Gaussian Mixture Clustering Using Relative Tests of Fit

Purvasha Chakravarti Sivaraman Balakrishnan
Larry Wasserman
Department of Statistics and Data Science,
Carnegie Mellon University,
Pittsburgh, PA, 15213.
{pchakrav, siva, larry}@stat.cmu.edu

October 8, 2019

Abstract

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 a significance guarantee. 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. One version of our test provides exact, finite sample control of the type I error. We show how our tests can be used for hierarchical clustering as well as in a sequential manner for model selection. We conclude with an extensive simulation study and a cluster analysis of a gene expression dataset.

1 Introduction

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 paper 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 paper 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.

1.1 Related Work

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 unimodal 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 nonregularity 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.

1.2 Outline

In Section 2 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 section also has results on the geometric properties of k-means clustering in a special case, which is a prelude to finding the power. In Section 3 we describe our new procedure. We also describe several other tests that are used for comparison. In Section 4 we show how to use our new tests in a hierarchical framework. Section 5 describes a sequential testing version of our approach which can be used for model selection for the GMM. We consider some examples in Section 6, and analyze a gene expression dataset in Section 7. Finally, concluding remarks are in Section 8. We defer the technical details of most proofs to the Appendix.

1.3 Notation

Throughout this paper 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.

2 Setup and the SigClust Procedure

We 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 $\mathbb{R}^d$. We recall the k-means clustering algorithm which chooses cluster centers $\mathbf{b_n} = (b_{n1}, \ldots, b_{nk}) \in \mathbb{R}^{d\times k}$ to minimize the within-cluster sum of squares,

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

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.

2.1 SigClust

In this section, we describe the SigClust procedure. Liu et al. (2008) define a cluster as a population sampled from a single Gaussian distribution. To capture the non-Gaussianity due to the presence of multiple clusters, Liu et al. (2008) consider a test statistic based on k-means, with $k = 2$. Specifically, define T_n to be the ratio between the within-class sum of squares and the total sum of squares as,

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

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.

Roughly, we reject the null for small values of this statistic. In order to estimate the p-value we 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.

2.2 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 B.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 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

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

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 B.1, we establish the following claims:

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

$$\tau^2 = 2\sum_{i=1}^{d} \sigma_i^4 - \frac{16\sigma_1^4}{\pi^2}. \tag{3}$$

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

Theorem 1. *For* $W(\boldsymbol{\mu})$ *and* τ^2 *defined in* (2) *and* (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 4 we analyze its power. The proof of Theorem 1 is quite long and technical. Most of the work is done in the Appendix. Here is a brief proof that leverages Lemma 1 which contains most of the technical details.

Proof. From Lemma 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. □

2.3 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

$$\{X_1, \ldots, X_n\} \sim \frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D), \tag{4}$$

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 symmertric centers. That is, we consider $\mathbf{b_n}^{(0)}$ that minimizes the within-cluster sum of squares,

$$W_n^{(0)}(t) := W_n(t, -t) = \frac{1}{n}\sum_{i=1}^{n} \min\{||X_i - t||^2, ||X_i + t||^2\}, \tag{5}$$

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 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,*

$$G_{ij} = \begin{cases} 2\mathbb{P}(A_i)\mathbf{I}_d - 2r_{ij}^{-1}\int_{M_{ij}} f(x)(x-\mu_i^*)(x-\mu_i^*)^T \, d\sigma(x) & \text{for} \quad i = j \\ -2r_{ij}^{-1}\int_{M_{ij}} f(x)(x-\mu_i^*)(x-\mu_j^*)^T \, d\sigma(x) & \text{for} \quad i \neq j, \end{cases} \tag{6}$$

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:

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

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

Theorem 3. *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*

$$\sigma_2^2 < \sigma_1^2 + \frac{a^2}{4}, \tag{7}$$

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*

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

2. *When*

$$\sigma_2^2 > \max \left\{ \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4 + \frac{a^4}{64}}}{2\sigma_1^2}, \frac{\pi}{2} \left(\sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\ \sigma_1^2}} + \frac{a}{2}\ P\left(|Z| < \frac{a}{2\ \sigma_1} \right) \right)^2 \right\}, \tag{9}$$

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*

$$W(\mu^*) = W(b^*) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_2^2. \tag{10}$$

In simpler words, the theorem implies that when the condition in (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 (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 (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 (9) holds, the optimal 2-means split at the population-level is along the second covariate.

Additionally we also have the following lemma:

Lemma 2. *In both the cases mentioned in Theorem 3, the matrix* G *given by equation (6) is positive definite.*

Therefore Theorem 3 and the above Lemma 2 combined together with Theorem 2 give the limiting distribution under the alternative.

2.4 Power

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

$$T_n^{(0)} := T_n^{(0)}\left(\mathbf{b_n}^{(0)}\right) = \frac{W_n^{(0)}(\mathbf{b_n}^{(0)})}{\frac{1}{n}\sum_{i=1}^n \|X_i - \overline{X}\|^2}. \tag{11}$$

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 1 and additionally on Theorem 2, we show the following result:

Theorem 4. *Suppose that samples are generated according to the model described in (4) and let $Z \sim N(0,1)$ then:*

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

$$\sigma_2^2 < \sigma_1^2 + \frac{a^2}{4}, \tag{12}$$

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

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

$$\sigma_2^2 > \max\left\{ \frac{2\sigma_1^4 + \frac{a^4}{16} + \frac{a^2}{2}\sqrt{\sigma_1^4 + \frac{a^4}{64}}}{2\sigma_1^2}, \frac{\pi}{2}\left(\sqrt{\frac{2}{\pi}}\,\sigma_1\,\exp\left(-\frac{a^2}{8\sigma_1^2}\right) + \frac{a}{2}\,P\left(|Z| < \frac{a}{2\,\sigma_1}\right)\right)^2 \right\} \tag{13}$$

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 $a > 0$:

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

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 section 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 D. At a high-level, the proof follows from Theorem 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, when the condition in (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 (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.

3 A Test For Relative Fit of Mixtures

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.

3.1 The Basic Test

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 paper.

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

$$\Gamma = K(p, \widehat{p}_1) - K(p, \widehat{p}_2), \tag{14}$$

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 $\mathcal{D}_1$,

$$H_0 : \Gamma \leq 0 \quad \text{versus} \quad H_1 : \Gamma > 0. \tag{15}$$

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

$$\widehat{\Gamma} = \frac{1}{n}\sum_{i\in\mathcal{D}_2} R_i \tag{16}$$

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

$$\sqrt{n}(\widehat{\Gamma} - \Gamma) \rightsquigarrow N(0, \tau^2)$$

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 $Z_1, \ldots, Z_n \sim N(0,1)$ 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 5. *Let $Z \sim N(0, \tau^2)$ where $\tau^2 = \mathbb{E}[(\widetilde{R}_i - \Gamma)^2 | \mathcal{D}_1]$. Then, under H_0*

$$\sup_t \left| \mathbb{P}(\sqrt{n}(\widehat{\Gamma} - \Gamma) \le t \,|\, \mathcal{D}_1) - \mathbb{P}(Z \le t) \right| \le \frac{C}{\sqrt{n}} \tag{17}$$

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 $\mathcal{D}_1$:

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

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,

$$\Gamma^* = \inf_{p_1 \in \mathcal{P}_1} K(p, p_1) - \inf_{p_2 \in \mathcal{P}_2} K(p, p_2) > 0, \tag{18}$$

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 6. *Suppose that Γ^* in (18) is strictly positive, then* RIFT *is asymptotically consistent, i.e. as $n \to \infty$.*

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

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 4 does not happen for our test.

3.2 Variants of RIFT

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

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 $W_i = \log(p(X_i)/q(X_i))$.

For relative fit we define

$$\widetilde{\Gamma} = \text{Median}_p[R] \tag{19}$$

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.

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 $Z_1, \ldots, Z_n \sim N(0,1)$ 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 7. *Let $Z \sim N(0, a^2)$ where $a^2 = var(\widetilde{U}_i)$. Then, under H_0,*

$$\sup_t |\mathbb{P}(\sqrt{n}(\widehat{\Theta} - \Theta) \leq t | \mathcal{D}_1) - \mathbb{P}(Z \leq t)| \leq \frac{\widetilde{C}}{\sqrt{n}}, \tag{20}$$

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.

3.3 Aside: A Test for Mixtures

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 8. *If $P \in \mathcal{P}_1$ then $\mathbb{P}(\widehat{\Gamma} > z_\alpha \widehat{\tau}/\sqrt{n}) = \alpha + o(1)$.*

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.

3.4 Truncation

If we use RIFT for top-down hierarchical clustering, as described later in Section 4, then after the first split, the null hypothesis will be a truncated Normal rather than a Normal since the test is now applied to the data in a cluster. 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.

3.5 Other Tests

Another way to decide whether to split the Normal 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.

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

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

2. NN Test. 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

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

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

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

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.

4 Hierarchical Clustering

To propose a hierarchical version of RIFT, we apply the procedure recursively. Figures 6 and 7 in the Appendix describe the details for the top-down approach and Figure 8 in the Appendix describes the details for the bottom-up approach. The final clustering is given by the leaf nodes of the tree derived by the algorithms.

In each case 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 estimate the parameters 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.

5 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

$$\max_s \widehat{\Gamma}_{js} > \frac{z_{\alpha/m_j} \widehat{\tau}_{js}}{\sqrt{n}} \tag{21}$$

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 $\widehat{\tau}_{js}^2 = \frac{1}{n} \sum_{i \in \mathcal{D}_2} (R_i - \overline{R})^2$.

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 3. *Under H_{0j},*

$$\limsup_{n \to \infty} \mathbb{P}(\text{rejecting } H_{0j}) \leq \alpha. \tag{22}$$

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

6 Simulations

In this section we compare SigClust and the RIFT variants we proposed through a variety of simulations. In Section 6.1 we investigate the asymptotic normality of the RIFT statistic defined in (16) under the null. In Section 6.2 we compare the power of various tests for detecting and splitting a mixture of two Gaussians. Finally, in Sections 6.3 and 6.4 we study hierarchical clustering using the RIFT statistic and evaluate model selection using the sequential RIFT procedure.

6.1 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 5. 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 $j = 2, \ldots, d$.

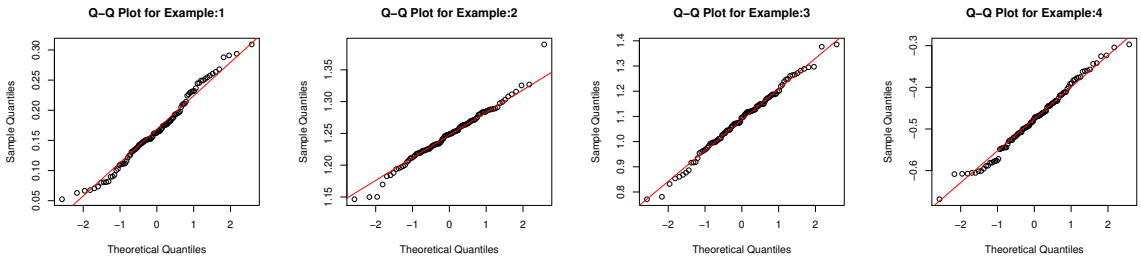

Figure 1: Q-Q plots to check Normality of the RIFT test statistic.

The test statistic, $\widehat{\Gamma}$ defined in (16) is computed for 100 simulations in each of these cases. Figure 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 5.

6.2 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 different tests in detecting the two clusters. 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 2 gives the proportion of tests that reject the null hypothesis that the underlying distribution has just

one cluster. 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.

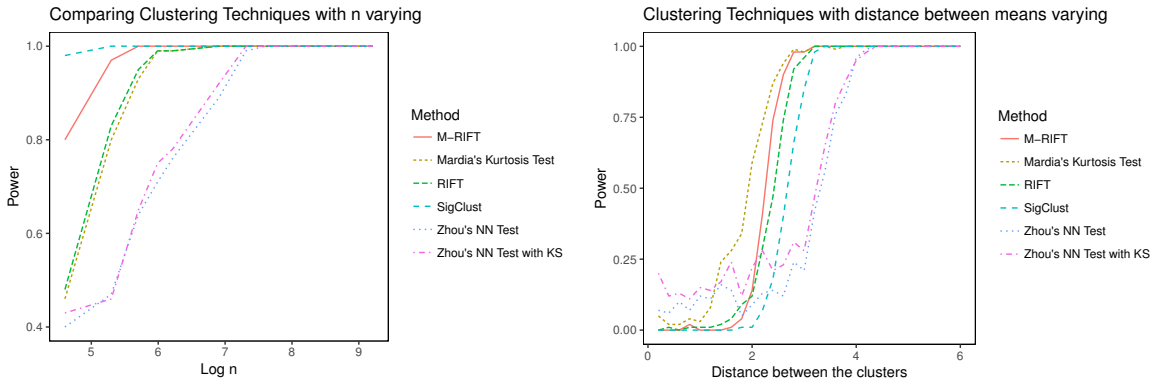

Figure 2: Comparing the power of the tests with varying number of observations, n and with increasing distance between the two mixture distributions (increasing a).

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. Figure 2 compares the proportion of times the tests detect the two distributions 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.

6.2.1 Signal in one direction

We compare the power of our test with SigClust while checking whether the tests control the type-I error. 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 3. We notice that the SigClust has very good power for both $a = 10$ and $a = 20$, whereas the RIFTs catch up for $a = 20$.

6.2.2 Signal in All Directions

Now we consider data with signal in all directions and compare the tests. 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 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.

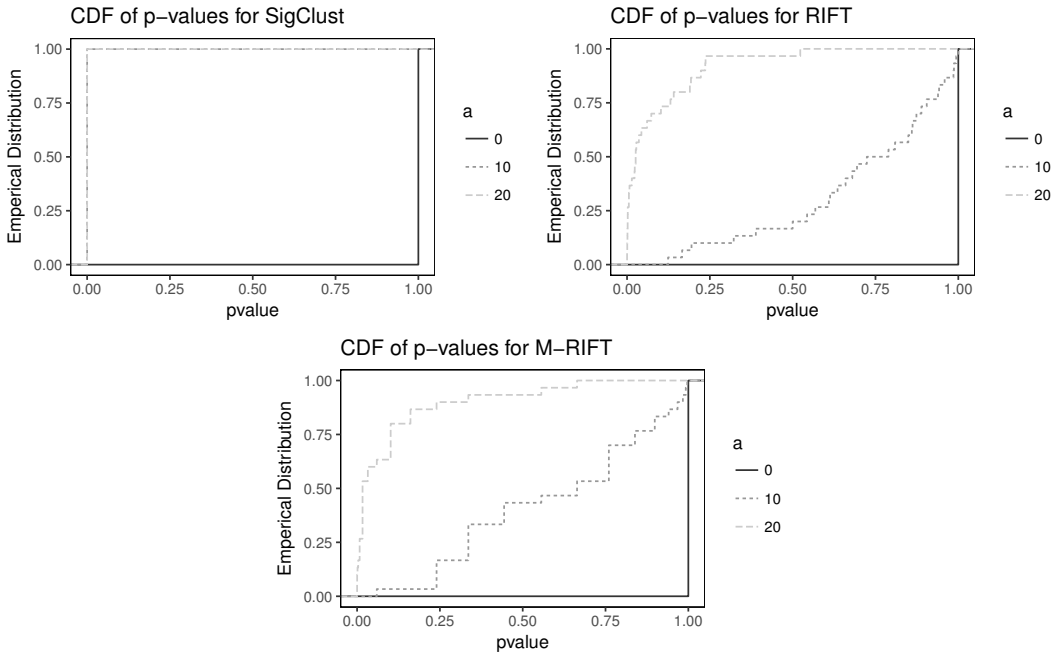

Figure 3: Comparing the empirical distribution of the p-values when signal is exactly in one direction.

6.2.3 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. 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 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.

6.3 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

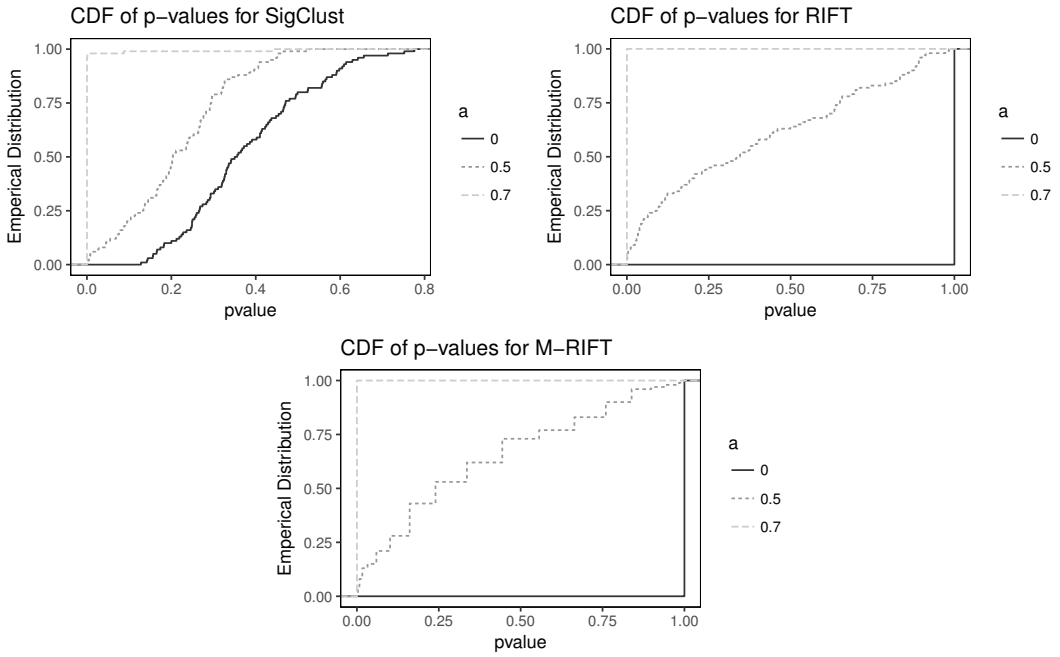

Figure 4: Comparing the empirical distribution of the p-values when signal is in all directions.

simulations.

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 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 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.

6.4 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.

First, 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 Gaussian components for 100 simulations. For each simulation, we use S-RIFT with the two different distances - Kullback-Leibler distance and ℓ_2 distance 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 2 gives the results of the simulations. 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

Table 1: Comparing the different algorithms for hierarchical clustering when samples are generated from a mixture of four Gaussian distributions. The table gives the number of simulations that identify the particular number of significant clusters over 100 replications.

Method	Algorithm type	Parameters			Number of clusters					
		d	δ	arr.	1	2	3	**4**	5	≥ 6
RIFT	Top-down	2	4	square	31	33	33	**3**	0	0
M-RIFT					5	9	43	**43**	0	0
SigClust					63	0	3	**15**	5	14
Trunc. SigClust					63	0	8	**8**	7	14
RIFT	Bottom-up	2	4	square	70	22	7	**1**	0	0
M-RIFT					16	36	41	**7**	0	0
SigClust					65	0	29	**5**	1	0
Trunc. SigClust					57	0	30	**5**	8	0
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

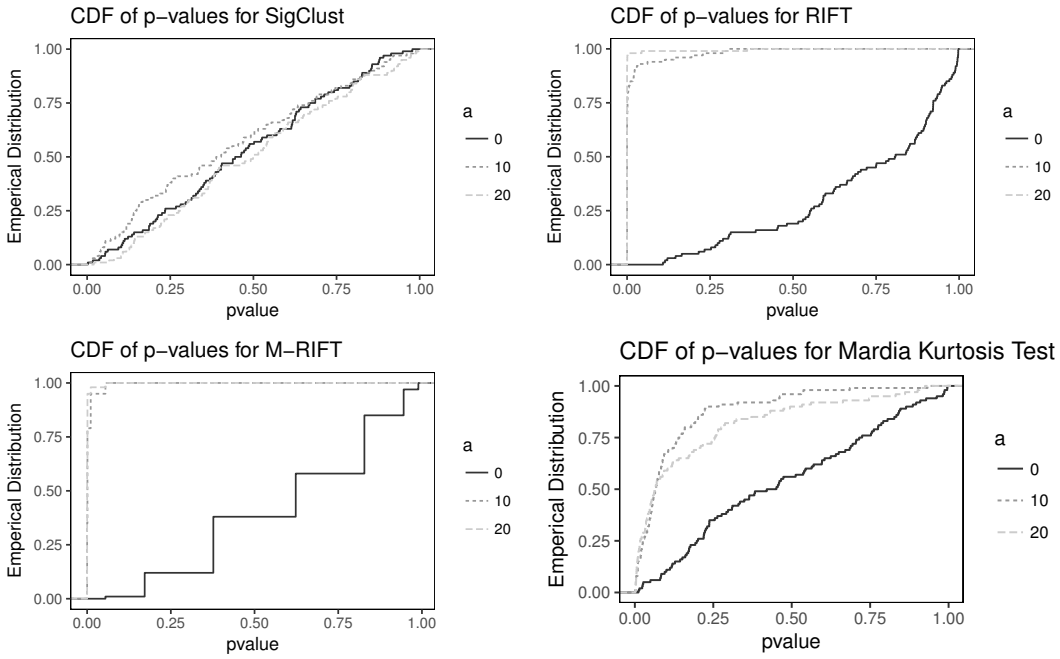

Figure 5: Comparing the power of the tests with higher signal in one direction and high variability in another.

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:

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

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.

The estimates of the number of clusters given by S-RIFT, AIC and BIC are recorded in Table 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.

6.5 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.

Table 2: Comparing the different algorithms for selecting the ideal number of clusters when samples are generated from a mixture of four Gaussian distributions. 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

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.

7 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 dataset. We consider a dataset 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

Table 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 4: Clusterings given by RIFT and SigClust for the multi-cancer gene expression dataset.

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

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 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 6.3, in this case also hierarchical clustering using Rift detects the true number of clusters much better than hierarchical clustering using SigClust.

8 Conclusion

We presented an analysis of the SigClust procedure of Liu et al. (2012) 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. (2012), 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 Normals 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.

References

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.

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.

Bock, H. H. (1985). On some significance tests in cluster analysis. *Journal of Classification*, 2(1):77–108.

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

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.

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

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

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.

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.

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

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.

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.

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

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*.

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

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

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.

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

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.

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

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

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

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.

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.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2012). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*.

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.

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

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.

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

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.

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.

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.

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

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

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

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.

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

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.

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

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

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.

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

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.

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.

A Hierarchical Clustering Algorithm Descriptions

In this Appendix, we summarize the algorithms. The summaries are in Figures 6 7, 8 and 9.

Top-Down Algorithm:

1. Split the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$. We estimate the parameters using $\mathcal{D}_1$ and create the tree using $\mathcal{D}_2$.
2. Set $\mathcal{D}_1^{T^{(0)}} = \mathcal{D}_1$, $\mathcal{D}_2^{T^{(0)}} = \mathcal{D}_2$. Use $\mathcal{D}_1^{T^{(0)}}$ to estimate the parameters required and $\mathcal{D}_2^{T^{(0)}}$ to test the null hypothesis that the data comes from a single cluster using the algorithms mentioned before.
3. If you accept the test at level $\alpha/2$, stop. If you reject the test at level $\alpha/2$, partition the sample space into two pieces $T_1^{(1)}$ and $T_2^{(1)}$. The partition also partitions $\mathcal{D}_1$ and $\mathcal{D}_2$, say into $\mathcal{D}_1^{T_1^{(1)}}$ and $\mathcal{D}_1^{T_1^{(1)}}$ and $\mathcal{D}_2^{T_2^{(1)}}$ and $\mathcal{D}_2^{T_2^{(1)}}$ respectively. Set Depth $= 1$.
4. If $\left\{i : T_i^{(\text{Depth})} \text{ is defined}\right\}$ is not an empty set, then for every $i \in \left\{i : T_i^{(\text{Depth})} \text{ is defined}\right\}$:
   (a) Set $T = T_i^{(\text{Depth})}$.
   (b) Use the corresponding sample sets $\mathcal{D}_1^T$ to estimate the parameters required and $\mathcal{D}_2^T$ to test the null hypothesis that the data comes from a single cluster using the algorithms mentioned before.
   (c) If you reject the test at level $\alpha/2^{2\text{Depth}+1}$, partition the sample space T into two pieces $T_{2i-1}^{(\text{Depth}+1)}$ and $T_{2i}^{(\text{Depth}+1)}$. Also partition $\mathcal{D}_1^T$ and $\mathcal{D}_2^T$, say into $\mathcal{D}_1^{T_{2i-1}^{(\text{Depth}+1)}}$ and $\mathcal{D}_1^{T_{2i}^{(\text{Depth}+1)}}$ and $\mathcal{D}_2^{T_{2i-1}^{(\text{Depth}+1)}}$ and $\mathcal{D}_2^{T_{2i}^{(\text{Depth}+1)}}$ respectively.
5. Set Depth $=$ Depth $+ 1$ and repeat step 4 till $\left\{i : T_i^{(\text{Depth})} \text{ is defined}\right\}$ is an empty set.

Figure 6: RIFT *applied in a top-down manner.*

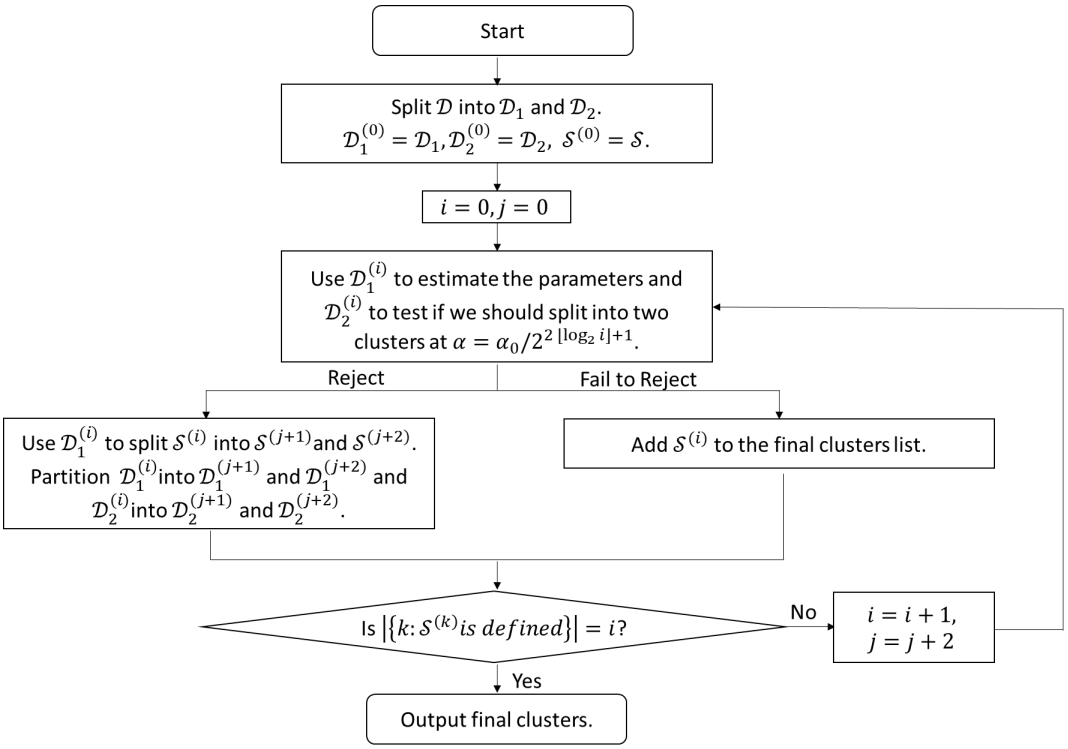

Figure 7: Top-down cluster tree algorithm.

Bottom-Up Algorithm:

1. Split the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$. We estimate the parameters using $\mathcal{D}_1$ and create the tree. Then using $\mathcal{D}_2$ we prune the tree by testing the significance of every split.

2. Let the entire sample space be denoted by $T_1^{(0)}$. Set $\mathcal{D}_1^{T_1^{(0)}} = \mathcal{D}_1$, $\mathcal{D}_2^{T_1^{(0)}} = \mathcal{D}_2$. Set Depth $= 0$.

3. If $\left\{i : T_i^{(\text{Depth})} \text{ is defined}\right\}$ is not an empty set, then for every $i \in \left\{i : T_i^{(\text{Depth})} \text{ is defined}\right\}$:

   (a) Set $T = T_i^{(\text{Depth})}$.

   (b) Use the corresponding sample sets $\mathcal{D}_1^T$ to estimate a mixture of two Gaussians and use that to partition the sample space T into two pieces $T_{2i-1}^{(\text{Depth}+1)}$ and $T_{2i}^{(\text{Depth}+1)}$. Also partition $\mathcal{D}_1^T$ and $\mathcal{D}_2^T$, say into $\mathcal{D}_1^{T_{2i-1}^{(\text{Depth}+1)}}$ and $\mathcal{D}_1^{T_{2i}^{(\text{Depth}+1)}}$ and $\mathcal{D}_2^{T_{2i-1}^{(\text{Depth}+1)}}$ and $\mathcal{D}_2^{T_{2i}^{(\text{Depth}+1)}}$ respectively.

4. Set Depth = Depth + 1 and repeat step 3 till $\left\{i : T_i^{(\text{Depth})} \text{ is defined}\right\}$ is an empty set.

5. If the total number of nodes in the tree is given by N_{nodes}, then set Depth = Depth $- 1$ and for every i such that $\{T_{2i-1}^{(\text{Depth})}, T_{2i}^{(\text{Depth})}\}$ is not empty:

   (a) Set $T = T_i^{(\text{Depth}-1)}$.

   (b) Use the sample set $\mathcal{D}_2^T$ to test the null hypothesis that the data comes from a single cluster using the algorithms mentioned before. If you fail to reject the test at level α/N_{nodes}, then delete $T_{2i-1}^{(\text{Depth})}$ and $T_{2i}^{(\text{Depth})}$. Also delete their sub-trees.

6. Set Depth = Depth $- 1$ and repeat step 5 till Depth $= 0$.

Figure 8: RIFT *applied in a bottom-up manner.*

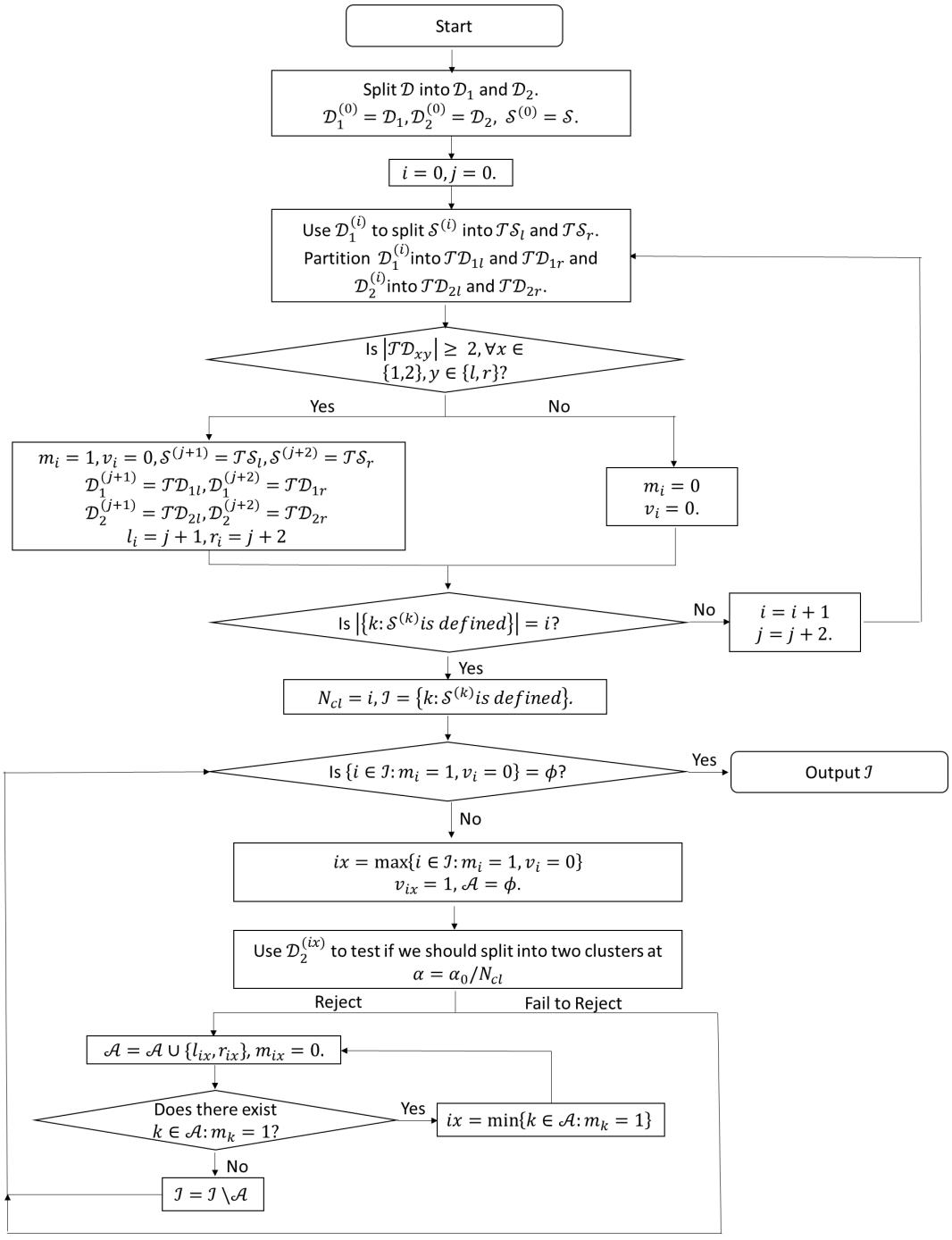

Figure 9: Bottom-up cluster tree algorithm.

B Proof of results presented under the null hypothesis of SigClust

In this Appendix, we prove Theorem 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 4 in Appendix B.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

$$V_i = 4\mathbb{E}\left[(X - \mu_i)(X - \mu_i)^T \mathbb{I}_{A_i}\right] \tag{23}$$

as its ith diagonal block and G_0 is analogously defined to the G as defined in equation (6) for the alternative. So, G_0 is a matrix made up of $d \times d$ matrixes of the form,

$$(G_0)_{ij} = \begin{cases} 2\mathbb{P}(A_i)\mathbf{I}_d - 2r_{ij}^{-1} \int_{M_{ij}} f(x)(x - \mu_i)(x - \mu_i)^T \, d\sigma(x) & \text{for} \quad i = j \\ -2r_{ij}^{-1} \int_{M_{ij}} f(x)(x - \mu_i)(x - \mu_j)^T \, d\sigma(x) & \text{for} \quad i \neq j, \end{cases} \tag{24}$$

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 4.
Now using Corollary 6.5 in Bock (1985), Lemma 1 follows immediately. In the next section, Appendix B.1, we prove Claims (2) and (3) which together with Lemma 1 give Theorem 1.

B.1 Proof of Claims (2) and (3)

Proof of Claim (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 (2).

Proof of Claim (3): In a similar fashion we can compute τ^2. Observe that,

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

$$\begin{aligned} &+ 2\sum_{i=2}^{d}\sum_{j=2,j\neq i}^{d}\mathbb{E}\left[X_i^2\right]\mathbb{E}\left[X_j^2\right] + \sum_{i=2}^{d}\mathbb{E}\left[X_i^4\right] \\ &= \sigma_1^4\left(3-\frac{4}{\pi}-\frac{12}{\pi^2}\right) + 2\sum_{i=2}^{d}\sigma_1^2\sigma_i^2\left(1-\frac{2}{\pi}\right) + 2\sum_{i=2}^{d}\sum_{j=2,j\neq i}^{d}\sigma_i^2\sigma_j^2 + 3\sum_{i=2}^{d}\sigma_i^4. \end{aligned}$$

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).

□

B.2 Proof of G_0 being positive definite

In order to use the result in Pollard (1982) to prove Lemma 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 4. *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 (24) is positive definite.

Proof of Lemma 4. The different blocks of the variance matrix are given by,

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

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

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 $i, j \in \{2, \ldots, d\}$,

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

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,

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

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 $2 \le j \le d$,

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

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.

□

B.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 $\mathbf{a} = (a_1, \dots, a_k)$ and the set of corresponding convex polyhedrons $A_1, \dots, A_k$ 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 9. *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 5 and 6 stated and proved below. □

Lemma 5. *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 $\forall\ 0 \le \beta < 2\pi$,*

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

Proof of Lemma 5. 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 6. *The asymptotic covariance between $\sqrt{n}B_n(\beta)$ and $\sqrt{n}B_n(\phi)$ is given by,*

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

Proof of Lemma 6. 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 $\alpha=|\beta-\phi|\le\pi$. 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

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

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,

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

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 $\beta = \tan^{-1}(X_{12}/X_{11})$ 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 $\lim_{n\to\infty} n\ \mathrm{V}(B_n(\beta)) = \frac{8}{\pi}\left(1-\frac{2}{\pi}\right)$. $\square$

C Proof of results presented under the alternate hypothesis of SigClust

In Section 2.3, we study the geometry of $k-$means under the alternative. Recall, under the alternative of SigClust we suppose that, we observe n samples from:

$$X \sim \frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D) \tag{25}$$

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 C.1 we first prove Theorem 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 (6).

We prove Lemma 2 that shows the positive definiteness of G in Appendix C.3. In Section C.2, we prove Theorem 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 (7) is true, the population-level optimal 2-means solution is unique and is given by:

$$\boldsymbol{\mu}^* = \left(\begin{bmatrix} \kappa \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} -\kappa \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right), \tag{26}$$

and when condition (9) is true then the population-level optimal 2-means solution is unique and is given by:

$$\boldsymbol{\mu}^* = \left(\begin{bmatrix} 0 \\ \sqrt{\frac{2}{\pi}}\sigma_2 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\sqrt{\frac{2}{\pi}}\sigma_2 \\ \vdots \\ 0 \end{bmatrix}, \right). \tag{27}$$

Now the reason that we have $\sqrt{\frac{2}{\pi}}\sigma_2$ in equation (27) is because $E[X_2|X_2 > 0] = \sqrt{\frac{2}{\pi}}\sigma_2$ and similarly we have κ in equation (26) because $E[X_1|X_1 > 0] = \kappa$, which is given by the following lemma:

Lemma 7. *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 seperating 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 8. *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 9. *For any fixed* $i \geq 2$ *and a separating hyperplane* $\mathcal{H}(b) = \{y \in \mathbb{R}^d : b^T y = 0\}$*, when*

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

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

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

Further more, to prove Theorem 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 10. *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:*

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

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

Lemma 11.

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

We include the proofs of all of these additional lemmas that help us prove Theorem 3 in Appendix C.4.

C.1 Proof of Theorem 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 $x \in \mathbb{R}^d$,

$$\phi(x, \mathbf{t}) = \min\{\|x - t_1\|^2, \|x - t_2\|^2\}, \tag{28}$$

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

$$\widetilde{\phi}(x, t^*) = \phi(x, (t^*, -t^*)) = \min\{\|x - t^*\|^2, \|x + t^*\|^2\}. \tag{29}$$

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 differentiables 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:

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

□

C.2 Proof of Theorem 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:

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

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 8, 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 8,

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

where $\alpha = P\left(Q = 0|b^T Y > 0\right) = 1 - \Phi\left(ab_1/2\sqrt{b^T D b}\right)$ as shown in equation 48, 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 10. 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 10, we get that the i^{th} coordinate of the projection of C_1 onto the separating hyperplane $\mathcal{H}(b)$ is given by:

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

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 $i = 1$,
$$(1-2\alpha)\left[\frac{a}{2} - \frac{ab_1^2\sigma_1^2}{2b^TDb}\right] + \frac{b_1\left(\sigma_1^2 - b^TDb\right)}{b^TDb}\mathbf{A} = 0, \tag{30}$$
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 $i \geq 2$,
$$-(1-2\alpha)\left[\frac{ab_1b_i\sigma_i^2}{2b^TDb}\right] + \frac{b_i\left(\sigma_i^2 - b^TDb\right)}{b^TDb}\mathbf{A} = 0. \tag{31}$$

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 (30) is greater than or equal to 0. Also $b_1^2\sigma_1^2 \leq b^TDb$, therefore the first expression in equation (30) is also greater than or equal to 0. Now for equation (30) 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 (30) 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 (31) holds for all $i \geq 2$. But if $b_1 = 0$, equation (31) 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 (30) and (31) 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 (31) for $i = d$. $b^TDb \geq \min_i \sigma_i^2 = \sigma_d^2$, therefore the second expression in equation (31) has the same sign as b_d. The first expression in equation (31) also has the same sign as b_d. Therefore, for their sum to be zero, i.e., for equation (31) 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 (31) 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 (31) simplifies to $b_i\left(\sigma_i^2 - b^TDb\right) = 0$ which holds if and only if $b_i = 0$ or $b^TDb = \sigma_i^2$. 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 (31) for $i = d-1$. Since $b_d = 0$, $b^TDb \geq \min_{1 \leq i \leq d-1} \sigma_i^2 = \sigma_{d-1}^2$, therefore the second expression in equation (31) has the same sign as b_{d-1}. The first

expression in equation (31) also has the same sign as b_{d-1}. Therefore, for equation (31) 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 (31) 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 (30) and equation (31) 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 (30) and equation (31) for $i = 2$ we need to look at $\mathbf{A}$. Now recall that if $Y \sim f$, then $b^T Y \sim N(-ab_1/2, b^T Db)$ and if $Y \sim g$, then $b^T Y \sim N(ab_1/2, b^T Db)$. Therefore,

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

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

Therefore,

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

Plugging this into the L.H.S. of equation (30) we get:

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

Similarly simplifying L.H.S. of equation (31) 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 (30) and equation (31) for $i = 2$ hold if and only if

$$
\left(2\;\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)-1\right)\frac{a}{2}-\frac{2b_1\left(\sigma_2^2-\sigma_1^2\right)}{\sqrt{b^TDb}}\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)=0 \tag{32}
$$

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 $\sigma_2^2 \le \sigma_1^2 + \frac{a^2}{4}$, the L.H.S. of equation (32) becomes

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

Therefore if $\sigma_2^2 \le \sigma_1^2 + \frac{a^2}{4}$, then equation (32) can not hold and hence either $b_1 = 0$ or $b_2 = 0$. Putting everything together, this implies equations (30) and (31) hold iff

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

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 $b^TDb = \sigma_i^2$ for all $i \ge 2$ and if $b_1 > 0$, we require $b_i = 0$ for all $i \ge 3$. Therefore equations (30) and (31) hold iff

$$
\begin{aligned}
&\{b_1 = 1 \;\text{ and }\; b_i = 0,\; i \ge 2\} \;\text{ OR }\; \left\{b_1 = 0 \;\text{ and }\; \left(b_i = 0 \;\text{ or }\; b^TDb = \sigma_i^2,\; i \ge 2\right)\right\}\\
&\text{OR }\; \left\{0 < b_1 < 1,\; b_1^2 + b_2^2 = 1,\; b_i = 0,\; i \ge 3,\;\; \text{and Eq 32 holds }\right\}.
\end{aligned}
$$

From cases 1 and 2 we finally come to the conclusion that for equations (30) and (31) to hold for $\sigma_2^2 \le \sigma_1^2 + \frac{a^2}{4}$, we need

$$
\{b_1 = 1 \;\text{ and }\; b_i = 0,\; i \ge 2\} \;\text{ OR }\; \left\{b_1 = 0 \;\text{ and }\; \left(b_i = 0 \;\text{ or }\; b^TDb = \sigma_i^2,\; i \ge 2\right)\right\}. \tag{33}
$$

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

$$\{b_1 = 1 \text{ and } b_i = 0,\ i \geq 2\} \text{ OR } \left\{b_1 = 0 \text{ and } \left(b_i = 0 \text{ or } b^T D b = \sigma_i^2,\ i \geq 2\right)\right\}$$
$$\text{OR } \left\{0 < b_1 < 1,\ b_1^2 + b_2^2 = 1,\ b_i = 0,\ i \geq 3, \text{ and Eq 32 holds }\right\}. \quad (34)$$

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 (33) to hold, for some unique $i = i_0$,

   $$b^T D b = \sigma_{i_0}^2,\ b_{i_0} = 1 \text{ and } b_j = 0 \text{ for } j \neq i_0.$$

   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 41 to find the corresponding within sum of squares as

   $$W_1^* := W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \left(\sqrt{\frac{2}{\pi}}\ \sigma_1\ e^{-\frac{a^2}{8\sigma_1^2}} + \frac{a}{2}\ P\left(|Z| < \frac{a}{2\sigma_1}\right)\right)^2. \quad (35)$$

   For $i_0 = 2$, that is, when $b_2 = 1$ and $b_i = 0$ for $i \neq 2$ we can similarly use Lemma 41 to find the corresponding within sum of squares as

   $$W_2^* := W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_2^2. \quad (36)$$

   From Lemma 9 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

   $$W_{i_0}^* := W(b) = \sum_{j=1}^{d} \sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_{i_0}^2. \quad (37)$$

   Now using the lower bound given by Lemma 11 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 hyperlane 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 9, 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 (34), 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 (35), is given by W_1^*.

   Following similar reasonings as presented in Case 1 for the second possibility (34), 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 (36)) and for $i_0 \geq 2$, it is $W_{i_0}^*$ (Equation (37)).

   To study the third possibility in 34, we use Lemma 8 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 (32) 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 (32) 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 (32).

$$\begin{aligned}
&\left[\left(2\Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) - 1\right)\frac{a}{2} + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 + \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\\
&= \left[\frac{2b_1\left(\sigma_2^2 - \sigma_1^2\right)}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) + \frac{2b_1\sigma_1^2}{\sqrt{b^T Db}}\phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\right]^2 + \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\\
&= \frac{4b_1^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) + \frac{4b_2^2\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right)\\
&= \frac{4\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right).
\end{aligned}$$

Again by using Equation (32) 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 (32),

$$\begin{aligned}
\frac{4\sigma_2^4}{b^T Db}\phi^2\left(\frac{ab_1}{2\sqrt{b^T Db}}\right) &\leq 4\sigma_2^4\ \frac{1}{2\pi}\ \frac{a^2 b_1^2}{b^T Db}\ \frac{a^2}{4}\ \frac{1}{4b_1^2(\sigma_2^2 - \sigma_1^2)^2}\\
&= \frac{2}{\pi}\sigma_2^2\ \frac{\sigma_2^2}{b^T Db}\ \left(\frac{a^2/4}{\sigma_2^2 - \sigma_1^2}\right)^2\\
&\leq \frac{2}{\pi}\sigma_2^2\ \frac{\sigma_2^2}{\sigma_1^2}\ \left(\frac{a^2/4}{\sigma_2^2 - \sigma_1^2}\right)^2,
\end{aligned}$$

since $\sigma_1^2 < \sigma_2^2$ and therefore $\sigma_1^2 \leq b^T Db$. 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:

$$\sigma_1^2(x - \sigma_1^2)^2 - x\ \frac{a^4}{16} = 0 \iff \sigma_1^2 x^2 - \left(2\sigma_1^4 + \frac{a^4}{16}\right)x + \sigma_1^6 = 0 \tag{38}$$

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

$$\frac{4\sigma_2^4}{b^T D b}\phi^2\left(\frac{ab_1}{2\sqrt{b^T D b}}\right) \leq \frac{2}{\pi}\sigma_2^2.$$

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 (32),

$$W(b) \geq \sum_{i=1}^{d}\sigma_i^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_2^2 = W_2^*.$$

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 (34), 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 $\mathcal{H}(b) = \{y \in \mathbb{R}^d : y_2 = 0\}$.

□

C.3 Proof of Lemma 2

This proof is analogous to the proof of the matrix being positive definite for the null case as shown in Lemma 4. We find the matrix G in the two different cases and show that it is positive definite.

1. When condition (7) is true, Theorem 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 4, we know that for $x \in M_{12}$,

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

$$(x-\mu_1^*)(x-\mu_2^*)^T = \begin{pmatrix} -\mu_{11}^{*2} & -\mu_{11}^* x_2 & \dots & -\mu_{11}^* x_d \\ -\mu_{21}^* x_2 & x_2^2 & \dots & x_2 x_d \\ \vdots & \vdots & \ddots & \vdots \\ -\mu_{21}^* x_d & x_2 x_d & \dots & x_d^2 \end{pmatrix}.$$

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 $i, j \in \{2, \dots, d\}$,

$$\int_{M_{12}} x_i \, x_j \, f(x)\, d\sigma(x) = \mathbb{E}[X_i] \, \mathbb{E}[X_j] \, \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} = 0.$$

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:

$$m_1 = \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1} \, \mathbb{E}[X_1 | X_1 > 0], \; m_j = \frac{\sigma_j^2}{\mathbb{E}[X_1 | X_1 > 0]} \, \frac{e^{-a^2/8\sigma_1^2}}{\sqrt{2\pi}\sigma_1}, \; j \neq 1. \tag{39}$$

Then the diagonal entries of G_{11} are given by $1 - m_j$, for $j = 1, \ldots, d$. 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 $\mathbb{E}[X_1|X_1 > 0]$.

$$\mathbb{E}[X_1|X_1 > 0] \leq \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_1}\left(\sigma_1^2 + \frac{a^2}{4}\right), \quad \mathbb{E}[X_1|X_1 > 0] \geq \sqrt{\frac{2}{\pi}} \frac{e^{-a^2/8\sigma_1^2}}{\sigma_1}\left(\sigma_1^2 + \frac{a^2}{4}\right) \quad (40)$$

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 $j = 1, \ldots, d$. 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 (9) is true, Theorem 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 $x \in M_{12}$,

$$(x - \mu_1^*)(x - \mu_1^*)^T = \begin{pmatrix} x_1^2 & -\mu_{11}^* x_1 & x_1 x_3 & \ldots & -\mu_{11}^* x_d \\ -\mu_{11}^* x_1 & \mu_{11}^{*2} & -\mu_{11}^* x_3 & \ldots & -\mu_{11}^* x_d \\ x_1 x_3 & -\mu_{11}^* x_3 & x_3^2 & \ldots & x_3 x_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1 x_d & -\mu_{11}^* x_d & x_3 x_d & \ldots & x_d^2 \end{pmatrix},$$

$$(x - \mu_1^*)(x - \mu_2^*)^T = \begin{pmatrix} x_1^2 & -\mu_{21}^* x_1 & x_1 x_3 & \ldots & -\mu_{11}^* x_d \\ -\mu_{11}^* x_1 & -\mu_{11}^{*2} & -\mu_{11}^* x_3 & \ldots & -\mu_{11}^* x_d \\ x_1 x_3 & -\mu_{21}^* x_3 & x_3^2 & \ldots & x_3 x_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1 x_d & -\mu_{21}^* x_d & x_3 x_d & \ldots & x_d^2 \end{pmatrix}.$$

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 $j \neq 2$,

$$\mu_{11}^* \int_{M_{12}} x_j \; f(x) \; d\sigma(x) = \mu_{11}^* \; \mathbb{E}[X_j] \; \frac{1}{\sqrt{2\pi}\sigma_2} = 0,$$
$$\mu_{21}^* \int_{M_{12}} x_j \; f(x) \; d\sigma(x) = \mu_{21}^* \; \mathbb{E}[X_j] \; \frac{1}{\sqrt{2\pi}\sigma_2} = 0,$$
$$\int_{M_{12}} x_j^2 \; f(x) \; d\sigma(x) = \mathbb{E}[X_j^2] \; \frac{1}{\sqrt{2\pi}\sigma_2}.$$

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 (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.

□

C.4 Proofs of Additional Lemmas Supporting Theorem 3

C.4.1 Proof of Lemma 7

Let Z be a random variable generated from $N(0,1)$. Then,

$$\begin{aligned}
\mathbb{E}\left[Y|Y>0\right] &= 2\mathbb{E}[Y\mathbb{I}_{\{Y>0\}}] \\
&= \int_0^\infty x f(x)dx + \int_0^\infty x g(x)dx \\
&= \frac{1}{\sqrt{2\pi}\sigma_1}\left[\int_0^\infty x e^{-\frac{1}{2\sigma_1^2}(x+a/2)^2}dx + \int_0^\infty x e^{-\frac{1}{2\sigma_1^2}(x-a/2)^2}dx\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_1}\left[\int_0^\infty \left(x+\frac{a}{2}\right) e^{-\frac{1}{2\sigma_1^2}(x+a/2)^2}dx + \int_0^\infty \left(x-\frac{a}{2}\right) e^{-\frac{1}{2\sigma_1^2}(x-a/2)^2}dx\right] + \frac{a}{2}\mathbb{P}\left(|Z|<\frac{a}{2\sigma_1}\right) \\
&= \frac{\sigma_1}{\sqrt{2\pi}}\left[e^{-\frac{1}{2\sigma_1^2}(x+a/2)^2}\Big|_\infty^0 + e^{-\frac{1}{2\sigma_1^2}(x-a/2)^2}\Big|_\infty^0\right] + \frac{a}{2}\mathbb{P}\left(|Z|<\frac{a}{2\sigma_1}\right) \\
&= \sqrt{\frac{2}{\pi}}\,\sigma_1\, e^{-\frac{a^2}{8\sigma_1^2}} + \frac{a}{2}\,\mathbb{P}\left(|Z|<\frac{a}{2\sigma_1}\right).
\end{aligned}$$

□

C.4.2 Proof of Lemma 8

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,

$$W(b) = E\left[\,(X_1-E[X_1|b^TX>0])^2\big|\,b^TX>0\right]+E\left[\,(X_2-E[X_2|b^TX>0])^2\big|\,b^TX>0\right]+\sum_{j=3}^d \sigma_j^2. \tag{41}$$

We know that,

$$E\left[\,(X_1-E[X_1|b^TX>0])^2\big|\,b^TX>0\right] = E\left[\,X_1^2\big|\,b^TX>0\right] - \left(E[X_1|b^TX>0]\right)^2,$$

and similarly,

$$E\left[\,(X_2-E[X_2|b^TX>0])^2\big|\,b^TX>0\right] = E\left[\,X_2^2\big|\,b^TX>0\right] - \left(E[X_2|b^TX>0]\right)^2.$$

Since $-X \stackrel{d}{=} X$, we can write,

$$\begin{aligned} E\left[X_1^2\right] &= P\left(b^T X > 0\right) E\left[X_1^2 \middle| b^T X > 0\right] + P\left(b^T X < 0\right) E\left[X_1^2 \middle| b^T X < 0\right] \\ &= P\left(b^T X > 0\right) E\left[X_1^2 \middle| b^T X > 0\right] + P\left(b^T X < 0\right) E\left[X_1^2 \middle| b^T X > 0\right] \\ &= E\left[X_1^2 \middle| b^T X > 0\right]. \end{aligned}$$

Hence,

$$E\left[X_1^2 \middle| b^T X > 0\right] = E\left[X_1^2\right] = \sigma_1^2 + \frac{a^2}{4}, \tag{42}$$

and following similar arguments,

$$E\left[X_2^2 \middle| b^T X > 0\right] = E\left[X_2^2\right] = \sigma_2^2. \tag{43}$$

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,

$$\begin{aligned} E\left[X_1 \middle| b^T X > 0\right] &= E\left[Y_1 \middle| b^T Y > 0\right] \\ &= E\left[E[Y_1 | b^T Y > 0, Q] \middle| b^T Y > 0\right] \\ &= P\left(Q = 0 | b^T Y > 0\right)\ E_f[Y_1 | b^T Y > 0, Q = 0] \\ &\quad + P\left(Q = 1 | b^T Y > 0\right)\ E_g[Y_1 | b^T Y > 0, Q = 1] \\ &= \alpha\ E_f[Y_1 | b^T Y > 0] + (1 - \alpha)\ E_g[Y_1 | b^T Y > 0], \end{aligned}$$

where $\alpha = P\left(Q = 0 | b^T Y > 0\right)$, 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,

$$E\left[X_2 \middle| b^T X > 0\right] = \alpha\ E_f[Y_2 | b^T Y > 0] + (1 - \alpha)\ E_g[Y_2 | b^T Y > 0]$$

To simplify further, we consider each of these terms separately. We first start with $E_f[Y_1 | b^T Y > 0]$ and to compute it we define random variable V such that when $Y \sim f$, that is, $Y \sim N(-\theta_1, D)$,

$$V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} := \begin{pmatrix} \frac{Y_1 + a/2}{\sigma_1} \\ \frac{b_1 Y_1 + b_2 Y_2 + a b_1/2}{\sqrt{b^T D b}} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \frac{b_1 \sigma_1}{\sqrt{b^T D b}} \\ \frac{b_1 \sigma_1}{\sqrt{b^T D b}} & 1 \end{pmatrix} \right).$$

Note that $b^T Y = b_1 Y_1 + b_2 Y_2 > 0$ is equivalent to $V_2 > \frac{a b_1}{2\sqrt{b^T D b}}$. In order to find $E\left[V_1 \middle| V_2 > \frac{a b_1}{2\sqrt{b^T D b}}\right]$ we use moments derived for truncated bivariate normal distribution, presented in Rosenbaum (1961). We get

$$\begin{aligned} E\left[V_1 \middle| V_2 > \frac{a b_1}{2\sqrt{b^T D b}}\right] &= \frac{b_1 \sigma_1}{\sqrt{b^T D b}} \left(\frac{\phi\left(\frac{a b_1}{2\sqrt{b^T D b}}\right)}{P\left(V_2 > \frac{a b_1}{2\sqrt{b^T D b}}\right)} \right) \\ &= \frac{b_1 \sigma_1}{\sqrt{b^T D b}} \left(\frac{\phi\left(\frac{a b_1}{2\sqrt{b^T D b}}\right)}{1 - \Phi\left(\frac{a b_1}{2\sqrt{b^T D b}}\right)} \right). \end{aligned}$$

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,

$$E_f\left[Y_1|\, b^TY>0\right]=-\frac{a}{2}+\frac{b_1\sigma_1^2}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{1-\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right). \tag{44}$$

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,

$$E\left[V_1^*\left|V_2>\frac{ab_1}{2\sqrt{b^TDb}}\right.\right]=E_f\left[\frac{Y_2}{\sigma_2}\middle|\, b^TY>0\right].$$

Therefore,

$$E_f\left[Y_2|\, b^TY>0\right]=\frac{b_2\sigma_2^2}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{1-\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right). \tag{45}$$

In order to find $E_g\left[Y_1|\, b^TY>0\right]$ and $E_g\left[Y_2|\, b^TY>0\right]$, 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 (44) and (45) will give us the corresponding expectations when $Y\sim g$. So we get

$$E_g\left[Y_1|\, b^TY>0\right]=\frac{a}{2}+\frac{b_1\sigma_1^2}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right) \tag{46}$$

$$E_g\left[Y_2|\, b^TY>0\right]=\frac{b_2\sigma_2^2}{\sqrt{b^TDb}}\left(\frac{\phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}{\Phi\left(\frac{ab_1}{2\sqrt{b^TDb}}\right)}\right) \tag{47}$$

To find α we note that,

$$\begin{aligned}\alpha=P\left(Q=0|b^TY>0\right)&=\frac{P\left(Q=0,b^TY>0\right)}{P\left(Q=0,b^TY>0\right)+P\left(Q=1,b^TY>0\right)}\\&=\frac{P_f\left(b^TY>0\right)}{P_f\left(b^TY>0\right)+P_g\left(b^TY>0\right)}.\end{aligned}$$

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,

$$\alpha = P_f\left(b^T Y > 0\right) = P_f\left(\frac{b^T Y + ab_1/2}{\sqrt{b^T Db}} > \frac{ab_1/2}{\sqrt{b^T Db}}\right) = 1 - \Phi\left(\frac{ab_1}{2\sqrt{b^T Db}}\right). \tag{48}$$

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 (42) and (43) into equation (41), we get that:

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

□

C.4.3 Proof of Lemma 9

The within sum of squares can be written as:

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

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,

$$\begin{aligned} W(b) &= Var(X_i|X_i > 0) + E[X_1^2] + \sum_{j \neq i, j \neq 1} E[X_j^2] \\ &= \sigma_i^2 - \frac{2}{\pi}\sigma_i^2 + \sigma_1^2 + \frac{a^2}{4} + \sum_{j \neq i, j \neq 1}\sigma_j^2 \\ &= \sum_{j=1}^{d}\sigma_j^2 + \frac{a^2}{4} - \frac{2}{\pi}\sigma_i^2. \end{aligned}$$

□

C.4.4 Proof of Lemma 10

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:

$$\mathcal{P}_i = \frac{a}{2} \mathbb{I}\{i = 1\} - \frac{a b_1 b_i}{2} + \frac{b_i Var(Y_i) - b_i \left(b^T D b\right)}{b^T D b} \left(E\left[Z_2 | \, Z_2 > 0\right] - \frac{a b_1}{2}\right).$$

□

C.4.5 Proof of Lemma 11

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 $u \geq 0$:

$$\exp(-u^2/2) \geq 1 - \frac{u^2}{2} + \frac{u^4}{8} - \frac{u^6}{48} + \frac{u^8}{384} - \frac{u^{10}}{3840}$$
$$\mathbb{P}(|Z| \leq u) \geq \sqrt{\frac{2}{\pi}}\left[u - \frac{u^3}{6} + \frac{u^5}{40} - \frac{u^7}{336} + \frac{u^9}{3456} - \frac{u^{11}}{42240}\right],$$

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.

□

D Proof of Theorem 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:

$$\{X_1, \ldots, X_n\} \sim \frac{1}{2}N(-\theta_1, D) + \frac{1}{2}N(\theta_1, D) \tag{49}$$

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 (7),

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

holds, SigClust is asymptotically consistent, and when condition (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 gave a characterization of the population-level optimal symmetric 2-means solution in this model. Under the model in (49) described above, the population-level optimal 2-means solution is unique and is given by (26) and (27).

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:

$$D_0 = \begin{bmatrix} \sigma_1^2 + \frac{a^2}{4} & 0 & 0 \ldots & 0 & \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & \sigma_d^2 \end{bmatrix}.$$

and use $\mathbb{P}_1$ to denote the distribution in (49). 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 12. *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 D.1, but note that it follows from straightforward calculations based on Lemma 1 and Theorem 2. As a consequence of this result, we have the following characterization of SigClust:

Lemma 13. *Suppose that for some constant $C > 0$,*

$$\frac{\tau_0}{\tau_1} \leq C \quad \text{and,} \tag{50}$$

$$\sqrt{n}\frac{W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)}{\tau_1} \to \infty, \quad \text{as } n \to \infty, \tag{51}$$

then SigClust is asymptotically consistent. On the other hand if,

$$\frac{\tau_0}{\tau_1} \leq C \quad \text{and,} \tag{52}$$

$$W_1(\boldsymbol{\mu}_1) = W_0(\boldsymbol{\mu}_0) \tag{53}$$

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 (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 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 (5). This is easily provided by an analogous version of Theorem 2 for a single Normal distribution as follows:

Lemma 14. *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) and (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 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 4.

So given the expressions for τ_0 and $W_0(\boldsymbol{\mu}_0)$ in (3) and (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 Thorem 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 15. *There are universal constants $0 < c \leq C$ such that:*

1. *If (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 (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 D.2. To complete the proof of the Theorem we need to put together Lemmas 13 and 15 to show the consistency and inconsistency of SigClust in different regimes.

We note that using (55) and the result of Lemma 15 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 (50) and (52). Thus, Lemmas 13 and 15 directly yield the inconsistency of SigClust when condition (9) holds. On the other hand, in order to establish consistency when (7) holds, to verify Condition (51) we note that for some constant $c > 0$,

$$\begin{aligned}\sqrt{n}\frac{W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)}{\tau_1} &\geq c\sqrt{n}\,[W_0(\boldsymbol{\mu}_0) - W_1(\boldsymbol{\mu}_1)] \\ &= c\sqrt{n}\underbrace{\left[\kappa^2 - \frac{2}{\pi}\max\left\{\left(\sigma_1^2 + \frac{a^2}{4}\right), \sigma_2^2\right\}\right]}_{T},\end{aligned}$$

so to complete the proof of the Theorem it suffices to lower bound the term $T > 0$ as $n \to \infty$, when (7) holds. Clearly in this regime $\kappa^2 - 2\sigma_2^2/\pi > 0$ so Lemma 11 as stated in Appendix C.2, completes the proof of our Theorem.

D.1 Proof of Lemma 12

Under the null, the distribution of the statistic follows from Theorem 1 and Lemma 14. Concretely, for

$$W_0(\boldsymbol{\mu}_0) = \widetilde{\sigma}^2 - \frac{2}{\pi}\max\left\{\left(\sigma_1^2 + \frac{a^2}{4}\right), \sigma_2^2\right\}, \tag{54}$$

$$\tau_0^2 = 2\sum_{i=2}^{d}\sigma_i^4 + 2\left(\sigma_1^2 + \frac{a^2}{4}\right)^2 - \frac{16}{\pi^2}\left[\max\left\{\left(\sigma_1^2 + \frac{a^2}{4}\right), \sigma_2^2\right\}\right]^2, \tag{55}$$

we have by a combination of Theorem 1 and Lemma 14 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 2 to obtain that,

$$\sqrt{n}\Big(T_n^{(0)} - \frac{W_1(\boldsymbol{\mu}_1)}{\widetilde{\sigma}^2}\Big) \rightsquigarrow N\Big(0, \Big[\frac{\tau_1}{\widetilde{\sigma}^2}\Big]^2\Big), \tag{56}$$

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 (49) 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 (56).

D.2 Proof of Lemma 15

We divide our analysis into two cases, according to the optimal 2-means solution.

When condition (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 (8), and it only remains to bound τ_1^2.

To lower bound τ_1^2 we note that,

$$\begin{aligned}
\tau_1^2 &= \sum_{i=1}^{2} \mathbb{P}_1(A_i)\mathbb{E}_{X\sim\mathbb{P}_1}[\|X - \mathbb{E}[X|X \in A_i]\|^4 | X \in A_i] - [W_1(\boldsymbol{\mu}_1)]^2 \\
&= 3\sum_{j=2}^{d} \sigma_j^4 + \sum_{i,j\neq 1, j\neq i} \sigma_i^2\sigma_j^2 + \mathbb{E}[(X_1-\kappa)^2|X_1 \geq 0]\sum_{j\neq 1}\sigma_j^2 \\
&\quad + \mathbb{E}[(X_1-\kappa)^4|X_1\geq 0] - \left[\mathbb{E}[(X_1-\kappa)^2|X_1\geq 0] + \sum_{j=2}^{d}\sigma_j^2\right]^2 \\
&= 2\sum_{j=2}^{d}\sigma_j^4 + \operatorname{var}\left((X_1-\kappa)^2|X_1\geq 0\right).
\end{aligned}$$

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 (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 (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.

E Proof of our main results for RIFT

In this Appendix, we collect the proofs of the main results for the RIFTs in our paper. In Sections E.1 and E.3 we consider the limiting distributions of the RIFT statistic, and its ℓ_2 counterpart under the null and prove Theorems 5 and 7. In Section E.2 we consider Thorem 6 and analyze the power of the RIFT and finally in Section E.4 we consider Theorem 8 where we verify the validity of the modified RIFT to test for mixtures of two Normals.

E.1 Proof of Theorem 5

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,

$$\rho = \mathbb{E}\left[\left|\widetilde{R}_i - \Gamma\right|^3\right] \quad \text{and} \quad \tau^2 = \mathbb{E}\left[\left(\widetilde{R}_i - \Gamma\right)^2\right]. \tag{57}$$

Now $\Gamma = \mathbb{E}\left[\widetilde{R}_i\right]$, therefore,

$$\tau^2 = \text{Var}\left(\widetilde{R}_i\right) = \text{Var}\left(R_i + \delta Z_i\right) \geq \delta^2 \text{Var}(Z_i) = \delta^2 > 0.$$

Now note that,

$$\begin{aligned}
|R_i| &= \left|\log\left(\frac{\widehat{p}_2(X_i)}{\widehat{p}_1(X_i)}\right)\right| = |\log \widehat{p}_2(X_i) - \log \widehat{p}_1(X_i)| \\
&\leq \max_x \left\{\log \widehat{p}_2(x) - \log \widehat{p}_1(x), \log \widehat{p}_1(x) - \log \widehat{p}_2(x)\right\} \\
&\leq \max_{x, i \in \{1,2\}} \left\{\log \widehat{f}_i(x) - \log \widehat{p}_1(x), \log \widehat{p}_1(x) - \log \widehat{f}_i(x)\right\}.
\end{aligned}$$

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

$$\begin{aligned}
\widehat{\mu}^T \widehat{\Sigma}^{-1} \widehat{\mu} - \left(\widehat{\Sigma}^{-1}\widehat{\mu} - \widehat{\Sigma}_i^{-1}\widehat{\mu}_i\right)^T \left(\widehat{\Sigma}^{-1} - \widehat{\Sigma}_i^{-1}\right)^{-1} \left(\widehat{\Sigma}^{-1}\widehat{\mu} - \widehat{\Sigma}_i^{-1}\widehat{\mu}_i\right) \leq k, \\
\widehat{\mu}_i^T \widehat{\Sigma}_i^{-1} \widehat{\mu}_i - \left(\widehat{\Sigma}_i^{-1}\widehat{\mu}_i - \widehat{\Sigma}^{-1}\widehat{\mu}\right)^T \left(\widehat{\Sigma}_i^{-1} - \widehat{\Sigma}^{-1}\right)^{-1} \left(\widehat{\Sigma}_i^{-1}\widehat{\mu}_i - \widehat{\Sigma}^{-1}\widehat{\mu}\right) \leq k.
\end{aligned}$$

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. □

E.2 Proof of Theorem 6

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

E.3 Proof of Theorem 7

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

$$a^2 = \text{Var}\left(\widetilde{U}_i\right). \tag{58}$$

Now $\Theta = \mathbb{E}[U_i] = \mathbb{E}\left[\widetilde{U}_i\right]$, therefore,

$$a^2 = \text{Var}\left(\widetilde{U}_i\right) = \text{Var}\left(U_i + \delta Z_i\right) \geq \delta^2 \text{Var}(Z_i) = \delta^2.$$

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 $\widetilde{C} = \frac{C_0}{\delta^3}\left[8C_2^3 + \delta\left(12C_2^2\sqrt{\frac{2}{\pi}} + 6C_2\delta + 2\sqrt{\frac{2}{\pi}}\delta^2\right)\right]$. □

E.4 Proof of Theorem 8

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.

Recall that when we fit H_1, we constrain the solution to satisfy $K(p, \widehat{p}_2) > \Delta$. Under H_0, $\mathbb{P}(A_n) \to 1$ where A_n is the event:

$$A_n = \left\{K(p, \widehat{p}_1) < \Delta + \frac{\tau \delta_n}{\sqrt{n}}\right\}$$

and δ_n is any sequence such that $\delta_n = o(1)$. (In fact, Δ can also be taken to be a slowly decreasing sequence.) On the event A_n we have that

$$\begin{aligned}
z_\alpha - \frac{\sqrt{n}\Gamma}{\tau} &= z_\alpha - \frac{\sqrt{n}(K(p,\widehat{p}_1) - K(p,\widehat{p}_2))}{\tau} \\
&\geq z_\alpha - \frac{\sqrt{n}(K(p,\widehat{p}_1) - \Delta_n)}{\tau} \geq z_\alpha - \delta_n.
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}\left[\overline{\Phi}\left(z_\alpha - \frac{\sqrt{n}\Gamma}{\tau}\right)\right] &= \mathbb{E}\left[\overline{\Phi}\left(z_\alpha - \frac{\sqrt{n}\Gamma}{\tau}\right)\mathbb{I}_{A_n}\right] + \mathbb{E}\left[\overline{\Phi}\left(z_\alpha - \frac{\sqrt{n}\Gamma}{\tau}\right)\mathbb{I}_{A_n^c}\right] \\
&\leq \mathbb{E}\left[\overline{\Phi}\left(z_\alpha - \delta_n\right)\mathbb{I}_{A_n}\right] + \mathbb{P}(A_n^c) \\
&= \alpha + o(1).
\end{aligned}$$

□