \frac{a^4}{240\sigma^2\pi} & \textit{for} \quad 0 \leq a \leq 4\sigma_1 \\ \frac{a^2}{40} & \textit{for} \quad a \geq 4\sigma_1. \end{cases}$$

We include the proofs of all of these additional lemmas that help us prove Theorem 3.3 in Appendix A.2.4.

A.2.1 Proof of Theorem 3.2

In order to prove this Theorem, we trace the steps followed by Pollard (1982) to prove their main theorem. First we define for every vector $\mathbf{t} = [t_1, t_2] \in \mathbb{R}^{2d}$ and every $x \in \mathbb{R}^d$,

$$\phi(x, \mathbf{t}) = \min\{\|x - t_1\|^2, \|x - t_2\|^2\}, \tag{A.6}$$

as defined by Pollard (1982). We additionally define a symmetric version of the function for $t^* \in \mathbb{R}^d$ and every $x \in \mathbb{R}^d$,

$$\widetilde{\phi}(x, t^*) = \phi(x, (t^*, -t^*)) = \min\{\|x - t^*\|^2, \|x + t^*\|^2\}. \tag{A.7}$$

Let us also define a map T from $\mathbb{R}^d$ to $\mathbb{R}^{2d}$ as $T(t^*) = (t^*, -t^*)$. Then,

$$\widetilde{\phi}(x, t^*) = \phi(x, T(t^*)).$$

Now note that an analogous version of Lemma A in Pollard (1982) also holds for the map $t^* \to \widetilde{\phi}(\cdot, t^*)$ as the composite function of two differentiable functions is also differentiable. Now Lemma B in Pollard (1982) also holds for $\widetilde{\phi}(\cdot, t^*)$ since the class of functions that are to be considered now is a subset of the class of functions ($\mathcal{G}$) considered for $\phi(x, \mathbf{t})$, as we only consider $\mathbf{t}$ of the form $\mathbf{t} = (t^*, -t^*)$ for some $t^* \in \mathbb{R}^d$. Since a subset of a Donsker class is also a Donsker class, an analogous Lemma B holds for $\widetilde{\phi}(\cdot, t^*)$.

We can also derive an analogous version of Lemma C and Lemma D by just using the chain rule for finding the second derivative of a composite function and using the results as obtained in Lemma A and B. Now putting them all together we can derive an analogous version of the main theorem.

Note that the assumptions (i) and (v) of the main theorem in Pollard (1982) are the same as the ones assumed here. The assumptions (ii) - (iv) are met by the mixtures of two Normals assumed in the statement. Therefore, following the arguments presented in the proof of the main theorem in Pollard (1982) and in the proof of Theorem 6.4 (b) on page 101 of Bock (1985), we arrive at the result that:

$$\sqrt{n}(W_n^{(0)}(\mathbf{b_n}^{(0)}) - W(\mu^*)) \rightsquigarrow N(0, \tau^{*2}).$$

□

A.2.2 Proof of Theorem 3.3

To prove this lemma, we extend projection arguments made in Qiu (2010) to the $d-$dimensional scenario. As shown earlier, for any separating hyperplane, $\mathcal{H}(b) = \{y \in \mathbb{R}^d : b^T y = 0\}$ when $b_1 \geq 0$ and $\sum_{i=1}^d b_i^2 = 1$, the corresponding within sum of squares is given by:

$$\begin{aligned} W(b) &= P(b^T X > 0) E[\|X - E[X|b^T X > 0]\|^2 | b^T X > 0] \\ &\quad + P(b^T X < 0) E[\|X - E[X|b^T X < 0]\|^2 | b^T X < 0] \\ &= E\left[\|X - E[X|b^T X > 0]\|^2 \middle| b^T X > 0\right], \qquad (\text{Since, } -X \overset{d}{=} X). \end{aligned}$$

The corresponding cluster centers are given by C_1 and C_2 where $C_1 = E[X|b^T X > 0]$ and $C_2 = E[X|b^T X < 0] = -C_1$, due to symmetry.

Step 1: Finding the projection of the $2-$means clustering centers onto the separating hyperplane.

As in the proof of Lemma A.1.4, we define f to be the pdf of $N(-\theta_1, D)$ and g to be the pdf of $N(\theta_1, D)$. We also define a latent variable $Q \sim \text{Ber}(0.5)$ and $Y \sim f$ if $Q = 0$ and $Y \sim g$ if $Q = 1$. Then $X \stackrel{d}{=} Y$ and similar to the proof of Lemma A.1.4,

$$C_1 = E[X|b^T X > 0] = \alpha \; E_f[Y|b^T Y > 0] + (1-\alpha) \; E_g[Y|b^T Y > 0],$$

where $\alpha = P\left(Q = 0|b^T Y > 0\right) = 1 - \Phi\left(ab_1/2\sqrt{b^T Db}\right)$ as shown in equation A.26, E_f is the expectation when the distribution of Y has a pdf f and E_g is the expectation when the pdf is g.

To find the projection of C_1 and C_2 onto the separating hyperplane $\mathcal{H}(b)$, we use lemma A.1.6. Since C_1 is the weighted mean of $E_f[Y|b^T Y > 0]$ and $E_g[Y|b^T Y > 0]$, the projection of C_1 is the weighted mean of their projections. Using Lemma A.1.6, we get that the i^{th} coordinate of the projection of C_1 onto the separating hyperplane $\mathcal{H}(b)$ is given by:

$$\begin{aligned}
\mathcal{P}(C_1)_i &= \alpha \left[-\frac{a}{2}\mathbb{I}\{i=1\} + \frac{ab_1 b_i}{2} + \frac{b_i\sigma_i^2 - b_i\left(b^T Db\right)}{b^T Db} \; \left(E_f\left[b^T Y \middle| b^T Y > 0\right] + \frac{ab_1}{2} \right) \right] \\
&\quad + (1-\alpha) \left[\frac{a}{2}\mathbb{I}\{i=1\} - \frac{ab_1 b_i}{2} + \frac{b_i\sigma_i^2 - b_i\left(b^T Db\right)}{b^T Db} \; \left(E_g\left[b^T Y \middle| b^T Y > 0\right] - \frac{ab_1}{2} \right) \right] \\
&= (1-2\alpha) \left[\frac{a}{2}\mathbb{I}\{i=1\} - \frac{ab_1 b_i}{2} - \frac{b_i\sigma_i^2 - b_i\left(b^T Db\right)}{b^T Db} \; \frac{ab_1}{2} \right] \\
&\quad + \frac{b_i\sigma_i^2 - b_i\left(b^T Db\right)}{b^T Db} \left(\alpha E_f\left[b^T Y \middle| b^T Y > 0\right] + (1-\alpha) E_g\left[b^T Y \middle| b^T Y > 0\right] \right) \\
&= (1-2\alpha) \left[\frac{a}{2}\mathbb{I}\{i=1\} - \frac{ab_1 b_i \sigma_i^2}{2b^T Db} \right] \\
&\quad + \frac{b_i\left(\sigma_i^2 - b^T Db\right)}{b^T Db} \left(\alpha E_f\left[b^T Y \middle| b^T Y > 0\right] + (1-\alpha) E_g\left[b^T Y \middle| b^T Y > 0\right] \right),
\end{aligned}$$

where E_f is the expectation when the distribution of Y has a pdf f and E_g is the expectation when the pdf is g.

Step 2: The projection of the optimum centers has to be the origin.

Since $\mathcal{H}(b)$ is a separating hyperplane for 2-means clustering, $b^T y > 0$ for some $y \in \mathbb{R}^d$ if and only if $\|y - C_1\| > \|y - C_2\|$. This gives that the line joining the centers C_1 and C_2 is perpendicular to the separating hyperplane and the separating hyperplane bisects the line segment joining the centers. Since the midpoint

of the centers $(C_1 + C_2)/2 = 0$, the projection of C_1 and C_2 onto the separating hyperplane is the origin. Therefore, $\mathcal{P}(C_1)_i = 0$ for every i. This implies that for the optimal separating hyperplane:

1. For $i = 1$,
$$(1-2\alpha)\left[\frac{a}{2} - \frac{ab_1^2\sigma_1^2}{2b^TDb}\right] + \frac{b_1\left(\sigma_1^2 - b^TDb\right)}{b^TDb}\mathbf{A} = 0, \tag{A.8}$$
where $\mathbf{A} = \alpha E_f\left[b^TY \middle| b^TY > 0\right] + (1-\alpha)E_g\left[b^TY \middle| b^TY > 0\right] > 0$.

2. For $i \geq 2$,
$$-(1-2\alpha)\left[\frac{ab_1b_i\sigma_i^2}{2b^TDb}\right] + \frac{b_i\left(\sigma_i^2 - b^TDb\right)}{b^TDb}\mathbf{A} = 0. \tag{A.9}$$

Step 3: Finding values of $b \in \mathbb{R}^d$ that satisfy the above equations.

Since we assume $b_1 \geq 0$, $\alpha = 1 - \Phi\left(ab_1/(2\sqrt{b^TDb})\right) \leq 0.5$, where equality occurs if and only if $b_1 = 0$. Hence $1 - 2\alpha \geq 0$. Now let us consider two cases. One when $\sigma_1^2 \geq \sigma_2^2$ and the other when $\sigma_2^2 > \sigma_1^2$.

1. **Case 1: $\sigma_1^2 \geq \sigma_2^2$**

   Note that $b^TDb \leq \max_i \sigma_i^2 = \sigma_1^2$ and therefore the second expression in equation (A.8) is greater than or equal to 0. Also $b_1^2\sigma_1^2 \leq b^TDb$, therefore the first expression in equation (A.8) is also greater than or equal to 0. Now for equation (A.8) to hold, we need both the expressions to be zero. Now the first expression is zero if either $b_1 = 0$ or $b_1^2\sigma_1^2 = b^TDb$. Note that $b_1^2\sigma_1^2 = b^TDb \iff b_1 = 1$ and $b^TDb = \sigma_1^2$. The second expression is also zero if either $b_1 = 0$ or $\{b_1 = 1$ and $b^TDb = \sigma_1^2\}$. So for equation (A.8) to hold, we need

   $$\{b_1 = 0\} \text{ OR } \{b_1 = 1 \text{ and } b^TDb = \sigma_1^2\}.$$

   If $b_1 = 1$ and $b^TDb = \sigma_1^2$, then $b_i = 0$ for all $i \geq 2$, so equation (A.9) holds for all $i \geq 2$. But if $b_1 = 0$, equation (A.9) simplifies to $b_i\left(\sigma_i^2 - b^TDb\right) = 0$ which holds if and only if $b_i = 0$ or $b^TDb = \sigma_i^2$. Therefore equations (A.8) and (A.9) hold iff

   $$\{b_1 = 1 \text{ and } b_i = 0,\ i \geq 2\} \text{ OR } \left\{b_1 = 0 \text{ and } \left(b_i = 0 \text{ or } b^TDb = \sigma_i^2,\ i \geq 2\right)\right\}.$$

2. **Case 2: $\sigma_1^2 < \sigma_2^2$**

   First, we consider equation (A.9) for $i = d$. $b^TDb \geq \min_i \sigma_i^2 = \sigma_d^2$, therefore the second expression in equation (A.9) has the same sign as b_d. The first expression in equation (A.9) also has the same sign as b_d. Therefore, for their sum to be zero, i.e., for equation (A.9) to hold for $i = d$, we need both the expressions to be zero. The first expression is zero if either $b_1 = 0$ or $b_d = 0$. The second expression is zero if either $b_i = 0$ or $b^TDb = \sigma_d^2$. So for $i = d$, equation (A.9) holds iff $\left\{b_1 = 0 \text{ and } \left(b_d = 0 \text{ or } b^TDb = \sigma_d^2,\ i \geq 2\right)\right\}$ OR $\{b_1 \neq 0 \text{ and } b_d = 0\}$.

If $b_1 = 0$ for $2 \leq i < d$, equation (A.9) simplifies to $b_i \left(\sigma_i^2 - b^T D b\right) = 0$ which holds if and only if $b_i = 0$ or $b^T D b = \sigma_i^2$. Now we concentrate on what happens when $b_1 \neq 0$, i.e. when $b_1 > 0$. We know so far that $b_d = 0$.

Now we consider equation (A.9) for $i = d-1$. Since $b_d = 0$, $b^T D b \geq \min_{1 \leq i \leq d-1} \sigma_i^2 = \sigma_{d-1}^2$, therefore the second expression in equation (A.9) has the same sign as b_{d-1}. The first expression in equation (A.9) also has the same sign as b_{d-1}. Therefore, for equation (A.9) to hold for $i = d-1$, we need both the expressions to be zero. The first expression is zero iff $b_{d-1} = 0$ since $b_1 > 0$. Similar arguments can be made by considering equation (A.9) for $i = d-2, \ldots, 3$ sequentially. We can show that if $b_1 > 0$, $b_i = 0$ for $i \geq 3$. Therefore the separating hyperplane is of the form

$$\mathcal{H}(b) = \{y \in \mathbb{R}^d : b^T y = 0\} \text{ where } b_1 > 0,\ b_1^2 + b_2^2 = 1,\ b_j = 0\ \forall\ j \geq 3.$$

We now consider equation (A.8) and equation (A.9) for $i = 2$. Note that $b_1 = 1$ and $b_2 = 0$ is a feasible solution.

Let us now consider $b_1 > 0$ and $0 < b_2^2 < 1$. In this case to study the feasibility of equation (A.8) and equation (A.9) for $i = 2$ we need to look at $\mathbf{A}$. Now recall that if $Y \sim f$, then $b^T Y \sim N(-ab_1/2, b^T D b)$ and if $Y \sim g$, then $b^T Y \sim N(ab_1/2, b^T D b)$. Therefore,

$$\begin{aligned}
\mathbf{A} &= \alpha E_f \left[b^T Y \middle| b^T Y > 0 \right] + (1-\alpha) E_g \left[b^T Y \middle| b^T Y > 0 \right] \\
&= \alpha \left(-\frac{ab_1}{2} + \sqrt{\frac{2b^T D b}{\pi}}\ \frac{e^{-\frac{a^2 b_1^2}{8 b^T D b}}}{2\Phi\left(-\frac{ab_1}{2\sqrt{b^T D b}}\right)} \right) + (1-\alpha) \left(\frac{ab_1}{2} + \sqrt{\frac{2b^T D b}{\pi}}\ \frac{e^{-\frac{a^2 b_1^2}{8 b^T D b}}}{2\Phi\left(\frac{ab_1}{2\sqrt{b^T D b}}\right)} \right).
\end{aligned}$$

Recall that $\alpha = 1 - \Phi\left(\frac{ab_1}{2\sqrt{b^T D b}}\right) = \Phi\left(-\frac{ab_1}{2\sqrt{b^T D b}}\right)$.

Therefore,

$$\begin{aligned}
\mathbf{A} &= \alpha \left(-\frac{ab_1}{2} + \sqrt{\frac{2b^T D b}{\pi}}\ \frac{e^{-\frac{a^2 b_1^2}{8 b^T D b}}}{2\alpha} \right) + (1-\alpha) \left(\frac{ab_1}{2} + \sqrt{\frac{2b^T D b}{\pi}}\ \frac{e^{-\frac{a^2 b_1^2}{8 b^T D b}}}{2(1-\alpha)} \right) \\
&= (1-2\alpha)\frac{ab_1}{2} + \sqrt{\frac{2b^T D b}{\pi}}\ e^{-\frac{a^2 b_1^2}{8 b^T D b}} \\
&= (1-2\alpha)\frac{ab_1}{2} + 2\sqrt{b^T D b}\ \phi\left(\frac{ab_1}{2\sqrt{b^T D b}}\right).
\end{aligned}$$

Plugging this into the L.H.S. of equation (A.8) we get:

$$
\begin{aligned}
&(1-2\alpha)\left[\frac{a}{2}-\frac{ab_1^2\sigma_1^2}{2b^TDb}\right]+\frac{b_1\left(\sigma_1^2-b^TDb\right)}{b^TDb}\left((1-2\alpha)\frac{ab_1}{2}+2\sqrt{b^TDb}\;\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\right)\\
&=(1-2\alpha)\frac{a}{2}\left[1+\frac{-b_1^2\sigma_1^2+b_1^2\left(\sigma_1^2-b^TDb\right)}{b^TDb}\right]+\frac{2b_1\left(\sigma_1^2(b_1^2+b_2^2)-(b_1^2\sigma_1^2+b_2^2\sigma_2^2)\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\\
&=(1-2\alpha)\frac{a}{2}\left[1-b_1^2\right]-\frac{2b_1b_2^2\left(\sigma_2^2-\sigma_1^2\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\\
&=(1-2\alpha)\frac{a}{2}\;b_2^2-\frac{2b_1b_2^2\left(\sigma_2^2-\sigma_1^2\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\\
&=b_2^2\left[(1-2\alpha)\frac{a}{2}-\frac{2b_1\left(\sigma_2^2-\sigma_1^2\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\right].
\end{aligned}
$$

Similarly simplifying L.H.S. of equation (A.9) for $i=2$ we get:

$$
\begin{aligned}
&-(1-2\alpha)\left[\frac{ab_1b_2\sigma_2^2}{2b^TDb}\right]+\frac{b_2\left(\sigma_2^2-b^TDb\right)}{b^TDb}\left((1-2\alpha)\frac{ab_1}{2}+2\sqrt{b^TDb}\;\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\right)\\
&=(1-2\alpha)\frac{ab_1b_2}{2}\left[\frac{-\sigma_2^2+\sigma_2^2-b^TDb}{b^TDb}\right]+\frac{2b_2\left(\sigma_2^2(b_1^2+b_2^2)-(b_1^2\sigma_1^2+b_2^2\sigma_2^2)\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\\
&=\;-(1-2\alpha)\frac{ab_1b_2}{2}+\frac{2b_2b_1^2\left(\sigma_2^2-\sigma_1^2\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\\
&=\;-b_1b_2\left[(1-2\alpha)\frac{a}{2}-\frac{2b_1\left(\sigma_2^2-\sigma_1^2\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)\right].
\end{aligned}
$$

Since $b_1>0$, $0<b_2^2<1$ and $\alpha=1-\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)$, equation (A.8) and equation (A.9) for $i=2$ hold if and only if

$$
\left(2\;\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)-1\right)\frac{a}{2}-\frac{2b_1\left(\sigma_2^2-\sigma_1^2\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)=0 \tag{A.10}
$$

Now we know that for $x>0$, $2\Phi(x)-1>2x\phi(x)$. So since $b_1>0$,

$$
\left(2\;\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)-1\right)>2\;\frac{ab_1}{2\sqrt{b^TDb}}\;\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right).
$$

Therefore, if $\sigma_2^2 \leq \sigma_1^2 + \frac{a^2}{4}$, the L.H.S. of equation (A.10) becomes

$$\begin{aligned}
&\left(2\ \Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} - \frac{2b_1\left(\sigma_2^2 - \sigma_1^2\right)}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) \\
&> 2\ \frac{a^2 b_1}{4\sqrt{b^T Db}}\ \phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 2\ \frac{a^2 b_1}{4\sqrt{b^T Db}}\ \phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) \\
&= 0.
\end{aligned}$$

Therefore if $\sigma_2^2 \leq \sigma_1^2 + \frac{a^2}{4}$, then equation (A.10) cannot hold and hence either $b_1 = 0$ or $b_2 = 0$. Putting everything together, this implies equations (A.8) and (A.9) hold iff

$$\{b_1 = 1 \ \text{ and } \ b_i = 0,\ i \geq 2\} \ \text{ OR } \ \left\{b_1 = 0 \ \text{ and } \ \left(b_i = 0 \ \text{ or } \ b^T Db = \sigma_i^2,\ i \geq 2\right)\right\}.$$

For $\sigma_2^2 > \sigma_1^2 + \frac{a^2}{4}$, we have shown that if $b_1 = 0$, we require $b_i = 0$ or $b^T Db = \sigma_i^2$ for all $i \geq 2$ and if $b_1 > 0$, we require $b_i = 0$ for all $i \geq 3$. Therefore equations (A.8) and (A.9) hold iff

$$\{b_1 = 1 \ \text{ and } \ b_i = 0,\ i \geq 2\} \ \text{ OR } \ \left\{b_1 = 0 \ \text{ and } \ \left(b_i = 0 \ \text{ or } \ b^T Db = \sigma_i^2,\ i \geq 2\right)\right\}$$

$$\text{OR } \ \left\{0 < b_1 < 1,\ b_1^2 + b_2^2 = 1,\ b_i = 0,\ i \geq 3, \ \text{ and Eq A.10 holds } \right\}.$$

From cases 1 and 2 we finally come to the conclusion that for equations (A.8) and (A.9) to hold for $\sigma_2^2 \leq \sigma_1^2 + \frac{a^2}{4}$, we need

$$\{b_1 = 1 \ \text{ and } \ b_i = 0,\ i \geq 2\} \ \text{ OR } \ \left\{b_1 = 0 \ \text{ and } \ \left(b_i = 0 \ \text{ or } \ b^T Db = \sigma_i^2,\ i \geq 2\right)\right\}. \tag{A.11}$$

For $\sigma_2^2 > \sigma_1^2 + \frac{a^2}{4}$, we need

$$\begin{aligned}
&\{b_1 = 1 \ \text{ and } \ b_i = 0,\ i \geq 2\} \ \text{ OR } \ \left\{b_1 = 0 \ \text{ and } \ \left(b_i = 0 \ \text{ or } \ b^T Db = \sigma_i^2,\ i \geq 2\right)\right\} \\
&\qquad \text{OR } \ \left\{0 < b_1 < 1,\ b_1^2 + b_2^2 = 1,\ b_i = 0,\ i \geq 3, \ \text{ and Eq A.10 holds } \right\}.
\end{aligned} \tag{A.12}$$

Step 4: Among the possible values of b, finding b^* that gives the minimum within sum of squares.

1. **Case 1: $\sigma_2^2 < \sigma_1^2 + \frac{a^2}{4}$**

   If every σ_i^2 is distinct, then for (A.11) to hold, for some unique $i = i_0$,

$$b^T Db = \sigma_{i_0}^2,\ b_{i_0} = 1 \ \text{ and } \ b_j = 0 \text{ for } j \neq i_0.$$

Notice that in this case, the optimal separating hyperplane is $\mathcal{H}(b) = \{y \in \mathbb{R}^d : y_{i_0} = 0\}$ for some i_0. Now if $b_1 = 1$ and $b_i = 0$ for $i \geq 2$, we use Lemma A.19 to find the corresponding within sum of squares as

$$W_1^* := W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \left(\sqrt{\frac{2}{\pi}}\, \sigma_1\, e^{-\frac{a^2}{8\sigma_1^2}} + \frac{a}{2}\, P\left(|Z| < \frac{a}{2\sigma_1} \right) \right)^2. \tag{A.13}$$

For $i_0 = 2$, that is, when $b_2 = 1$ and $b_i = 0$ for $i \neq 2$ we can similarly use Lemma A.19 to find the corresponding within sum of squares as

$$W_2^* := W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_2^2. \tag{A.14}$$

From Lemma A.1.5 we get that the corresponding within sum of squares when $b_{i_0} = 1$ and $b_j = 0$ for $j \neq i_0$, corresponding to any $i_0 \geq 2$ is

$$W_{i_0}^* := W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_{i_0}^2. \tag{A.15}$$

Now using the lower bound given by Lemma A.1.7 we see that,

$$\left[\frac{a}{2}\mathbb{P}(|Z| \leq u) + \sqrt{\frac{2}{\pi}}\sigma_1 \exp(-u^2/2) \right]^2 \geq \frac{2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4} \right) > \frac{2}{\pi}\sigma_2^2 > \frac{2}{\pi}\sigma_2^2, \quad \text{for } j \geq 3,$$

since $\sigma_2^2 > \sigma_j^2$ for $j \geq 3$. Therefore, the minimum within sum of squares is achieved if $W(b^*) = W_1^*$, that is, if $b_1^* = 1$ and $b_i^* = 0$ for every $i \neq 1$. Therefore, the unique optimal separating hyperplane is given by $\mathcal{H}(b^*) = \{y \in \mathbb{R}^d : y_1 = 0\}$.

If σ_i^2 are not distinct. Since, $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$, suppose for some $i_0 \geq 3$ and $i_0 + m \leq d$, $\sigma_{i_0}^2 = \sigma_{i_0+1}^2 = \ldots = \sigma_{i_0+m}^2 = \sigma^2$. Then the optimal separating hyperplane can be given by either $\mathcal{H}(b) = \{y \in \mathbb{R}^d : y_i = 0\}$, where $i \notin [i_0, i_0 + m]$, that is, $b_i = 1$ for some $i \notin [i_0, i_0 + m]$ and $b_j = 0$ for all $j \neq i$ or by $\mathcal{H}(b) = \{y \in \mathbb{R}^d : \sum_{j=i_0}^{i_0+m} b_j y_j = 0\}$, where $\sum_{j=i_0}^{i_0+m} b_j^2 = 1$.

Suppose if the separating hyperplane is $\mathcal{H}(b) = \{y \in \mathbb{R}^d : \sum_{j=i_0}^{i_0+m} b_j y_j = 0\}$, where $\sum_{j=i_0}^{i_0+m} b_j^2 = 1$, then the corresponding within sum of squares is given by

$$\widetilde{W}_{i_0,m}^* := W(b) = \sum_{j \notin [i_0, i_0+m]} \sigma_j^2 + \frac{a^2}{4} + \sum_{k=i_0}^{i_0+m} E\left[\left(X_k - E\left[X_k | b^T X > 0 \right] \right)^2 \middle| b^T X > 0 \right].$$

Since b_i for $i \in [i_0, i_0+m]$ are not all zero and $\sum_{i=i_0}^{i_0+m} b_i^2 = 1$, we can construct a orthogonal matrix $\widetilde{A}$ of dimension $(m+1)\times(m+1)$ whose first row is given by $(b_{i_0}, \ldots, b_{i_0+m})$. Now define a rotated space such that $v = (y_{i_0}, \ldots, y_{i_0+m})$ is transformed to $u = \widetilde{A}v$ and define $U = \widetilde{A}V$, where $V = (X_{i_0}, \ldots, X_{i_0+m})$. Then the separating hyperplane now becomes $\mathcal{H}(b) = \{u \in \mathbb{R}^d : u_{i_0} = 0\}$ since $u_{i_0} = \sum_{j=i_0}^{i_0+m} b_j y_j = 0$. As $V \sim N_{m+1}(0, \sigma^2 I)$, we have that $U \sim N_{m+1}(0, \sigma^2 I)$ and therefore, we can write the within sum of squares as:

$$\begin{aligned}
\widetilde{W}^*_{i_0,m} &= \sum_{j \notin [i_0, i_0+m]} \sigma_j^2 + \frac{a^2}{4} + \sum_{k=1}^{m} E\left[\left(U_k - E\left[U_k | U_1 > 0\right]\right)^2 \middle| U_1 > 0 \right] \\
&= \sum_{j \notin [i_0, i_0+m]} \sigma_j^2 + \frac{a^2}{4} + \sum_{k=i_0}^{i_0+m} E\left[\left(X_k - E\left[X_k | X_{i_0} > 0\right]\right)^2 \middle| X_{i_0} > 0 \right] \\
&= \sum_{j \neq i_0} \sigma_j^2 + \frac{a^2}{4} + E\left[\left(X_{i_0} - E\left[X_{i_0} | X_{i_0} > 0\right]\right)^2 \middle| X_{i_0} > 0 \right].
\end{aligned}$$

Similar to the calculations in Lemma A.1.5, we get

$$\widetilde{W}^*_{i_0,m} = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_{i_0}^2.$$

Now similar to the previous case, $\left(\sqrt{\frac{2}{\pi}}\, \sigma_1\, e^{-\frac{a^2}{8\sigma_1^2}} + \frac{a}{2}\, P\left(|Z| < \frac{a}{2\sigma_1}\right)\right)^2 \geq \frac{2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4}\right) > \frac{2}{\pi}\sigma_j^2$ for $j \geq 2$. Therefore, the minimum within sum of squares is achieved for b^* if $b_1^* = 1$ and $b_i^* = 0$ for every $i \neq 1$ and the optimal separating hyperplane is given by $\mathcal{H}(b) = \{y \in \mathbb{R}^d : y_1 = 0\}$.

2. **Case 2: $\sigma_2^2 > \sigma_1^2 + \frac{a^2}{4}$**

   Looking at the first possibility in Expression (A.12), that is, if $b_1 = 1$ and $b_i = 0$ for every $i \neq 1$, the corresponding minimum within sum of squares, as seen in the previous case in Equation (A.13), is given by W_1^*.

   Following similar reasonings as presented in Case 1 for the second possibility (A.12), we argue that if every σ_i^2 is distinct, then $b_1 = 0$ and for a unique $i_0 \geq 2$, $b^T D b = \sigma_{i_0}^2$, $b_{i_0} = 1$ and $b_j \to 0$ for $j \neq i_0$. As seen in the previous case, if $i_0 = 2$ the corresponding minimum within sum of squares is W_2^* (Equation (A.14)) and for $i_0 \geq 2$, it is $W_{i_0}^*$ (Equation (A.15)).

   To study the third possibility in A.12, we use Lemma A.1.4 to get the within sum of squares for any b such that $0 < b_1 < 1, b_1^2 + b_2^2 = 1$ and find the minimum possible $W(b)$ such that Equation (A.10)

holds.

$$W(b) = \sum_{i=1}^{d} \sigma_i^2 + \frac{a^2}{4} - \left[\left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 - \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right).$$

In order to minimize $W(b)$ notice that we have to maximize

$$\left[\left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 + \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right),$$

and for Equation (A.10) to hold, we have

$$\left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} = \frac{2b_1\left(\sigma_2^2 - \sigma_1^2\right)}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right).$$

Therefore, we have to maximize the following given b_1 satisfies Equation (A.10).

$$\begin{aligned}
&\left[\left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 + \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\\
&= \left[\frac{2b_1\left(\sigma_2^2 - \sigma_1^2\right)}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 + \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\\
&= \frac{4b_1^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) + \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\\
&= \frac{4\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right).
\end{aligned}$$

Again by using Equation (A.10) we get that

$$\frac{4\sigma_2^4}{b^T Db}\ \phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) = 4\sigma_2^4\ \left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)^2 \frac{a^2}{4}\ \frac{1}{4b_1^2(\sigma_2^2 - \sigma_1^2)^2}.$$

Since $2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1 \leq \frac{1}{\sqrt{2\pi}}\ \left(\frac{ab_1}{\sqrt{b^T Db}}\right)$, if b_1 satisfies Equation (A.10),

$$\begin{aligned}
\frac{4\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) &\leq 4\sigma_2^4\ \frac{1}{2\pi}\ \frac{a^2 b_1^2}{b^T Db}\ \frac{a^2}{4}\ \frac{1}{4b_1^2(\sigma_2^2 - \sigma_1^2)^2}\\
&= \frac{2}{\pi}\sigma_2^2\ \frac{\sigma_2^2}{b^T Db}\ \left(\frac{a^2/4}{\sigma_2^2 - \sigma_1^2}\right)^2\\
&\leq \frac{2}{\pi}\sigma_2^2\ \frac{\sigma_2^2}{\sigma_1^2}\ \left(\frac{a^2/4}{\sigma_2^2 - \sigma_1^2}\right)^2,
\end{aligned}$$

since $\sigma_1^2 < \sigma_2^2$ and therefore $\sigma_1^2 \leq b^T D b$. We now show that for large enough σ_2^2, $\frac{\sigma_2^2}{\sigma_1^2}\ \left(\frac{a^2/4}{\sigma_2^2-\sigma_1^2}\right)^2 \leq 1$, that is, we want to show $\sigma_1^2(\sigma_2^2-\sigma_1^2)^2 \geq \sigma_2^2\ \frac{a^4}{16}$. In order to show, we plug in x instead of σ_2^2 and consider the equation:

$$\sigma_1^2(x-\sigma_1^2)^2 - x\ \frac{a^4}{16} = 0 \iff \sigma_1^2 x^2 - \left(2\sigma_1^4 + \frac{a^4}{16}\right) x + \sigma_1^6 = 0 \tag{A.16}$$

Note that the larger solution to this equation is given by:

$$x = \frac{2\sigma_1^4 + \frac{a^4}{16} + \sqrt{\left(2\sigma_1^4 + \frac{a^4}{16}\right)^2 - 4\sigma_1^8}}{2\sigma_1^2} = \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4 + \frac{a^4}{64}}}{2\sigma_1^2}.$$

Therefore for all $x > \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4+\frac{a^4}{64}}}{2\sigma_1^2}$, $\sigma_1^2(x-\sigma_1^2)^2 - x\ \frac{a^4}{16} > 0$. Therefore, for

$$\sigma_2^2 > \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4 + \frac{a^4}{64}}}{2\sigma_1^2} \implies \frac{\sigma_2^2}{\sigma_1^2}\ \left(\frac{a^2/4}{\sigma_2^2-\sigma_1^2}\right)^2 < 1,$$

which gives

$$\frac{4\sigma_2^4}{b^T D b}\phi^2\left(\frac{a b_1}{2\sqrt{b^T D b}}\right) \leq \frac{2}{\pi}\sigma_2^2.$$

Therefore, when $\sigma_2^2 > \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4+\frac{a^4}{64}}}{2\sigma_1^2}$, for any b_1 satisfying Equation (A.10),

$$W(b) \geq \sum_{i=1}^{d}\sigma_i^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_2^2 = W_2^*.$$

If we additionally have that

$$\sigma_2^2 > \frac{\pi}{2}\left(\sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\ \sigma_1^2}} + \frac{a}{2}\ P\left(|Z| < \frac{a}{2\ \sigma_1}\right)\right)^2,$$

then

$$W_2 * < W_1^* < W_{i_0}^*, \quad \text{for} \quad i_0 \geq 3.$$

Therefore, considering all the three possibilities in Expression (A.12), if

$$\sigma_2^2 > \max\left\{\frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4+\frac{a^4}{64}}}{2\sigma_1^2}, \frac{\pi}{2}\left(\sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\ \sigma_1^2}} + \frac{a}{2}\ P\left(|Z| < \frac{a}{2\ \sigma_1}\right)\right)^2\right\},$$

then the minimum within sum of squares is given by W_2^* which is achieved for the unique b^* such that $b_2^* = 1$ and $b_j^* = 0$ if $j \neq 2$ and the unique optimal separating hyperplane is given by $\mathcal{H}(b) = \{y \in \mathbb{R}^d : y_2 = 0\}$.

□

A.2.3 Proof of Lemma 3.3.1

This proof is analogous to the proof of the matrix being positive definite for the null case as shown in Lemma A.0.1. We find the matrix G in the two different cases and show that it is positive definite.

1. When condition (3.7) is true, Theorem 3.3 gives us the unique optimum as $\mu_1^* = -\mu_2^* = (\mathbb{E}[X_1|X_1 > 0], 0, \ldots, 0)$, $M_{12} = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 = 0\}$, $\mathbb{P}(A_1) = \mathbb{P}(A_2) = 0.5$ and $r_{12} = 2\mathbb{E}[X_1|X_1 > 0]$, where

$$\mathbb{E}[X_1|X_1 > 0] = \sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\sigma_1^2}} + \frac{a}{2}\ \mathbb{P}\left(|Z| < \frac{a}{2\sigma_1}\right).$$

From the proof of lemma A.0.1, we know that for $x \in M_{12}$,

$$(x-\mu_1^*)(x-\mu_1^*)^T = \begin{pmatrix} \mu_{11}^{*2} & -\mu_{11}^* x_2 & \ldots & -\mu_{11}^* x_d \\ -\mu_{11}^* x_2 & x_2^2 & \ldots & x_2 x_d \\ \vdots & \vdots & \ddots & \vdots \\ -\mu_{11}^* x_d & x_2 x_d & \ldots & x_d^2 \end{pmatrix},$$

$$(x-\mu_1^*)(x-\mu_2^*)^T = \begin{pmatrix} -\mu_{11}^{*2} & -\mu_{11}^* x_2 & \ldots & -\mu_{11}^* x_d \\ -\mu_{21}^* x_2 & x_2^2 & \ldots & x_2 x_d \\ \vdots & \vdots & \ddots & \vdots \\ -\mu_{21}^* x_d & x_2 x_d & \ldots & x_d^2 \end{pmatrix}.$$

As before, $\mathbb{I}_{M_{12}} = \mathbb{I}_{\{X_1=0\}}$. Therefore,

$$\mu_{11}^{*2} \int_{M_{12}} f(x)\ d\sigma(x) = \mathbb{E}[X_1|X_1 > 0]^2\ \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1}.$$

For $2 \le j \le d$,

$$\begin{aligned} \mu_{11}^* \int_{M_{12}} x_j\ f(x)\ d\sigma(x) &= \mu_{11}^*\ \mathbb{E}[X_j]\ \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} = 0, \\ \mu_{21}^* \int_{M_{12}} x_j\ f(x)\ d\sigma(x) &= \mu_{21}^*\ \mathbb{E}[X_j]\ \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} = 0, \\ \int_{M_{12}} x_j^2\ f(x)\ d\sigma(x) &= \mathbb{E}[X_j^2]\ \ \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} = \frac{\sigma_j^2}{\sqrt{2\pi}\sigma_1}\ e^{-a^2/8\sigma_1^2}. \end{aligned}$$

Let $i \neq j$, and $i, j \in \{2, \ldots, d\}$,

$$\int_{M_{12}} x_i\ x_j\ f(x)\ d\sigma(x) = \mathbb{E}[X_i]\ \mathbb{E}[X_j]\ \ \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} = 0.$$

Then the matrix G can be derived as,

$$G_{22} = G_{11} = \mathbf{I}_d - \frac{1}{\mathbb{E}[X_1|X_1>0]} \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} \begin{pmatrix} \mathbb{E}[X_1|X_1>0]^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{pmatrix},$$

$$G_{21} = G_{12} = \frac{1}{\mathbb{E}[X_1|X_1>0]} \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} \begin{pmatrix} \mathbb{E}[X_1|X_1>0]^2 & 0 & \dots & 0 \\ 0 & -\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -\sigma_d^2 \end{pmatrix}.$$

Boyd and Vandenberghe (2004) now gives that the symmetric matrix G is positive definite if and only if G_{11} and G/G_{11} (the Schur complement of G_{11} in G) are both positive definite. Let us first look at the diagonal entries of G_{11} and define the following:

$$m_1 = \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} \mathbb{E}[X_1|X_1>0], \; m_j = \frac{\sigma_j^2}{\mathbb{E}[X_1|X_1>0]} \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1}, \; j \neq 1. \tag{A.17}$$

Then the diagonal entries of G_{11} are given by $1 - m_j$, for $j = 1, \dots, d$. Next note that

$$\frac{1}{\sqrt{2\pi}} \frac{a}{\sigma_1} e^{-a^2/8\sigma_1^2} \leq \mathbb{P}\left(|Z| < \frac{a}{2\sigma_1}\right) \leq \frac{1}{\sqrt{2\pi}} \frac{a}{\sigma_1}$$

and therefore we can get the following bounds on $\mathbb{E}[X_1|X_1>0]$.

$$\mathbb{E}[X_1|X_1>0] \leq \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_1} \left(\sigma_1^2 + \frac{a^2}{4}\right), \quad \mathbb{E}[X_1|X_1>0] \geq \sqrt{\frac{2}{\pi}} \frac{e^{-a^2/8\sigma_1^2}}{\sigma_1} \left(\sigma_1^2 + \frac{a^2}{4}\right) \tag{A.18}$$

Using these inequalities and observing that $e^x \geq 1 + x$ for $x > 0$ along with the assumption that $\sigma_1^2 + \frac{a^2}{4} > \sigma_j^2$ for $j \neq 1$, one can easily verify that $1 - m_j > 0$ for all $j = 1, \dots, d$. Therefore, G_{11} is a positive definite matrix. To show G/G_{11} is also positive definite first we simplify it.

$$G/G_{11} = G_{22} - G_{21} \left[G_{11}\right]^{-1} G_{12}.$$

Since all of them are diagonal matrices, G/G_{11} is also a diagonal matrix with the j^{th} entry given by $1 - m_j - \frac{m_j^2}{1-m_j} = \frac{1-2m_j}{1-m_j}$. Since, we have already verified that $1 - m_j > 0$, we just have to verify that $1 - 2m_j$ is also greater than 0. This can again be easily verified with the properties as mentioned above. That is, by using the inequalities and the assumption. Therefore, G/G_{11} is also a positive definite matrix, which implies G itself is a positive definite matrix.

2. When condition (3.9) is true, Theorem 3.3 gives us the unique optimum as $\mu_1^* = -\mu_2^* = (0, \mathbb{E}[X_2|X_2 > 0], 0, \ldots, 0)$, $M_{12} = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_2 = 0\}$, $\mathbb{P}(A_1) = \mathbb{P}(A_2) = 0.5$ and $r_{12} = 2\mathbb{E}[X_2|X_2 > 0]$, where

$$\mathbb{E}[X_2|X_2 > 0] = \sqrt{\frac{2}{\pi}}\ \sigma_2.$$

Analogous to previous part we can show that for $x \in M_{12}$,

$$(x-\mu_1^*)(x-\mu_1^*)^T = \begin{pmatrix} x_1^2 & -\mu_{11}^* x_1 & x_1x_3 & \ldots & -\mu_{11}^* x_d \\ -\mu_{11}^* x_1 & \mu_{11}^{*2} & -\mu_{11}^* x_3 & \ldots & -\mu_{11}^* x_d \\ x_1x_3 & -\mu_{11}^* x_3 & x_3^2 & \ldots & x_3x_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1x_d & -\mu_{11}^* x_d & x_3x_d & \ldots & x_d^2 \end{pmatrix},$$

$$(x-\mu_1^*)(x-\mu_2^*)^T = \begin{pmatrix} x_1^2 & -\mu_{21}^* x_1 & x_1x_3 & \ldots & -\mu_{11}^* x_d \\ -\mu_{11}^* x_1 & -\mu_{11}^{*2} & -\mu_{11}^* x_3 & \ldots & -\mu_{11}^* x_d \\ x_1x_3 & -\mu_{21}^* x_3 & x_3^2 & \ldots & x_3x_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1x_d & -\mu_{21}^* x_d & x_3x_d & \ldots & x_d^2 \end{pmatrix}.$$

Now, $\mathbb{I}_{M_{12}} = \mathbb{I}_{\{X_2=0\}}$. Therefore,

$$\mu_{11}^{*2} \int_{M_{12}} f(x)\ d\sigma(x) = \frac{2\sigma_2^2}{\pi} \frac{1}{\sqrt{2\pi}\sigma_2} = \sqrt{\frac{2}{\pi^3}}\sigma_2.$$

For $j \neq 2$,

$$\begin{aligned} \mu_{11}^* \int_{M_{12}} x_j\ f(x)\ d\sigma(x) &= \mu_{11}^*\ \mathbb{E}[X_j]\ \frac{1}{\sqrt{2\pi}\sigma_2} = 0, \\ \mu_{21}^* \int_{M_{12}} x_j\ f(x)\ d\sigma(x) &= \mu_{21}^*\ \mathbb{E}[X_j]\ \frac{1}{\sqrt{2\pi}\sigma_2} = 0, \\ \int_{M_{12}} x_j^2\ f(x)\ d\sigma(x) &= \mathbb{E}[X_j^2]\ \ \frac{1}{\sqrt{2\pi}\sigma_2}. \end{aligned}$$

Therefore for $j \geq 3$,

$$\int_{M_{12}} x_j^2\ f(x)\ d\sigma(x) = \frac{\sigma_j^2}{\sqrt{2\pi}\sigma_2},$$

and for $j = 1$,

$$\int_{M_{12}} x_j^2\ f(x)\ d\sigma(x) = \frac{\sigma_1^2 + \frac{a^2}{4}}{\sqrt{2\pi}\sigma_2}.$$

Let $i \neq j$, and $i, j \in \{1, 3, \ldots, d\}$,

$$\int_{M_{12}} x_i \; x_j \; f(x) \; d\sigma(x) = \mathbb{E}[X_i] \; \mathbb{E}[X_j] \;\; \frac{1}{\sqrt{2\pi}\sigma_2} = 0.$$

Then the matrix G can be derived as,

$$G_{22} = G_{11} = \mathbf{I}_d - \frac{1}{2\sigma_2^2} \begin{pmatrix} \sigma_1^2 + \frac{a^2}{4} & 0 & 0 & \ldots & 0 \\ 0 & \frac{2\sigma_2^2}{\pi} & 0 & \ldots & 0 \\ 0 & 0 & \sigma_3^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sigma_d^2 \end{pmatrix},$$

$$G_{21} = G_{12} = \frac{1}{2\sigma_2^2} \begin{pmatrix} -\sigma_1^2 - \frac{a^2}{4} & 0 & 0 & \ldots & 0 \\ 0 & \frac{2\sigma_2^2}{\pi} & 0 & \ldots & 0 \\ 0 & 0 & -\sigma_3^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -\sigma_d^2 \end{pmatrix}.$$

Boyd and Vandenberghe (2004) now gives that the symmetric matrix G is positive definite if and only if G_{11} and G/G_{11} (the Schur complement of G_{11} in G) are both positive definite. Since $\sigma_2^2 > \sigma_1^2 + \frac{a^2}{4}$ under the assumption (3.9) and $\sigma_2^2 > \sigma_j^2$ for every $j \geq 3$, all the diagonal elements of G_{11} are strictly positive and therefore G_{11} is a positive definite matrix. Now as observed in the previous part, G/G_{11} is given by

$$G/G_{11} = G_{22} - G_{21} \left[G_{11}\right]^{-1} G_{12},$$

which ends up being a diagonal matrix with the j^{th} entry given by $\frac{1-2m_j}{1-m_j}$, where in this part,

$$m_1 = \frac{\sigma_1^2 + \frac{a^2}{4}}{2\sigma_2^2}, \;\; m_2 = \frac{1}{\pi}, \;\; m_j = \frac{\sigma_j^2}{2\sigma_2^2}, \;\; j \geq 3.$$

$1 - m_j$ are the diagonal entries of G_{11} which we have verified are greater than zero. Again since $\sigma_2^2 > \sigma_1^2 + \frac{a^2}{4}$ and $\sigma_2^2 > \sigma_j^2$ for every $j \geq 3$, $1 - 2m_j$ is greater than 0 for every j and therefore, G_0/G_{11} is also a positive definite matrix, which implies G itself is a positive definite matrix.

□

A.2.4 Proofs of Additional Lemmas Supporting Theorem 3.3

Proof of Lemma A.1.3

Let Z be a random variable generated from $N(0,1)$. Then,

$$\begin{aligned}
\mathbb{E}\left[Y|Y>0\right] &= 2\mathbb{E}[Y\mathbb{I}_{\{Y>0\}}] \\
&= \int_0^\infty x f(x)dx + \int_0^\infty x g(x)dx \\
&= \frac{1}{\sqrt{2\pi}\sigma_1}\left[\int_0^\infty x e^{-\frac{1}{2\sigma_1^2}(x+a/2)^2}dx + \int_0^\infty x e^{-\frac{1}{2\sigma_1^2}(x-a/2)^2}dx\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_1}\left[\int_0^\infty \left(x+\frac{a}{2}\right) e^{-\frac{1}{2\sigma_1^2}(x+a/2)^2}dx + \int_0^\infty \left(x-\frac{a}{2}\right) e^{-\frac{1}{2\sigma_1^2}(x-a/2)^2}dx\right] + \frac{a}{2}\mathbb{P}\left(|Z|<\frac{a}{2\sigma_1}\right) \\
&= \frac{\sigma_1}{\sqrt{2\pi}}\left[e^{-\frac{1}{2\sigma_1^2}(x+a/2)^2}\Big|_\infty^0 + e^{-\frac{1}{2\sigma_1^2}(x-a/2)^2}\Big|_\infty^0\right] + \frac{a}{2}\mathbb{P}\left(|Z|<\frac{a}{2\sigma_1}\right) \\
&= \sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\sigma_1^2}} + \frac{a}{2}\ \mathbb{P}\left(|Z|<\frac{a}{2\sigma_1}\right).
\end{aligned}$$

□

Proof of Lemma A.1.4

The within sum of squares can be written as:

$$\begin{aligned}
W(b) &= P(b^TX>0)E[||X-E[X|b^TX>0]||^2|b^TX>0] \\
&\quad + P(b^TX<0)E[||X-E[X|b^TX<0]||^2|b^TX<0] \\
&= E\left[\,||X-E[X|b^TX>0]||^2\big|\,b^TX>0\right], \qquad (\text{Since, } -X\overset{d}{=}X).
\end{aligned}$$

Since $b_j = 0\ \forall\ j\geq 3$, and X_j for $j\geq 3$ is independent of X_1 and X_2,

$$\begin{aligned}
W(b) &= E\left[\,||X-E[X|b^TX>0]||^2\big|\,b^TX>0\right] \\
&= E\left[\,(X_1-E[X_1|b^TX>0])^2\big|\,b^TX>0\right] + E\left[\,(X_2-E[X_2|b^TX>0])^2\big|\,b^TX>0\right] + \sum_{j=3}^d \sigma_j^2,
\end{aligned}$$

since for $j\geq 3$, $E[X_j|b^TX>0] = E[X_j] = 0$ and $E[X_j^2|b^TX>0] = E[X_j^2] = \sigma_j^2$. Therefore,

$$W(b) = E\left[\,(X_1-E[X_1|b^TX>0])^2\big|\,b^TX>0\right] + E\left[\,(X_2-E[X_2|b^TX>0])^2\big|\,b^TX>0\right] + \sum_{j=3}^d \sigma_j^2. \quad (A.19)$$

We know that,

$$E\left[(X_1 - E[X_1|b^TX > 0])^2 \middle| b^TX > 0\right] = E\left[X_1^2 \middle| b^TX > 0\right] - \left(E[X_1|b^TX > 0]\right)^2,$$

and similarly,

$$E\left[(X_2 - E[X_2|b^TX > 0])^2 \middle| b^TX > 0\right] = E\left[X_2^2 \middle| b^TX > 0\right] - \left(E[X_2|b^TX > 0]\right)^2.$$

Since $-X \stackrel{d}{=} X$, we can write,

$$\begin{aligned} E\left[X_1^2\right] &= P\left(b^TX > 0\right) E\left[X_1^2 \middle| b^TX > 0\right] + P\left(b^TX < 0\right) E\left[X_1^2 \middle| b^TX < 0\right] \\ &= P\left(b^TX > 0\right) E\left[X_1^2 \middle| b^TX > 0\right] + P\left(b^TX < 0\right) E\left[X_1^2 \middle| b^TX > 0\right] \\ &= E\left[X_1^2 \middle| b^TX > 0\right]. \end{aligned}$$

Hence,

$$E\left[X_1^2 \middle| b^TX > 0\right] = E\left[X_1^2\right] = \sigma_1^2 + \frac{a^2}{4}, \tag{A.20}$$

and following similar arguments,

$$E\left[X_2^2 \middle| b^TX > 0\right] = E\left[X_2^2\right] = \sigma_2^2. \tag{A.21}$$

To find the conditional first moments, for simplicity of notation, let us define f to be the pdf of $N(-\theta_1, D)$ and g to be the pdf of $N(\theta_1, D)$. Let us define a latent variable $Q \sim \text{Ber}(0.5)$ and define $Y \sim f$ if $Q = 0$ and $Y \sim g$ if $Q = 1$. Then $X \stackrel{d}{=} Y$ and by the law of total expectation,

$$\begin{aligned} E\left[X_1 \middle| b^TX > 0\right] &= E\left[Y_1 \middle| b^TY > 0\right] \\ &= E\left[E[Y_1|b^TY > 0, Q] \middle| b^TY > 0\right] \\ &= P\left(Q = 0|b^TY > 0\right)\ E_f[Y_1|b^TY > 0, Q = 0] \\ &\quad + P\left(Q = 1|b^TY > 0\right)\ E_g[Y_1|b^TY > 0, Q = 1] \\ &= \alpha\ E_f[Y_1|b^TY > 0] + (1 - \alpha)\ E_g[Y_1|b^TY > 0], \end{aligned}$$

where $\alpha = P\left(Q = 0|b^TY > 0\right)$, E_f is the expectation when the distribution of Y has a pdf f and E_g is the expectation when the pdf is g.

Similarly,

$$E\left[X_2|\, b^TX>0\right] = \alpha\ E_f[Y_2|b^TY>0] + (1-\alpha)\ E_g[Y_2|b^TY>0]$$

To simplify further, we consider each of these terms separately. We first start with $E_f[Y_1|b^TY>0]$ and to compute it we define random variable V such that when $Y\sim f$, that is, $Y\sim N(-\theta_1, D)$,

$$V=\begin{pmatrix} V_1\\ V_2\end{pmatrix} := \begin{pmatrix} \frac{Y_1+a/2}{\sigma_1}\\ \frac{b_1Y_1+b_2Y_2+ab_1/2}{\sqrt{b^TDb}}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} 1 & \frac{b_1\sigma_1}{\sqrt{b^TDb}}\\ \frac{b_1\sigma_1}{\sqrt{b^TDb}} & 1\end{pmatrix}\right).$$

Note that $b^TY = b_1Y_1+b_2Y_2>0$ is equivalent to $V_2>\frac{ab_1}{2\sqrt{b^TDb}}$. In order to find $E\left[V_1\left|V_2>\frac{ab_1}{2\sqrt{b^TDb}}\right.\right]$ we use moments derived for truncated bivariate normal distribution, presented in Rosenbaum (1961). We get

$$\begin{aligned} E\left[V_1\left|V_2>\frac{ab_1}{2\sqrt{b^TDb}}\right.\right] &= \frac{b_1\sigma_1}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{P\left(V_2>\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right)\\ &= \frac{b_1\sigma_1}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{1-\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right).\end{aligned}$$

Note that

$$E\left[V_1\left|V_2>\frac{ab_1}{2\sqrt{b^TDb}}\right.\right] = E_f\left[\frac{Y_1+a/2}{\sigma_1}\middle|\, b^TY>0\right].$$

Therefore,

$$E_f\left[Y_1|\, b^TY>0\right] = -\frac{a}{2}+\frac{b_1\sigma_1^2}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{1-\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right). \tag{A.22}$$

Similarly in order to find $E_f[Y_2|b^TY>0]$, we define random variable V^* such that when $Y\sim f$, that is, $Y\sim N(-\theta_1, D)$,

$$V^*=\begin{pmatrix} V_1^*\\ V_2\end{pmatrix} := \begin{pmatrix} \frac{Y_2}{\sigma_2}\\ \frac{b_1Y_1+b_2Y_2+ab_1/2}{\sqrt{b^TDb}}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} 1 & \frac{b_2\sigma_2}{\sqrt{b^TDb}}\\ \frac{b_2\sigma_2}{\sqrt{b^TDb}} & 1\end{pmatrix}\right).$$

Then again using moments derived for truncated bivariate normal distribution, presented in Rosenbaum (1961). We get

$$E\left[V_1^*\left|V_2>\frac{ab_1}{2\sqrt{b^TDb}}\right.\right] = \frac{b_2\sigma_2}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{1-\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right).$$

Again note that,

$$E\left[V_1^* \middle| V_2 > \frac{ab_1}{2\sqrt{b^T Db}}\right] = E_f\left[\frac{Y_2}{\sigma_2}\middle| b^T Y > 0\right].$$

Therefore,

$$E_f\left[Y_2|\, b^T Y > 0\right] = \frac{b_2\sigma_2^2}{\sqrt{b^T Db}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)}{1-\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)}\right). \tag{A.23}$$

In order to find $E_g\left[Y_1|\, b^T Y > 0\right]$ and $E_g\left[Y_2|\, b^T Y > 0\right]$, note that f is the density of $N(-\theta_1, D)$ and g is the density of $N(\theta_1, D)$, where $\theta = (a/2, 0, \ldots, 0)$. So just replacing a with $-a$ in equations (A.22) and (A.23) will give us the corresponding expectations when $Y \sim g$. So we get

$$E_g\left[Y_1|\, b^T Y > 0\right] = \frac{a}{2} + \frac{b_1\sigma_1^2}{\sqrt{b^T Db}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)}{\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)}\right) \tag{A.24}$$

$$E_g\left[Y_2|\, b^T Y > 0\right] = \frac{b_2\sigma_2^2}{\sqrt{b^T Db}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)}{\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)}\right) \tag{A.25}$$

To find α we note that,

$$\begin{aligned}\alpha = P\left(Q=0|b^T Y>0\right) &= \frac{P\left(Q=0, b^T Y>0\right)}{P\left(Q=0, b^T Y>0\right)+P\left(Q=1, b^T Y>0\right)} \\ &= \frac{P_f\left(b^T Y>0\right)}{P_f\left(b^T Y>0\right)+P_g\left(b^T Y>0\right)}.\end{aligned}$$

Now if $Y \sim g$, then $-Y \sim f$. So $P_g\left(b^T Y>0\right) = P_f\left(b^T Y<0\right) = 1 - P_f\left(b^T Y>0\right)$. Now if $Y \sim f$, then $b^T Y \sim N(-ab_1/2, b^T Db)$. Therefore,

$$\alpha = P_f\left(b^T Y>0\right) = P_f\left(\frac{b^T Y + ab_1/2}{\sqrt{b^T Db}} > \frac{ab_1/2}{\sqrt{b^T Db}}\right) = 1-\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right). \tag{A.26}$$

Using all of the above equations we get:

$$\begin{aligned}E\left[X_1|\, b^T X>0\right] &= \alpha E_f\left[Y_1|\, b^T Y>0\right] + (1-\alpha)E_g\left[Y_1|\, b^T Y>0\right] \\ &= \left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)-1\right)\frac{a}{2} + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right).\end{aligned}$$

Similarly,

$$\begin{aligned}E\left[X_2|\, b^T X>0\right] &= \alpha E_f\left[Y_2|\, b^T Y>0\right] + (1-\alpha)E_g\left[Y_2|\, b^T Y>0\right] \\ &= \frac{2b_2\sigma_2^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right).\end{aligned}$$

Plugging the expressions for $E\left[X_1 | b^T X > 0\right]$ and $E\left[X_2 | b^T X > 0\right]$ along with equations (A.20) and (A.21) into equation (A.19), we get that:

$$W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \left[\left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 - \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right).$$

□

Proof of Lemma A.1.5

The within sum of squares can be written as:

$$\begin{aligned}
W(b) &= P(b^T X > 0)E[||X - E[X|b^T X > 0]||^2 | b^T X > 0] \\
&\quad + P(b^T X < 0)E[||X - E[X|b^T X < 0]||^2 | b^T X < 0] \\
&= E\left[\,||X - E[X|b^T X > 0]||^2 \big|\, b^T X > 0\right], \qquad \text{(Since, } -X \stackrel{d}{=} X\text{)} \\
&= E\left[\,||X - E[X|X_i > 0]||^2 \big|\, b^T X_i > 0\right] \\
&= E\left[(X_i - E[X_i|X_i > 0])^2 \,|X_i > 0\right] + \sum_{j \neq i} E[X_j^2].
\end{aligned}$$

Recall that $X_i \sim N(0, \sigma_i^2)$, therefore $E[X_i|X_i > 0] = \sqrt{\frac{2}{\pi}}\sigma_i$. This implies that the within sum of squares is given by,

$$\begin{aligned}
W(b) &= Var(X_i|X_i > 0) + E[X_1^2] + \sum_{j \neq i, j \neq 1} E[X_j^2] \\
&= \sigma_i^2 - \frac{2}{\pi}\sigma_i^2 + \sigma_1^2 + \frac{a^2}{4} + \sum_{j \neq i, j \neq 1} \sigma_j^2 \\
&= \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_i^2.
\end{aligned}$$

□

Proof of Lemma A.1.6

First, for any point $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$, its projection, $u = (u_1, \ldots, u_d)$, onto the hyperplane $\mathcal{H}(b) = \{y \in \mathbb{R}^d : b^T y = 0\}$ is given by the following equations:

$$u_i = v_i + b_i t, \text{ for some } t, \ \forall\ i \text{ and } b^T u = 0,$$

solving which gives us that for every i,

$$u_i = v_i - b_i \sum_{i=1}^{d} b_i v_i = v_i - b_i(b^T v).$$

Now for simplicity, define $Z_{1i} = Y_i - b_i(b^T Y)$, the projection of Y onto the plane and $Z_2 = b^T Y$. Therefore, the i^{th} coordinate of the projection of $E[Y|b^T Y > 0]$ is given by:

$$\begin{aligned}
\mathcal{P}_i &= E\left[Y_i | Z_2 > 0\right] - b_i \left(b^T E\left[Y | Z_2 > 0\right]\right) \\
&= E\left[Y_i - b_i(b^T Y) | Z_2 > 0\right] \\
&= E\left[Z_{1i} | Z_2 > 0\right] \\
&= \frac{E\left[Z_{1i} I\{Z_2 > 0\}\right]}{P(Z_2 > 0)} \\
&= \frac{E\left[E\left[Z_{1i} I\{Z_2 > 0\} | Z_2\right]\right]}{P(Z_2 > 0)} \\
&= \frac{E\left[I\{Z_2 > 0\} E\left[Z_{1i} | Z_2\right]\right]}{P(Z_2 > 0)} \\
&= E\left[E\left[Z_{1i} | Z_2\right] | \, Z_2 > 0\right] \\
&= E\left[E[Z_{1i}] + \frac{Cov(Z_{1i}, Z_2)}{Var(Z_2)} \left(Z_2 - E[Z_2]\right) \middle| \, Z_2 > 0\right] \\
&= E[Z_{1i}] + \frac{Cov(Z_{1i}, Z_2)}{Var(Z_2)} \left(E\left[\, Z_2 | \, Z_2 > 0\right] - E[Z_2]\right).
\end{aligned}$$

Note that we can do this because Y is a multivariate normal random variable and therefore Z_{1i} and Z_2 are jointly normal. Now since $Y \sim N(\theta_1, D)$, where $\theta_1 = (a/2, 0, \ldots, 0)$, and D is a diagonal matrix,

$$E[Z_2] = E\left[b^T Y\right] = b^T E\left[Y\right] = \frac{a b_1}{2},$$

$$E[Z_{1i}] = E\left[Y_i - b_i(b^T Y)\right] = E\left[Y_i - b_i Z_2\right] = E\left[Y_i\right] - b_i E\left[Z_2\right] = \frac{a}{2}\mathbb{I}\{i = 1\} - \frac{a b_1 b_i}{2},$$

$$Var(Z_2) = Var\left(b^T Y\right) = b^T D \; b,$$

and

$$Cov(Z_{1i}, Z_2) = Cov\left(Y_i - b_i Z_2, Z_2\right) = Cov\left(Y_i, Z_2\right) - b_i Var\left(Z_2\right) = b_i Var(Y_i) - b_i \left(b^T D \; b\right).$$

Therefore the projection is given by:

$$\mathcal{P}_i = \frac{a}{2}\mathbb{I}\{i = 1\} - \frac{a b_1 b_i}{2} + \frac{b_i Var(Y_i) - b_i \left(b^T D b\right)}{b^T D b} \; \left(E\left[\, Z_2 | \, Z_2 > 0\right] - \frac{a b_1}{2}\right).$$

□

Proof of Lemma A.1.7

Our goal is to lower bound the term:

$$T := \left[\frac{a}{2}\mathbb{P}(|Z| \leq u) + \sqrt{\frac{2}{\pi}}\sigma_1 \exp(-u^2/2)\right]^2,$$

where $u = a/(2\sigma_1)$. Defining,

$$R := \left[\frac{a}{2}\mathbb{P}(|Z| \leq u) + \sqrt{\frac{2}{\pi}}\sigma_1 \exp(-u^2/2)\right],$$

we see that if we can lower bound R with k, i.e., find k such that $R \geq k > 0$, then k^2 is a lower bound on T^2. We consider two cases:

Case when $0 \leq u \leq 2$: We first claim that the following hold for all $u \geq 0$:

$$\begin{aligned}
\exp(-u^2/2) &\geq 1 - \frac{u^2}{2} + \frac{u^4}{8} - \frac{u^6}{48} + \frac{u^8}{384} - \frac{u^{10}}{3840} \\
\mathbb{P}(|Z| \leq u) &\geq \sqrt{\frac{2}{\pi}}\left[u - \frac{u^3}{6} + \frac{u^5}{40} - \frac{u^7}{336} + \frac{u^9}{3456} - \frac{u^{11}}{42240}\right],
\end{aligned}$$

and both lower bounds are positive for $0 \leq u \leq 2$. The first bound follows from the Taylor expansion of $\exp(x)$. For the second we use that:

$$\begin{aligned}
\mathbb{P}(|Z| \leq u) &= \sqrt{\frac{2}{\pi}}\int_0^u \exp(-x^2/2)dx \\
&\geq \sqrt{\frac{2}{\pi}}\int_0^u \left(1 - \frac{x^2}{2} + \frac{x^4}{8} - \frac{x^6}{48} + \frac{x^8}{384} - \frac{x^{10}}{3840}\right)dx.
\end{aligned}$$

The fact that the bounds are positive for the specified range can be directly verified.

Now, we can simply plug-in these estimates to obtain the following:

$$\begin{aligned}
\frac{-2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4}\right) + T &\geq \frac{-2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4}\right) + \frac{a^2}{2\pi}\left[\frac{1}{u} + \frac{u}{2} - \frac{u^3}{24} + \frac{u^5}{240} - \frac{u^7}{2688} + \frac{u^9}{34560} - \frac{u^{11}}{42240}\right]^2, \\
&= \frac{a^2u^2}{2\pi}\left[\frac{1}{6} - \frac{u^2}{30} + \frac{13u^4}{2520} - \frac{u^6}{1512} + \frac{797u^8}{26611200} - \frac{233u^{10}}{7983360} + \frac{42067u^{12}}{17882726400}\right. \\
&\quad \left. - \frac{559u^{14}}{2554675200} + \frac{1697u^{16}}{91968307200} - \frac{u^{18}}{729907200} + \frac{u^{20}}{1784217600}\right].
\end{aligned}$$

and using the fact that $u \leq 2$ to bound the negative terms (and dropping some positive terms) we obtain that,

$$\begin{aligned}\frac{-2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4}\right) + T &\geq \frac{a^2u^2}{2\pi}\left[\frac{1}{6} - \frac{u^2}{30} + \frac{19u^4}{7560} - \frac{6929u^8}{79833600}\right] \\ &\geq \frac{a^2u^2}{2\pi}\left[\frac{1}{6} - \frac{u^2}{30}\right] \geq \frac{a^2u^2}{60\pi},\end{aligned}$$

as desired.

Case when $u \geq 2$: Observe that if $u^2 \geq 4$, then:

$$\frac{-2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4}\right) + T \geq \frac{a^2}{40}.$$

Notice that since $u^2 \geq 4$, we obtain that,

$$\frac{2}{\pi}\left(\sigma_1^2 + \frac{a^2}{4}\right) \leq \frac{5a^2}{8\pi}.$$

We also notice that we can verify numerically that,

$$T \geq \frac{a^2}{4}(\mathbb{P}(|Z| \leq 2))^2 \geq \frac{9a^2}{40}.$$

Putting these two bounds together yields the desired result.

□

A.3 Proof of Theorem 3.4

Throughout this proof we use $c, C, c_1, C_1, \ldots$ to denote positive constants whose value may change from line to line. Recall, that in studying the power of SigClust we suppose that, we observe samples:

$$\{X_1, \ldots, X_n\} \sim \frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D) \tag{A.27}$$

where $\theta_1 = (a/2, 0, \ldots, 0) \in \mathbb{R}^d$ and $a > 0$. Furthermore, D is a diagonal matrix with elements $\Sigma_{jj} = \sigma_j^2$, such that $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$. Recall that our goal is to show that when condition (3.7),

$$\sigma_2^2 < \sigma_1^2 + \frac{a^2}{4}$$

holds, SigClust is asymptotically consistent, and when condition (3.9),

$$\sigma_2^2 > \max\left\{ \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4 + \frac{a^4}{64}}}{2\sigma_1^2}, \frac{\pi}{2}\ \kappa^2 \right\}$$

holds, SigClust is asymptotically inconsistent. Before we embark on the proof of the theorem we first recollect that Theorem 3.3 gave a characterization of the population-level optimal symmetric 2-means solution in this model. Under the model in (A.27) described above, the population-level optimal 2-means solution is unique and is given by (A.4) and (A.5).

Let us now first derive the power of the test in terms of the limiting distribution of the statistic under the null and alternate. We let $\mathbb{P}_0$ denote the Gaussian distribution with mean 0, diagonal covariance matrix:

$$D_0 = \begin{bmatrix} \sigma_1^2 + \frac{a^2}{4} & 0 & 0 \dots & 0 & \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & \sigma_d^2 \end{bmatrix}.$$

and use $\mathbb{P}_1$ to denote the distribution in (A.27). We let $W_0(\boldsymbol{\mu}_0)$ denote the population optimal 2-means value under $\mathbb{P}_0$, and let

$$\tau_0^2 = \sum_{i=1}^{2} \mathbb{P}_0(A_i)\mathbb{E}_{X\sim\mathbb{P}_0}[\|X - \mathbb{E}[X|X \in A_i]\|^4 | X \in A_i] - [W_0(\boldsymbol{\mu}_0)]^2.$$

Similarly, we let $W_1(\boldsymbol{\mu}_1)$ be the population optimal 2-means value under $\mathbb{P}_1$, and let

$$\tau_1^2 = \sum_{i=1}^{2} \mathbb{P}_1(A_i)\mathbb{E}_{X\sim\mathbb{P}_1}[\|X - \mathbb{E}[X|X \in A_i]\|^4 | X \in A_i] - [W_1(\boldsymbol{\mu}_1)]^2.$$

With this notation in place the following result characterizes the power of SigClust. We let Φ denote the standard normal CDF.

Lemma A.1.8. *SigClust has power:*

$$Power_n(a) = \Phi\left(\frac{\tau_0 \Phi^{-1}(\alpha)}{\tau_1} + \sqrt{n}\frac{W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)}{\tau_1} \right).$$

We prove this result in Appendix A.3.1, but note that it follows from straightforward calculations based on Lemma 3.0.1 and Theorem 3.2. As a consequence of this result, we have the following characterization of SigClust:

Lemma A.1.9. *Suppose that for some constant $C > 0$,*

$$\frac{\tau_0}{\tau_1} \leq C \quad \text{and,} \tag{A.28}$$

$$\sqrt{n}\frac{W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)}{\tau_1} \to \infty, \quad \text{as } n \to \infty, \tag{A.29}$$

then SigClust is asymptotically consistent. On the other hand if,

$$\frac{\tau_0}{\tau_1} \leq C \quad \text{and,} \tag{A.30}$$

$$W_1(\boldsymbol{\mu}_1) = W_0(\boldsymbol{\mu}_0) \tag{A.31}$$

then SigClust is asymptotically inconsistent.

This Lemma provides sufficient conditions for consistency and inconsistency respectively and we proceed to verify these conditions in the sequel. The proof of this Lemma is straightforward and is omitted.

To find the expression for $W_0(\boldsymbol{\mu}_0)$, note that in (3.2) we had calculated the population optimal within sum of squares for regular $2-$means clustering. But now since the test statistic considers a symmetric version of $2-$means clustering we need a version of Lemma 3.0.1 for the within sum of squares for symmetric $2-$means clustering,

$$W_n^{(0)}(t) = \frac{1}{n}\sum_{i=1}^{n} \min\{||X_i - t||^2, ||X_i + t||^2\},$$

as given by (3.5). This is easily provided by an analogous version of Theorem 3.2 for a single Normal distribution as follows:

Lemma A.1.10. *Let the data be generated from $N(0, D_0)$, as defined above, and τ_0 and $W_0(\boldsymbol{\mu}_0)$ be as given by* (3.3) *and* (3.2)*. Then as $n \to \infty$,*

$$\sqrt{n}(W_n^{(0)}(\mathbf{b_n}^{(0)}) - W_0(\mu_0)) \rightsquigarrow N(0, \tau_0^2),$$

where $W_n^{(0)}(\mathbf{b_n}^{(0)}) = \min_t W_n^{(0)}(t)$, the minimum within sum of squares for symmetric $2-$means clustering

We skip the proof of this lemma as it follows exactly along the lines of the proof of Theorem 3.2 along with the observation that the unique μ_0 that minimizes the within sum of squares for regular $2-$means is itself symmetric and hence it also minimizes the symmetric version. Additionally the positive definiteness of the corresponding matrix G_0 has already been shown in Lemma A.0.1.

So given the expressions for τ_0 and $W_0(\boldsymbol{\mu}_0)$ in (3.3) and (3.2), it now remains to calculate τ_1 and $W_1(\boldsymbol{\mu}_1)$ to analyze the power of SigClust. The following Lemma builds on Theorem 3.3 to calculate these quantities.

We analyze two cases which depend on whether the optimal population-level split occurs along the first or second coordinate.

Lemma A.1.11. *There are universal constants $0 < c \leq C$ such that:*

1. *If* (3.7) *holds, then:*

$$W_1(\boldsymbol{\mu}_1) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \kappa^2.$$
$$c \leq \tau_1^2 \leq C.$$

2. *If* (3.9) *holds, then:*

$$W_1(\boldsymbol{\mu}_1) = W_0(\boldsymbol{\mu}_0),$$
$$c \leq \tau_1^2 \leq C.$$

We prove this result in Appendix A.3.2. To complete the proof of the Theorem we need to put together Lemmas A.1.9 and A.1.11 to show the consistency and inconsistency of SigClust in different regimes.

We note that using (A.33) and the result of Lemma A.1.11 that both τ_0^2 and τ_1^2 are bounded by constants (recall that we take $\{a, \sigma_1^2, \ldots, \sigma_d^2\}$ to be fixed) as $n \to \infty$ verifying Conditions (A.28) and (A.30). Thus, Lemmas A.1.9 and A.1.11 directly yield the inconsistency of SigClust when condition (3.9) holds. On the other hand, in order to establish consistency when (3.7) holds, to verify Condition (A.29) we note that for some constant $c > 0$,

$$\begin{aligned}\sqrt{n}\frac{W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)}{\tau_1} &\geq c\sqrt{n}\left[W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)\right] \\ &= c\sqrt{n}\underbrace{\left[\kappa^2 - \frac{2}{\pi}\max\left\{\left(\sigma_1^2 + \frac{a^2}{4}\right), \sigma_2^2\right\}\right]}_{T},\end{aligned}$$

so to complete the proof of the Theorem it suffices to lower bound the term $T > 0$ as $n \to \infty$, when (3.7) holds. Clearly in this regime $\kappa^2 - 2\sigma_2^2/\pi > 0$ so Lemma A.1.7 as stated in Appendix A.2.2, completes the proof of our Theorem.

A.3.1 Proof of Lemma A.1.8

Under the null, the distribution of the statistic follows from Theorem 3.1 and Lemma A.1.10. Concretely, for

$$W_0(\boldsymbol{\mu}_0) = \widetilde{\sigma}^2 - \frac{2}{\pi}\max\left\{\left(\sigma_1^2 + \frac{a^2}{4}\right), \sigma_2^2\right\}, \tag{A.32}$$

$$\tau_0^2 = 2\sum_{i=2}^{d}\sigma_i^4 + 2\left(\sigma_1^2 + \frac{a^2}{4}\right)^2 - \frac{16}{\pi^2}\left[\max\left\{\left(\sigma_1^2 + \frac{a^2}{4}\right), \sigma_2^2\right\}\right]^2, \tag{A.33}$$

we have by a combination of Theorem 3.1 and Lemma A.1.10 that we would expect under the null that,

$$\sqrt{n}\Big(T_n^{(0)} - \frac{W_0(\boldsymbol{\mu}_0)}{\widetilde{\sigma}^2}\Big) \rightsquigarrow N\Big(0, \Big[\frac{\tau_0}{\widetilde{\sigma}^2}\Big]^2\Big), \quad \text{as} \quad n \to \infty.$$

Thus, we reject at level α, if:

$$\sqrt{n}\Big(T_n^{(0)} - \frac{W_0(\boldsymbol{\mu}_0)}{\widetilde{\sigma}^2}\Big) \leq \frac{\tau_0\Phi^{-1}(\alpha)}{\widetilde{\sigma}^2}.$$

Under the alternate we can once again use Theorem 3.2 to obtain that,

$$\sqrt{n}\Big(T_n^{(0)} - \frac{W_1(\boldsymbol{\mu}_1)}{\widetilde{\sigma}^2}\Big) \rightsquigarrow N\Big(0, \Big[\frac{\tau_1}{\widetilde{\sigma}^2}\Big]^2\Big), \tag{A.34}$$

where $W_1(\boldsymbol{\mu}_1)$ denotes the optimal 2-means objective under the alternate, and

$$\tau_1^2 = \sum_{i=1}^{2}\mathbb{P}_1(A_i)\mathbb{E}_{X\sim\mathbb{P}_1}[\|X - \mathbb{E}[X|X\in A_i]\|^4|X\in A_i] - [W_1(\boldsymbol{\mu}_1)]^2,$$

where $\{A_1, A_2\}$ denotes the Voronoi partition induced by $\boldsymbol{\mu}_1$. Accordingly letting $\mathbb{P}_1$ denote the distribution in (A.27) we have that,

$$\begin{aligned}
\text{Power}_n(a) &= \mathbb{P}_1\left(\sqrt{n}\Big(T_n^{(0)} - \frac{W_0(\boldsymbol{\mu}_0)}{\widetilde{\sigma}^2}\Big) \leq \frac{\tau_0\Phi^{-1}(\alpha)}{\widetilde{\sigma}^2}\right) \\
&= \mathbb{P}_1\left(\frac{\sqrt{n}\widetilde{\sigma}^2}{\tau_1}\Big(T_n^{(0)} - \frac{W_1(\boldsymbol{\mu}_1)}{\widetilde{\sigma}^2}\Big) \leq \frac{\tau_0\Phi^{-1}(\alpha)}{\tau_1} + \sqrt{n}\frac{W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)}{\tau_1}\right) \\
&\overset{\text{(i)}}{=} \Phi\left(\frac{\tau_0\Phi^{-1}(\alpha)}{\tau_1} + \sqrt{n}\frac{W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)}{\tau_1}\right),
\end{aligned}$$

where (i) follows from (A.34).

A.3.2 Proof of Lemma A.1.11

We divide our analysis into two cases, according to the optimal 2-means solution.

When condition (3.7) **holds:** In this case, the population optimal 2-means split is along the first coordinate. The expression for $W_1(\boldsymbol{\mu}_1)$ follows from (3.8), and it only remains to bound τ_1^2. To lower bound τ_1^2 we note that,

$$\begin{aligned}
\tau_1^2 &= \sum_{i=1}^{2} \mathbb{P}_1(A_i)\mathbb{E}_{X\sim\mathbb{P}_1}[\|X - \mathbb{E}[X|X\in A_i]\|^4|X\in A_i] - [W_1(\boldsymbol{\mu}_1)]^2 \\
&= 3\sum_{j=2}^{d}\sigma_j^4 + \sum_{i,j\neq 1, j\neq i}\sigma_i^2\sigma_j^2 + \mathbb{E}[(X_1-\kappa)^2|X_1\geq 0]\sum_{j\neq 1}\sigma_j^2 \\
&\qquad + \mathbb{E}[(X_1-\kappa)^4|X_1\geq 0] - \left[\mathbb{E}[(X_1-\kappa)^2|X_1\geq 0] + \sum_{j=2}^{d}\sigma_j^2\right]^2 \\
&= 2\sum_{j=2}^{d}\sigma_j^4 + \operatorname{var}\left((X_1-\kappa)^2|X_1\geq 0\right).
\end{aligned}$$

Using the fact that the variances and a are all fixed and bounded above and below we obtain that for two universal constants $0 < c \leq C$,

$$c \leq \tau_1^2 \leq C.$$

When condition (3.9) **holds:** In this case, the population optimal 2-means split is along the second coordinate. The expression for $W_1(\boldsymbol{\mu}_1)$ follows from (3.10), and once again it only remains to bound τ_1^2. In this case,

$$\tau_1^2 = \sum_{i=1}^{2} \mathbb{P}_1(A_i)\mathbb{E}_{X\sim\mathbb{P}_1}[\|X - \mathbb{E}[X|X\in A_i]\|^4|X\in A_i] - [W_1(\boldsymbol{\mu}_1)]^2.$$

Noting that,

$$\mathbb{E}[X_1^2] = \sigma_1^2 + \frac{a^2}{4}, \quad \mathbb{E}[X_1^4] = \frac{a^4}{16} + 3\sigma_1^4 + 3\sigma_1^2\frac{a^2}{2},$$

we obtain

$$\begin{aligned}
\tau_1^2 &= \left[\frac{a^4}{16} + 3\sigma_1^4 + 3\sigma_1^2\frac{a^2}{2}\right] + 2\sum_{j=3}^{d}\sigma_j^4 + \operatorname{var}[(X_2 - \sqrt{2}\pi\sigma_2)^4|X_2\geq 0] - \left[\sigma_1^2 + \frac{a^2}{4}\right]^2 \\
&= 2\sum_{j\neq 2}\sigma_j^4 + \operatorname{var}[(X_2 - \sqrt{2}\pi\sigma_2)^4|X_2\geq 0] + \sigma_1^2 a^2.
\end{aligned}$$

Once again using the fact that the variances and a are all fixed and bounded above and below we obtain that for two universal constants $0 < c \leq C$,

$$c \leq \tau_1^2 \leq C,$$

as desired.

A.4 Proof of the Main Results for RIFT

In this Appendix, we collect the proofs of the main results for the RIFTs in the thesis. In Sections A.4.1 and A.4.3 we consider the limiting distributions of the RIFT statistic, and its ℓ_2 counterpart under the null and prove Theorems 4.1 and 4.3. In Section A.4.2 we consider Theorem 4.2 and analyze the power of the RIFT and finally in Section A.4.4 we consider Theorem 4.4 where we verify the validity of the modified RIFT to test for mixtures of two Normals.

A.4.1 Proof of Theorem 4.1

In the following proof all probabilities and expectations are taken conditioned on $\mathcal{D}_1$. By the Berry-Esseen theorem, given $\mathcal{D}_1$,

$$\sup_t |\mathbb{P}(\sqrt{n}(\widehat{\Gamma} - \Gamma) \leq t) - \mathbb{P}(Z \leq t)| \leq \frac{C_0 \rho}{\tau^3 \sqrt{n}},$$

where C_0 is a constant,

$$\rho = \mathbb{E}\left[\left|\widetilde{R}_i - \Gamma\right|^3\right] \quad \text{and} \quad \tau^2 = \mathbb{E}\left[\left(\widetilde{R}_i - \Gamma\right)^2\right]. \tag{A.35}$$

Now $\Gamma = \mathbb{E}\left[\widetilde{R}_i\right]$, therefore,

$$\tau^2 = \mathrm{Var}\left(\widetilde{R}_i\right) = \mathrm{Var}\left(R_i + \delta Z_i\right) \geq \delta^2 \mathrm{Var}(Z_i) = \delta^2 > 0.$$

Now note that,

$$\begin{aligned}
|R_i| &= \left|\log\left(\frac{\widehat{p}_2(X_i)}{\widehat{p}_1(X_i)}\right)\right| = |\log \widehat{p}_2(X_i) - \log \widehat{p}_1(X_i)| \\
&\leq \max_x \{\log \widehat{p}_2(x) - \log \widehat{p}_1(x), \log \widehat{p}_1(x) - \log \widehat{p}_2(x)\} \\
&\leq \max_{x, i \in \{1,2\}} \left\{\log \widehat{f}_i(x) - \log \widehat{p}_1(x), \log \widehat{p}_1(x) - \log \widehat{f}_i(x)\right\}.
\end{aligned}$$

Then we get,

$$\begin{aligned}
|R_i| \leq \max_{i \in \{1,2\}} \Bigg\{ &\frac{1}{2}\log\left(\frac{|\widehat{\Sigma}|}{|\widehat{\Sigma}_i|}\right) + \frac{1}{2}\left(\widehat{\mu}^T \widehat{\Sigma}^{-1} \widehat{\mu} - \left(\widehat{\Sigma}^{-1}\widehat{\mu} - \widehat{\Sigma}_i^{-1}\widehat{\mu}_i\right)^T \left(\widehat{\Sigma}^{-1} - \widehat{\Sigma}_i^{-1}\right)^{-1} \left(\widehat{\Sigma}^{-1}\widehat{\mu} - \widehat{\Sigma}_i^{-1}\widehat{\mu}_i\right)\right), \\
&\frac{1}{2}\log\left(\frac{|\widehat{\Sigma}_i|}{|\widehat{\Sigma}|}\right) + \frac{1}{2}\left(\widehat{\mu}_i^T \widehat{\Sigma}_i^{-1} \widehat{\mu}_i - \left(\widehat{\Sigma}_i^{-1}\widehat{\mu}_i - \widehat{\Sigma}^{-1}\widehat{\mu}\right)^T \left(\widehat{\Sigma}_i^{-1} - \widehat{\Sigma}^{-1}\right)^{-1} \left(\widehat{\Sigma}_i^{-1}\widehat{\mu}_i - \widehat{\Sigma}^{-1}\widehat{\mu}\right)\right)\Bigg\}.
\end{aligned}$$

Now we notice that since for $i = 1, 2$, $\mu, \widehat{\mu}_i \in \mathcal{A}$ and the eigenvalues of $\widehat{\Sigma}, \widehat{\Sigma}_i$ lie in a bounded set, there exists a constant $k \geq 0$ such that

$$\widehat{\mu}^T \widehat{\Sigma}^{-1} \widehat{\mu} - \left(\widehat{\Sigma}^{-1}\widehat{\mu} - \widehat{\Sigma}_i^{-1}\widehat{\mu}_i\right)^T \left(\widehat{\Sigma}^{-1} - \widehat{\Sigma}_i^{-1}\right)^{-1} \left(\widehat{\Sigma}^{-1}\widehat{\mu} - \widehat{\Sigma}_i^{-1}\widehat{\mu}_i\right) \leq k,$$
$$\widehat{\mu}_i^T \widehat{\Sigma}_i^{-1} \widehat{\mu}_i - \left(\widehat{\Sigma}_i^{-1}\widehat{\mu}_i - \widehat{\Sigma}^{-1}\widehat{\mu}\right)^T \left(\widehat{\Sigma}_i^{-1} - \widehat{\Sigma}^{-1}\right)^{-1} \left(\widehat{\Sigma}_i^{-1}\widehat{\mu}_i - \widehat{\Sigma}^{-1}\widehat{\mu}\right) \leq k.$$

Also note that,

$$\left|\log\left(\frac{|\widehat{\Sigma}|}{|\widehat{\Sigma}_i|}\right)\right| \leq d \log\left(\frac{c_2}{c_1}\right) \quad \text{and} \quad \left|\log\left(\frac{|\widehat{\Sigma}_i|}{|\widehat{\Sigma}|}\right)\right| \leq d \log\left(\frac{c_2}{c_1}\right).$$

Therefore

$$|R_i| \leq \frac{1}{2}\left(d \log\left(\frac{c_2}{c_1}\right) + k\right) = C_1.$$

So we can also say,

$$|\Gamma| = |\mathbb{E}\left[R_i\right]| \leq C_1.$$

Then,

$$\begin{aligned}
\rho = \mathbb{E}\left[\left|\widetilde{R}_i - \Gamma\right|^3\right] &\leq \mathbb{E}\left[\left(|\widetilde{R}_i| + |\Gamma|\right)^3\right] \\
&\leq \mathbb{E}\left[\left(2C_1 + \delta|Z_i|\right)^3\right] \\
&= 8C_1^3 + 12C_1^2\delta\mathbb{E}[|Z_i|] + 6C_1\delta^2\mathbb{E}[Z_i^2] + \delta^3\mathbb{E}[|Z_i|^3] \\
&= 8C_1^3 + \delta\left[12C_1^2\sqrt{\frac{2}{\pi}} + 6C_1\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right].
\end{aligned}$$

Therefore,

$$\frac{C_0\rho}{\tau^3} \leq \frac{C_0}{\delta^3}\left[8C_1^3 + \delta\left(12C_1^2\sqrt{\frac{2}{\pi}} + 6C_1\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right)\right].$$

Hence,

$$\sup_t |\mathbb{P}(\sqrt{n}(\widehat{\Gamma} - \Gamma) \leq t) - \mathbb{P}(Z \leq t)| \leq \frac{C}{\sqrt{n}},$$

where $C = \frac{C_0}{\delta^3}\left[8C_1^3 + \delta\left(12C_1^2\sqrt{\frac{2}{\pi}} + 6C_1\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right)\right]$. Since the upper bound does not depend on $\mathcal{D}_1$, the result holds unconditionally as well. □

A.4.2 Proof of Theorem 4.2

Let $p \in \mathcal{P}_2 - \mathcal{P}_1$. Conditional on $\mathcal{D}_1$, $\mathbb{E}[\widehat{\Gamma}|\mathcal{D}_1] = \Gamma = K(p, \widehat{p}_1) - K(p, \widehat{p}_2)$. There exists $\gamma > 0$ such that $K(p, p_1) \geq \gamma > 0$ for all $p_1 \in \mathcal{P}_1$. It follows from the law of large numbers, with probability 1, that

$\liminf_{n\to\infty} K(p,\widehat{p}_1) > \gamma/2$. Since $\widehat{p}_2$ is consistent, $K(p,\widehat{p}_2) = o_{\mathbb{P}}(1)$. Thus, with probability 1, $\Gamma > \gamma/2$ for all large n. Also, with probability 1, $\widehat{\tau}/\tau = 1 + o(1)$. Combining these facts with the Berry-Esseen result, we have that

$$\begin{aligned}
\mathbb{P}\left(\widehat{\Gamma} > \frac{z_\alpha \widehat{\tau}}{\sqrt{n}} \,\middle|\, \mathcal{D}_1\right) &= \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\Gamma}-\Gamma)}{\tau} > (1+o(1))z_\alpha - \frac{\sqrt{n}\Gamma}{\tau} \,\middle|\, \mathcal{D}_1\right) \\
&= \mathbb{P}(Z > (1+o(1))z_\alpha - \sqrt{n}\Gamma/\tau|\mathcal{D}_1) + \frac{C}{\sqrt{n}} \\
&\geq \mathbb{P}(Z > (1+o(1))z_\alpha - \sqrt{n}\gamma/(2\tau)) + \frac{C}{\sqrt{n}}
\end{aligned}$$

where $Z \sim N(0,1)$ and C is a constant that does not depend on $\mathcal{D}_1$. It follows that $\mathbb{P}(\widehat{\Gamma} > z_\alpha\widehat{\tau}/\sqrt{n}) \to 1$. □

A.4.3 Proof of Theorem 4.3

In the following proof all probabilities and expectations are taken conditioned on $\mathcal{D}_1$. By Berry-Esseen theorem, given $\mathcal{D}_1$,

$$\sup_t |\mathbb{P}(\sqrt{n}(\widehat{\Theta}-\Theta) \leq t) - \mathbb{P}(Z \leq t)| \leq \frac{C_0 \mathbb{E}\left[\left|\widetilde{U}_i - \Theta\right|^3\right]}{a^3\sqrt{n}},$$

where C_0 is a constant and $a^2 = \mathbb{E}\left[\left(\widetilde{U}_i - \Theta\right)^2\right]$. Let

$$a^2 = \mathrm{Var}\left(\widetilde{U}_i\right). \tag{A.36}$$

Now $\Theta = \mathbb{E}[U_i] = \mathbb{E}\left[\widetilde{U}_i\right]$, therefore,

$$a^2 = \mathrm{Var}\left(\widetilde{U}_i\right) = \mathrm{Var}\left(U_i + \delta Z_i\right) \geq \delta^2 \mathrm{Var}(Z_i) = \delta^2.$$

Now note that,

$$\begin{aligned}
|U_i| &= |\widehat{p}_1(X_i) - \widehat{p}_2(X_i)| \\
&\leq |\widehat{p}_1(X_i)| + |\widehat{p}_2(X_i)| \\
&\leq \frac{1}{(2\pi)^{d/2}|\widehat{\Sigma}|^{1/2}} + \max_{i\in\{1,2\}} \frac{1}{(2\pi)^{d/2}|\widehat{\Sigma}_i|^{1/2}} \leq \frac{2}{(2\pi c_1)^{d/2}} = C_2.
\end{aligned}$$

Therefore we also have that,

$$|\Theta| = |\mathbb{E}[U_i]| \leq C_2.$$

Then following the same arguments as before while finding a bound for ρ, we can see that

$$\mathbb{E}\left[\left|\widetilde{U}_i - \Theta\right|^3\right] \leq 8C_2^3 + \delta\left[12C_2^2\sqrt{\frac{2}{\pi}} + 6C_2\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right].$$

Therefore,

$$\frac{C_0\mathbb{E}\left[\left|\widetilde{U}_i - \Theta\right|^3\right]}{a^3} \leq \frac{C_0}{\delta^3}\left[8C_2^3 + \delta\left(12C_2^2\sqrt{\frac{2}{\pi}} + 6C_2\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right)\right].$$

Hence,

$$\sup_t |\mathbb{P}(\sqrt{n}(\widehat{\Gamma} - \Gamma) \leq t) - \mathbb{P}(Z \leq t)| \leq \frac{\widetilde{C}}{\sqrt{n}},$$

where $\widetilde{C} = \frac{C_0}{\delta^3}\left[8C_2^3 + \delta\left(12C_2^2\sqrt{\frac{2}{\pi}} + 6C_2\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right)\right]$. □

A.4.4 Proof of Theorem 4.4

Suppose H_0 is true. Crucially, the error from the Berry-Esseen theorem does not depend on $\mathcal{D}_1$. The unconditional type I error of the split test is thus

$$\begin{aligned}
\mathbb{P}\left(\widehat{\Gamma} > \frac{z_\alpha\widehat{\tau}}{\sqrt{n}}\right) &= \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\Gamma} - \Gamma)}{\widehat{\tau}} > z_\alpha - \frac{\sqrt{n}\Gamma}{\widehat{\tau}}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\frac{\sqrt{n}(\widehat{\Gamma} - \Gamma)}{\widehat{\tau}} > z_\alpha - \frac{\sqrt{n}\Gamma}{\widehat{\tau}} \;\middle|\; \mathcal{D}_1\right)\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(Z > z_\alpha - \frac{\sqrt{n}\Gamma}{\tau} \;\middle|\; \mathcal{D}_1\right)\right] + O(n^{-1/2}) \\
&= \mathbb{E}\left[\overline{\Phi}\left(z_\alpha - \frac{\sqrt{n}\Gamma}{\tau}\right)\right] + O(n^{-1/2})
\end{aligned}$$

where $Z \sim N(0,1)$, $\overline{\Phi} = 1 - \Phi$ and Φ is the normal cdf.

(i) Since $\widehat{p}_1$ is a consistent estimator of the true density p under H_0, $K(p, \widehat{p}_1)$ converges to zero. More precisely, $K(p, \widehat{p}_1) = o_p(c_n/\sqrt{n})$, where c_n is some appropriately chosen, slowly diverging sequence. (ii) By construction, $K(p, \widehat{p}_2) > \Delta$, where $\Delta > 0$ is a fixed positive constant. From (i) and (ii), it follows that with probability tending to 1,

$$\frac{\sqrt{n}\Gamma}{\tau} := \frac{\sqrt{n}(K(p, \widehat{p}_1) - K(p, \widehat{p}_1))}{\tau} \leq \frac{\sqrt{n}(\{c_n/\sqrt{n}\} - \Delta)}{\tau} \to -\infty$$

and thus

$$z_\alpha - \frac{\sqrt{n}\Gamma}{\tau} \to \infty$$

(where $\tau \geq \delta > 0$ as shown on p.45). Hence it follows that

$$P\left(\sqrt{n}\frac{\widehat{\Gamma}}{\widehat{\tau}} > z_\alpha\right) = E\left[\overline{\Phi}\left(z_\alpha - \frac{\sqrt{n}\Gamma}{\tau}\right)\right] + O(n^{-1/2}) = o(1)$$

(where Φ is the standard normal cdf). $\square$

Appendix B

Exploratory Data Analysis of the Higgs Boson Data

As mentioned in Chapter 9, the data set is from a machine learning challenge hosted by Kaggle at https://www.kaggle.com/c/higgs-boson which consists of simulated data provided by the ATLAS experiment at CERN to optimize the analysis of the Higgs boson.

We analyze the training set provided by the challenge to demonstrate the performance of the classifier tests as well as understand the behavior of the signal. The training set has 250K observations, and $d = 30$ features whose individual details can be found in the Appendix B of Adam-Bourdarios et al. (2014).

As mentioned in Chapter 9 we will only be looking at the "raw" quantities measured by the detector, i.e. features prefixed with `PRI` (for PRImitives), since the derived features are just functions of the primitive features. Additionally, in order to avoid missing data, we only consider the events that have two jets (`PRI_jet_num`= 2) which results in 50,379 events, 24,645 background events and 25,734 signal events.

Among the primitive features, five of them provide the azimuth angle ϕ of the particles generated in the event (variables ending with `_phi`). These features are rotation invariant in the sense that the event doesn't change if all of them are rotated together with any angle. The first row of Figure B.1 demonstrates the uniform distribution of the phi variables. The phi variables themselves don't contain any information, but the difference between the angles is what contains the information. Hence to interpret these variables more easily using the active subspace methods, we remove the invariance of the azimuth angle variables by rotating all the ϕ's and setting the azimuth angle of the leading jet at 0 (`PRI_leading_phi`= 0). The bottom row of Figure B.1 demonstrates the importance of the change in the distribution of the angles after the rotation. The symmetry of the distributions is expected as a difference of $-\pi$ radians is the same as a difference of $-\pi$ radians.

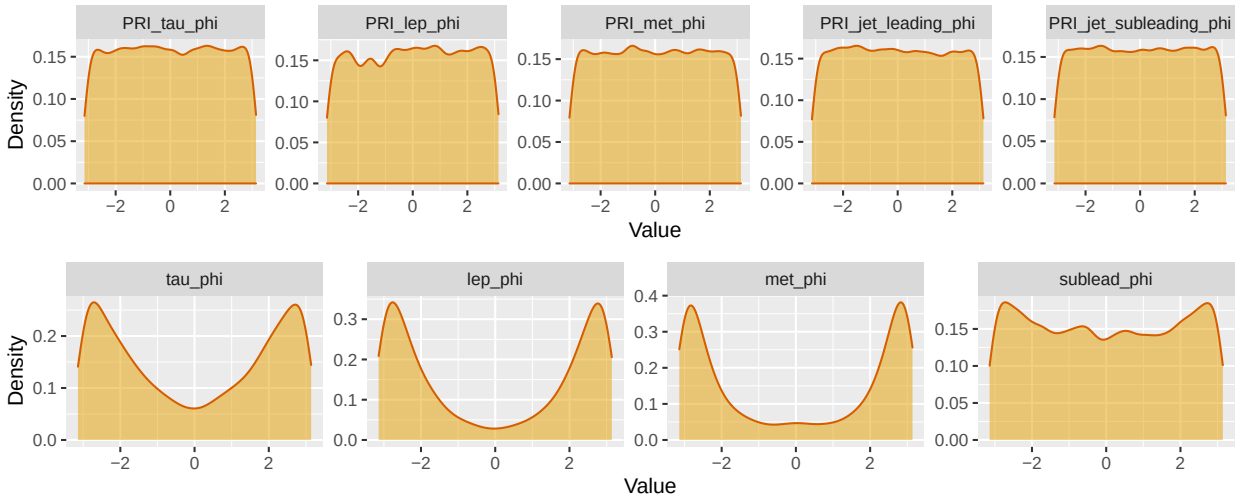


Figure B.1: Top row gives the phi variables before rotation. Bottom row gives the phi variables after rotation such that the phi of the leading jet is set to 0.

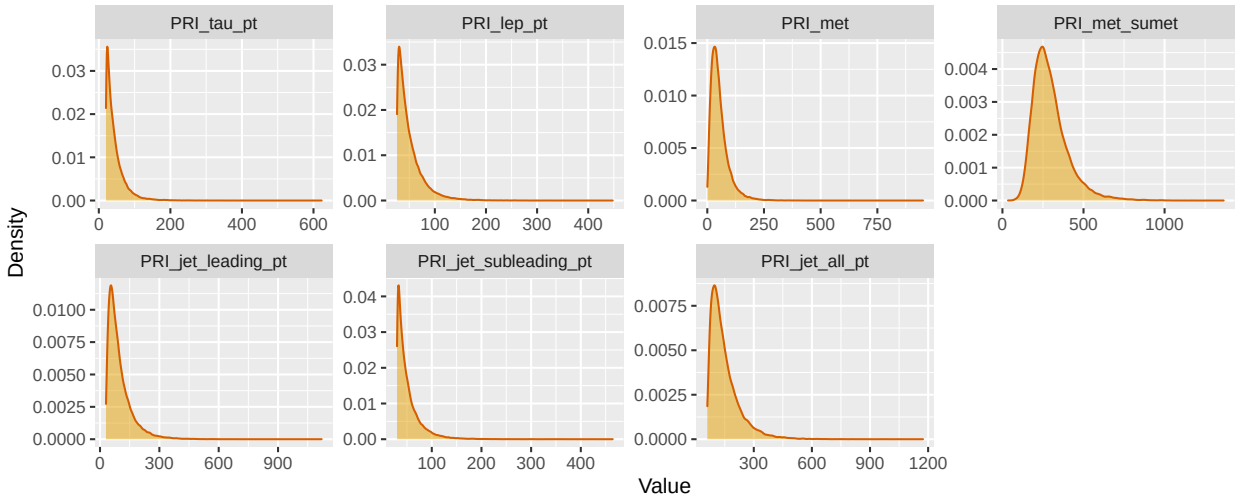


Figure B.2: Distribution of the variables for which we consider a log transformation.

Additionally, we take logarithmic transformations of the variables that give the transverse momentum of the particles produced (variables ending with `_pt`), the missing transverse energy (`PRI_met`) and the total transverse energy in the detector (`PRI_met_sumet`). This is done so our analysis is not affectced by the skewness demonstrated by these variables in Figure B.2. Taking a log transformation of these variables fixes the problem upto some extent.

Our goal is to detect the presence of the Higgs boson signal in the experimental data, using this data set. The difficulty of this problem is demonstrated by Figure B.3 which shows that the distributions of the signal and the background data are not very different. Particularly, when we are searching for signal that

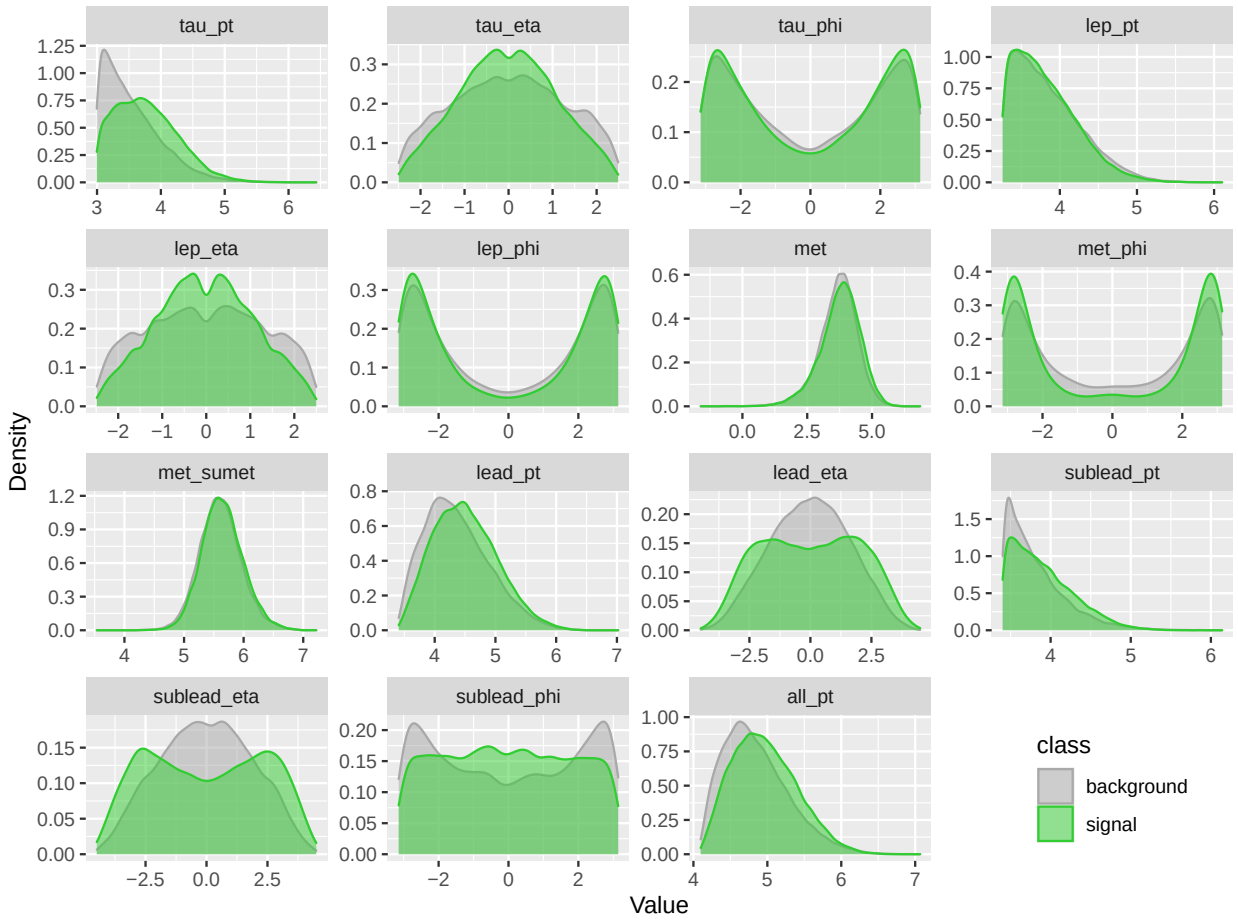


Figure B.3: Histograms of all the variables for signal (green) data as well as background (grey) data.

is just around 10% of the experimental data or even less, these minute differences are difficult to detect. In the next section, we explore the random forest classifier trained for a single random simulation (one of the 100 simulations explored in Section 9.2) when $\lambda = 0.1$, i.e., 10% of the experimental data is from the signal sample.

B.1 Analysis of a Single Semi-Supervised Simulation

As described in Section 9.2, for the semi-supervised methods, we consider a training set of m_1 = 7,322 background events and N_1 =7,323 experimental events, which contains $\lfloor N_1\lambda \rfloor = 732$ signal events. We test for the presence of signal using a test set of m_2 = 5,000 background events and N_1 =5,000 experimental events, which contains $\lfloor N_2\lambda \rfloor = 500$ signal events. We train a random forest classifier on the training data to differentiate between the background and the experimental events. These two data sets differ very slightly from each other as can be seen in Figure B.4; visually they are almost indistinguishable from the histograms.

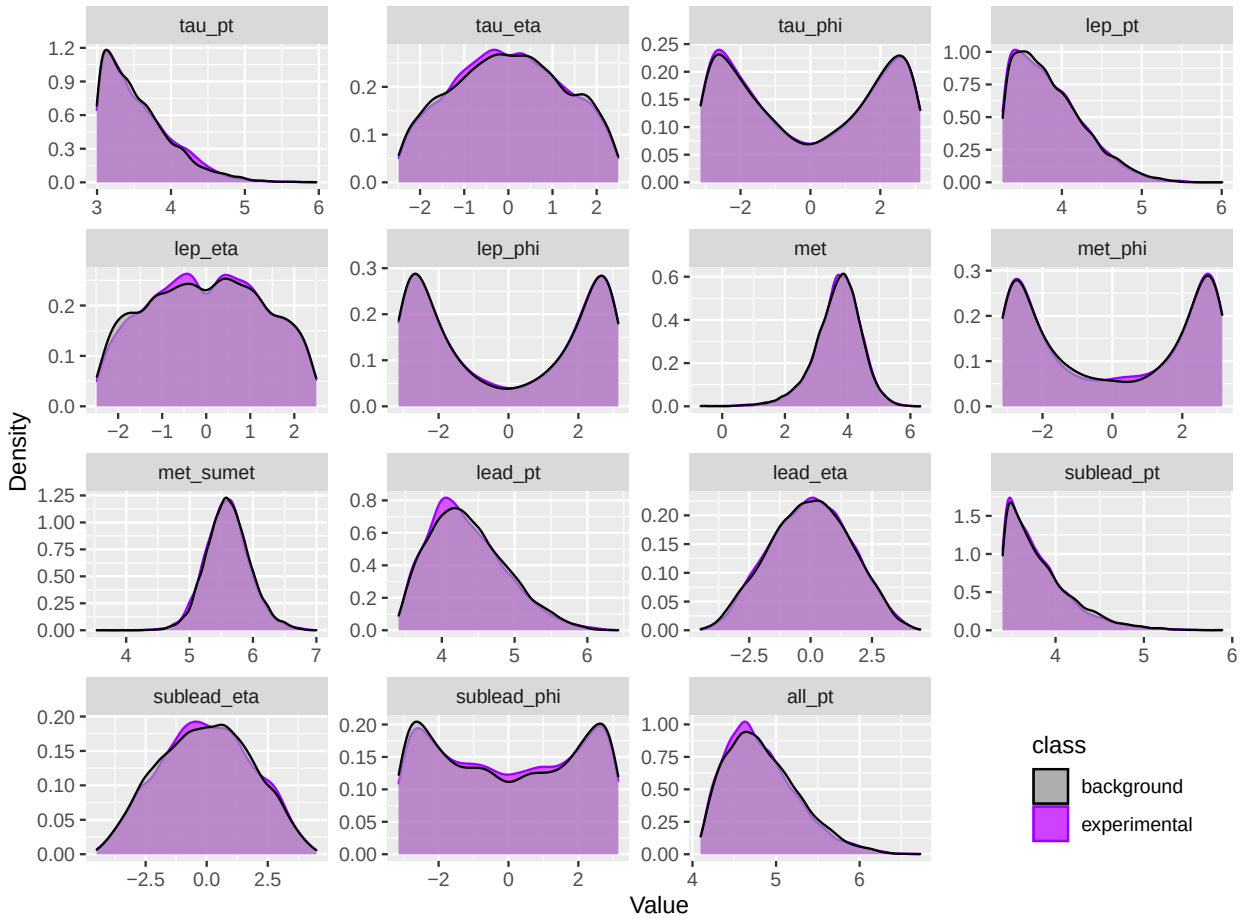
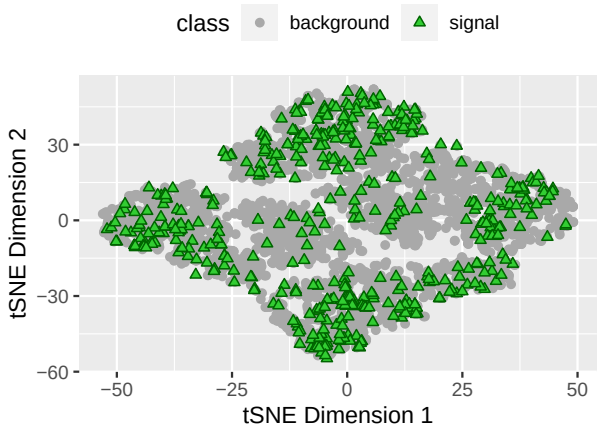


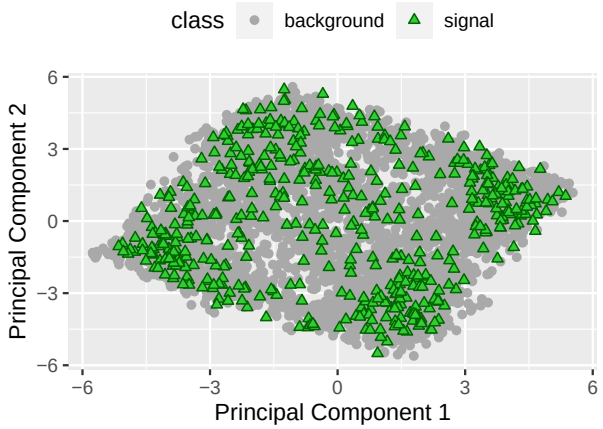
Figure B.4: Histograms of all the variables for training data: experimental (purple) data and background (grey) data.

To demonstrate this further, we incorporate the dependence of the variables on each other as well. We demonstrate two different approaches. First we use Principal Component Analysis on just the background data to find the two principal components of the background data and then project the test experimental data on those axes. Figure B.5c shows that the signal is not very distinguishable from the background. We then use t-distributed stochastic neighbor embedding proposed by Maaten and Hinton (2008) to visualize the data in two dimensions. First we train the algorithm to distinguish experimental data from the background data. When this fails, as shown in Figure B.5a, we directly train the algorithm to distinguish signal from background data. As shown in Figure B.5b, this approach fails as well. This emphasizes the difficulty of the problem to detect differences between the background data and the experimental data.



(a) Semi-supervised tSNE

(b) Supervised tSNE



(c) PCA

Figure B.5: Experimental and background test data containing signal events (green) and background events (grey). (a) t-distributed stochastic neighbor embedding (tSNE) trained on experimental versus background training samples. (b) t-distributed stochastic neighbor embedding (tSNE) trained on signal versus background training samples. (c) Principal component analysis (PCA) trained on background training samples.

Since in the case of $\lambda = 0.1$, the random forests demonstrate some power in detecting the signal, we want recognize the variables affecting the membership probabilities. This is not an easy task as demonstrated by Figure B.6, which shows the random forest classifier output (experimental membership probabilities) as a function of each of the variables in the data.

We notice that the random classifier seems to depend on the transverse momentums of all the particles produced (variables ending with `_pt`), as well as the missing transverse energy (`met`) and the total transverse energy in the detector (`met_sumet`).

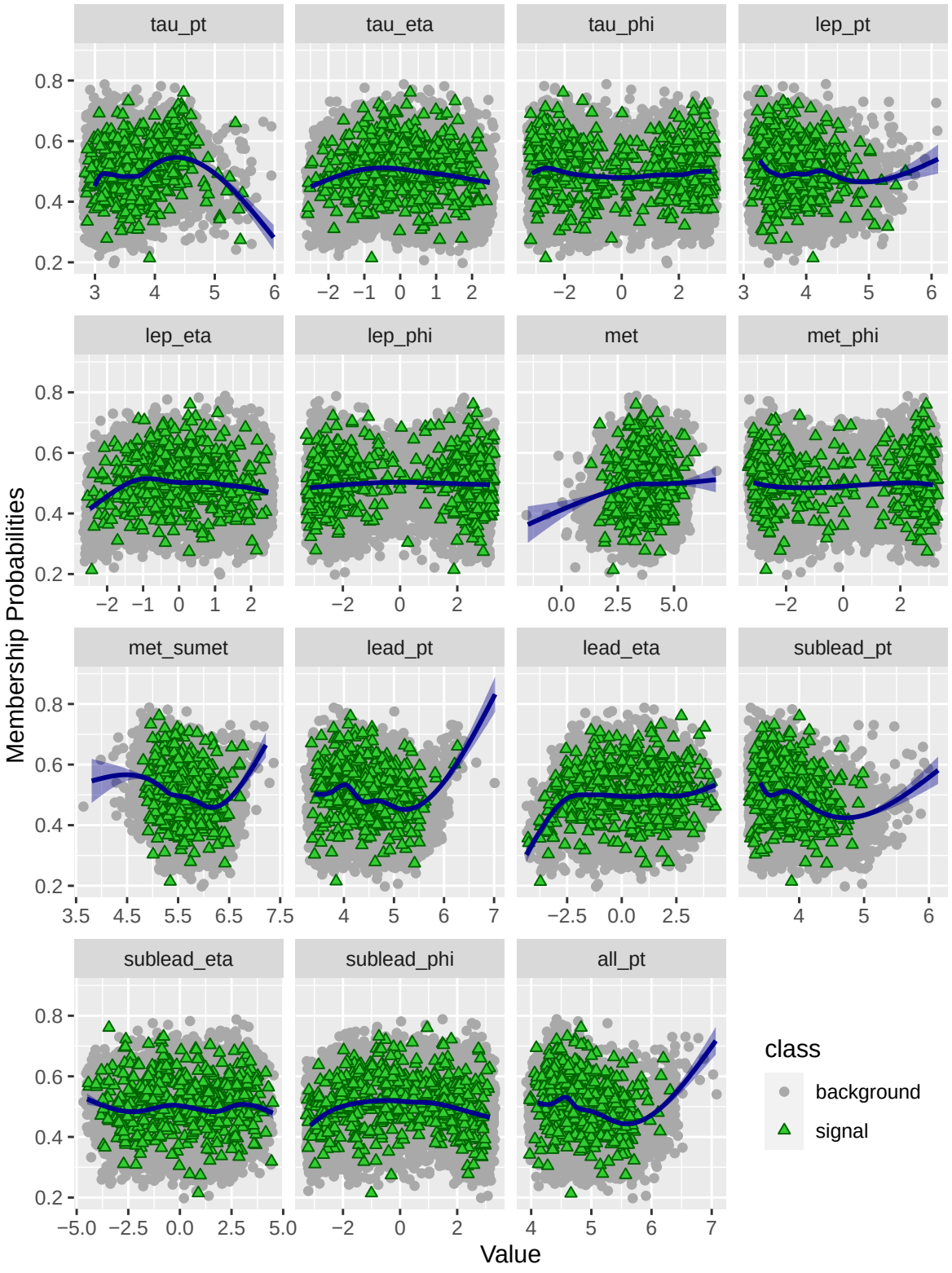


Figure B.6: Experimental membership probabilities (the random forest output) versus all the variables for the test data sets. Signal events in green and background events in grey.

To understand better how these variables affect the classifier we use active subspace methods introduced in Section 8.5 and show the results in Section 9.3. In order to select the smoothing parameter for the local linear smoother being used, we use the standard deviation of the variables scaled by a factor as the bandwidth. We explore a few scaling factors and calculate the gradients as well as the mean of the gradients for all of them. Following Method 8.5.1, we find the mean projection corresponding to the mean gradient. We choose a scaling factor that demonstrates maximum amount of difference between the signal and the background distributions when the data is projected along that direction.

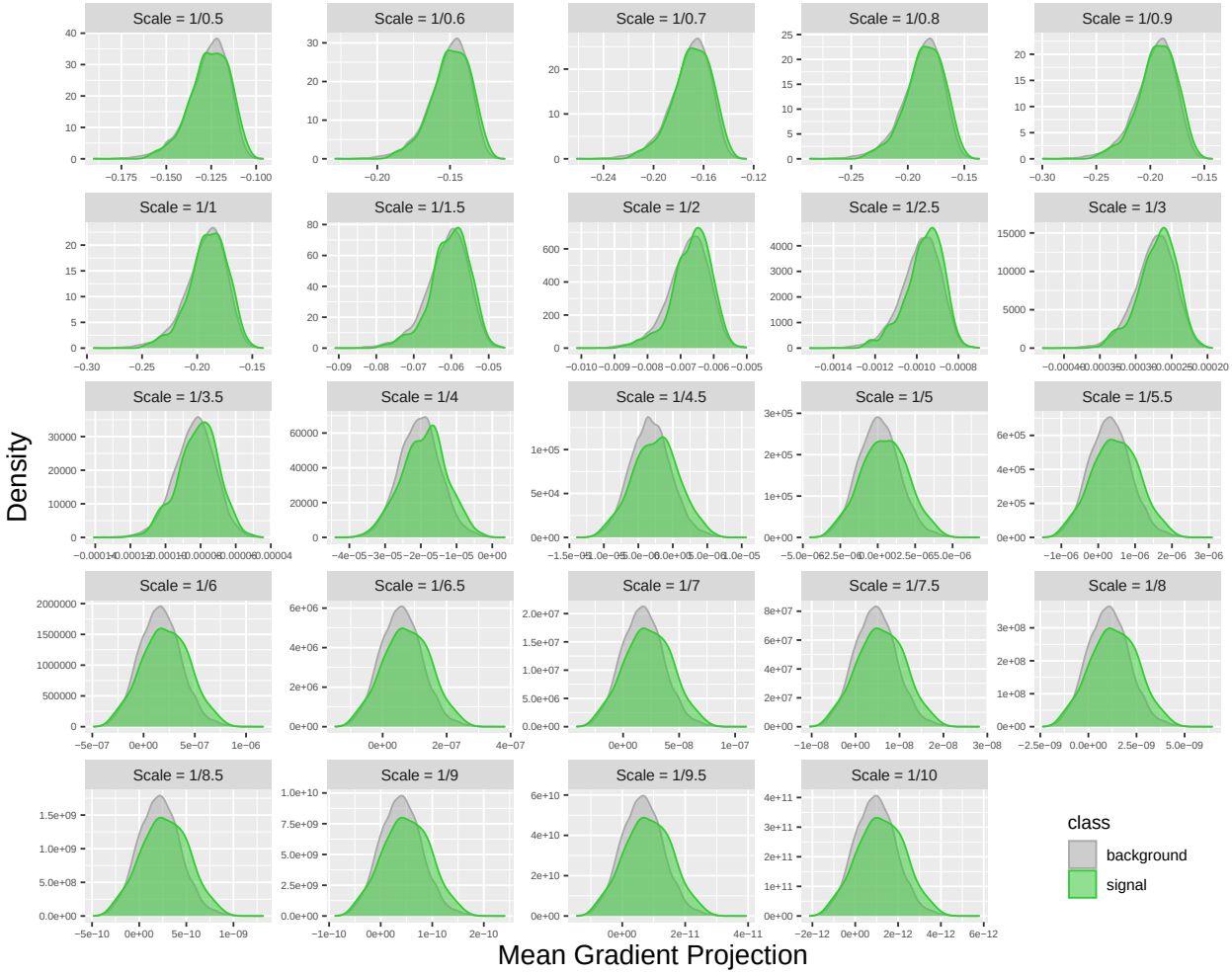


Figure B.7: Histograms of the signal (green) and background (grey) data projected onto the mean projection vector when the standard deviation of the variables scaled by a factor is used as the bandwidth for the local linear smoother

Figure B.7 shows that scaling the standard deviation by anything larger than 4.5 seems to be similar. So we consider dividing by 5 and use that as the bandwidth for the results presented in Section 9.3. The mean projection vector as well as the first two active subspace vectors are presented in Section 9.3. Figure B.8

presents the third, fourth and the fifth active subspace vectors using PCA. Sparse PCA results in only two non-zero principal components.

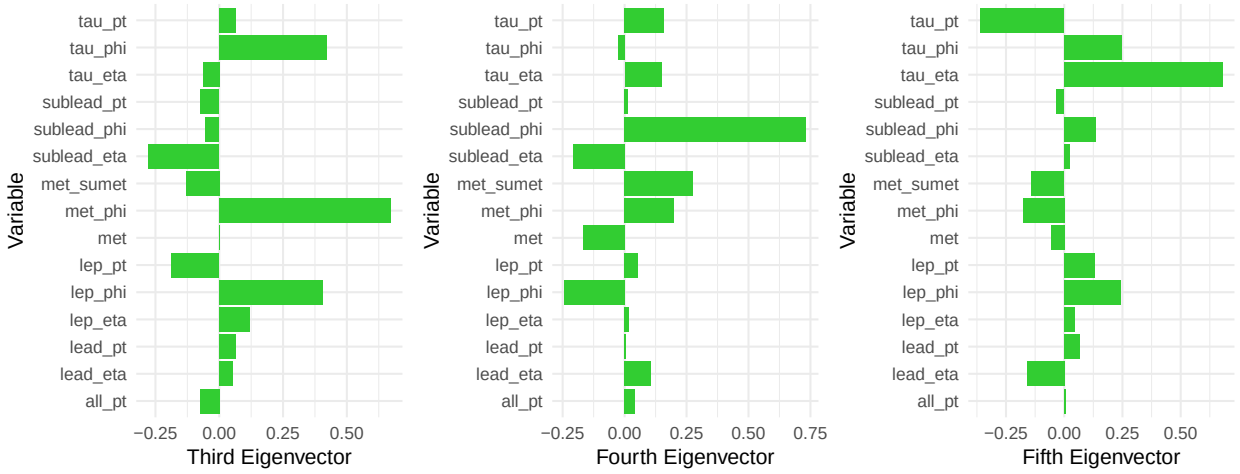


Figure B.8: Third to fifth active subspace variables.

We further use the active subspace methods for a simulation with $\lambda = 1.5$. Figure B.9 shows that scaling the standard deviation by 1.5 or 2 seems to result in the most difference between the signal and background distributions. So we consider dividing the standard deviations of the variables by 2 and using that as the bandwidth for the results presented in Section 9.3. The mean projection vector as well as the first two active subspace vectors are presented in Section 9.3. Figure B.8 presents the third, fourth and the fifth active subspace vectors using PCA. Sparse PCA results in only two non-zero principal components.

We note that the sparse PCA identifies the transverse momentum of the hadron tau (`tau_pt`) and the relationship between the phi angles between the leading jet and the missing transverse energy (`met_phi`) as additional eigen vectors important for the detection of the signal.

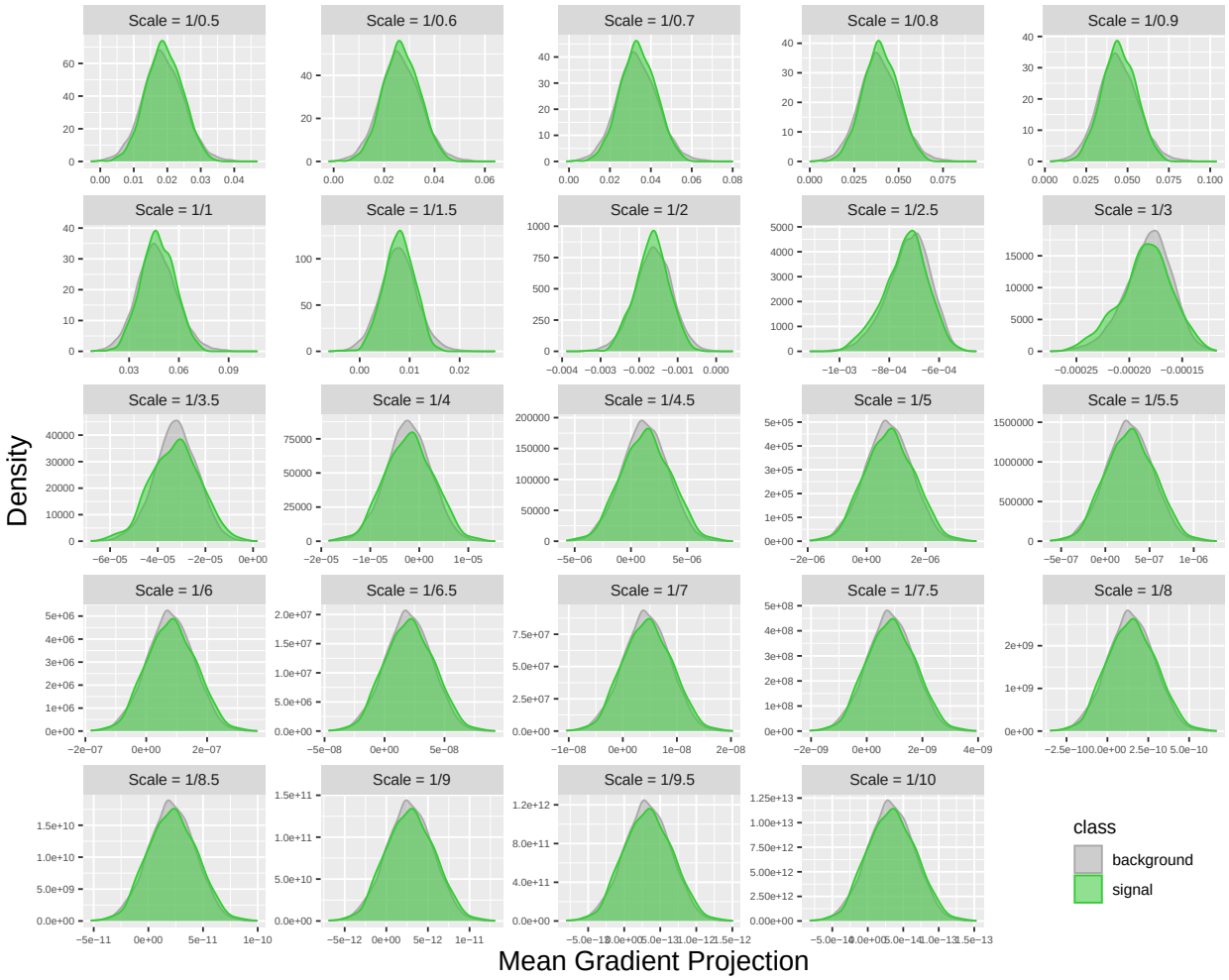


Figure B.9: Histograms of the signal (green) and background (grey) data projected onto the mean projection vector when the standard deviation of the variables scaled by a factor is used as the bandwidth for the local linear smoother

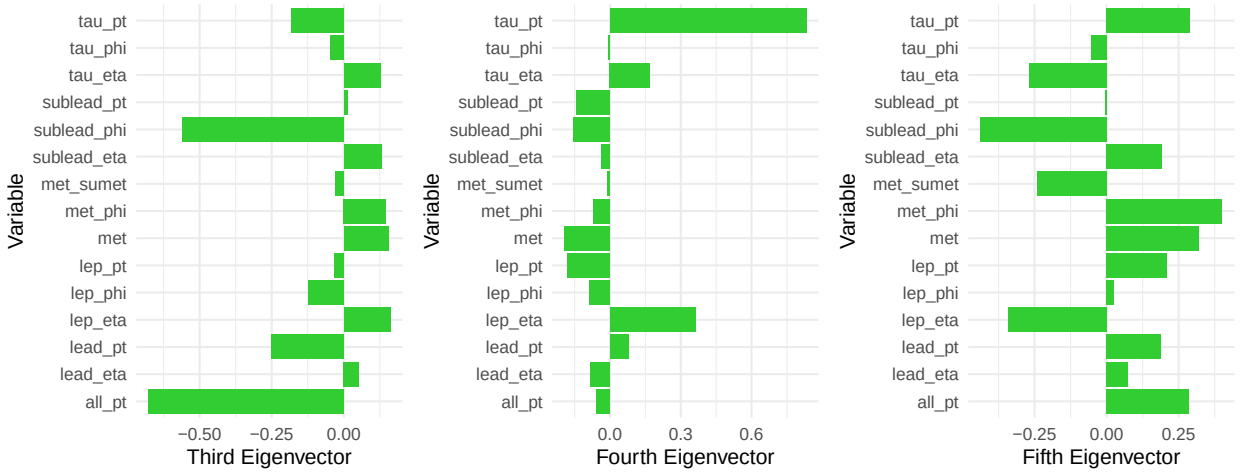


Figure B.10: Third to fifth active subspace variables.

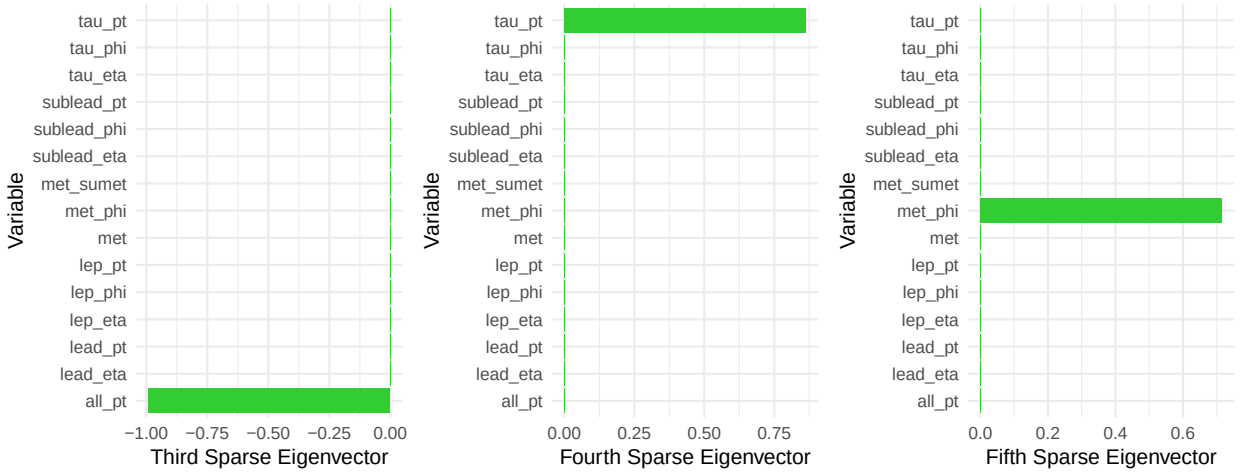


Figure B.11: Third to fifth sparse active subspace variables.

Vita

EDUCATION

Department of Statistics & Data Science, Carnegie Mellon University *May 2020*

PhD in Statistics & Data Science *(expected)*

Dissertation Title: Inference for Clustering and Anomaly Detection

Thesis committee: Larry Wasserman (Advisor), Sivaraman Balakrishnan, Mikael Kuusela, Andrew B. Nobel, Rebecca Nugent, Alessandro Rinaldo

Machine Learning Department, Carnegie Mellon University *May 2018*

Secondary Masters in ML

Indian Statistical Institute, Kolkata *May 2014*

Master of Statistics

Specialization: Mathematical Statistics and Probability

Indian Statistical Institute, Kolkata *May 2012*

Bachelor of Statistics (*Hons.*)

PUBLICATIONS

Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests

Purvasha Chakravarti, *Mikael Kuusela and Larry Wasserman*

In preparation, 2020.

Gaussian Mixture Clustering Using Relative Tests of Fit

Purvasha Chakravarti, *Sivaraman Balakrishnan and Larry Wasserman*

In revision, Journal of the American Statistical Association (JASA) Theory and Methods, 2020. Preprint - arXiv:1910.02566

A Generalization of Convolutional Neural Networks to Graph-Structured Data

Yotam Hechtlinger, **Purvasha Chakravarti** *and Jining Qin*

Preprint - arXiv:1704.08165

Spatially Adaptive Kernel Regression Using Risk Estimation

Sunder Ram Krishnan, Chandra Sekhar Seelamantula and **Purvasha Chakravarti**

Published in IEEE Signal Processing Letters, 2014

SELECTED TALKS

Gaussian Mixture Clustering Using Relative Tests of Fit 2019

Contributed Talk at Joint Statistical Meetings, Denver, Colorado.

A Generalization of Convolutional Neural Networks to Graph-Structured Data 2019

Poster at The Science of Deep Learning, National Academy of Sciences Arthur M. Sackler Colloquium, Washington, D.C.

Hierarchical Significance Testing for Gaussian Mixture Clustering 2018

Contributed Talk at Joint Statistical Meetings, Vancouver, Canada.

Gaussian Mixture Clustering Using Relative Tests of Fit (RIFTs) 2018

Working Group on Model-Based Clustering Summer Session, Ann Arbor, Michigan.

Statistical Significance of k-Means Clustering 2017

Contributed Talk at Joint Statistical Meetings, Baltimore, Maryland.

Statistical Analysis of the Chikungunya Fever 2016

Women in Statistics and Data Science Conference, Charlotte, North Carolina.

Women in Statistics at Carnegie Mellon University 2016

Women in Statistics and Data Science Conference, Charlotte, North Carolina.

Jointly presented with Shannon Gallagher.

AWARDS AND HONORS

Honorable Mention for the 2019 Do-Bui Travel Award

Received an Honorable Mention for Do-Bui Travel Award given by the Caucus for Women in Statistics (CWS), Joint Statistical Meetings, 2019.

National Level Scholarship

Obtained the INSPIRE Scholarship offered by the Department of Science and Technology (DST), Government of India from 2009 - 2014.

Cyber Olympiad

Secured All India rank 19 in 5th National Cyber Olympiad held on 19th February, 2006.

Summer Fellowship

Received Indian Academy of Science Fellowship, 2012.

TEACHING EXPERIENCE

Instructor:

CMU 36-225 Introduction to Probability Theory Summer 2019

9 credits undergraduate course with 43 students

Evaluation (5.00) (Response Rate: 91%): Overall = 4.1; Teaching = 4.15

CMU 36-200 Reasoning with Data Summer 2018

9 credits undergraduate course with 8 students

Evaluation (5.00) (Response Rate: 25%): Overall = 5; Teaching = 3.5

CMU 36-226 Introduction to Statistical Inference Summer 2017, 2016

9 credits undergraduate course with 31 students and 26 students respectively.

Evaluation (5.00) (Response Rate: 57%, 67%): Overall = 4.69, 4.42; Teaching = 4.75, 4.42

Teaching Assistant:

CMU 36-303 Sampling, Survey and Society Spring 2019

CMU 36-401 Modern Regression Fall 2018, 2014

CMU 36-402 Advanced Methods for Data Analysis Spring 2018, 2017, 2016

CMU 36-705 Intermediate Statistics Fall 2017, 2016

CMU 36-225 Introduction to Probability Theory Fall 2015

CMU 36-625 Probability and Mathematical Statistics (Hons.) Spring 2015

PROFESSIONAL AFFILIATION AND SERVICE

Member: CMU Women in Statistics, American Statistical Association, Institute of Electrical and Electronics Engineers.

Volunteer for Women in Data Science Pittsburgh @CMU 2018, 2019

Women in Data Science Pittsburgh @CMU Conference, Pittsburgh, PA

Panelist for Women in Statistics at Carnegie Mellon University 2016

Women in Statistics and Data Science Conference, Charlotte, NC

Gave an Outreach Talk on Opportunities in Statistics 2016

Winchester Thurston, Pittsburgh, PA

Cultural Chair (Music), Indian Graduate Student Association 2016-present

Carnegie Mellon University, Pittsburgh, PA

Convenor of Irene J.Curie Hall 2011-2013

Indian Statistical Institute, Kolkata, India