Assessing the Statistical Significance of Clusters Using SigClust

Makena Grigsby and Sharvee Joshi

Contents

Abstract

The goal of this project is to study when a two–cluster structure in high–dimensional data is statistically meaningful versus simply an artifact of running a clustering algorithm. We focus on the SigClust procedure of Liu et al. and the soft–thresholding covariance estimation proposed by Huang et al. (2015), which tests H_0 : one Gaussian cluster against H_1 : more than one cluster. We replicate a subset of the simulation settings in Huang et al. using high–dimensional Gaussian data with a spiked covariance structure. First, we generate data from a single Gaussian distribution, apply k-means with $k = 2$, and use SigClust to test whether the resulting partition corresponds to real clusters. We then focus on implementing the test, examining the behavior under more complex scenarios dealing with breast cancer data, similar to the data used in the paper. Specifically, we simulate high-dimensional Gaussian data created to mimic the data found in the paper. Our findings align with those in Huang et al. (2015) and complement the BRCA cancer subtype analysis in the methodology section. SigClust correctly rejects under strong separation and struggles under a weak dense signal.

1 Introduction

In high-dimensional settings, such as genomics, visual clustering tends to be misleading. Datasets that are drawn from a single Gaussian distribution often appear to multiple subgroups, as seen in past homework assignments. The paper our team chose for this project focuses on using SigClust, which is a hypothesis test designed to determine whether an observed two-cluster partition reflects true structure or arises from sampling variability. In this project, we aim to explore how SigClust behaves under controlled simulations, as well as reproduce the simulations found in the paper. These simulations mimic the high-dimension, low sample-size regime.

Clustering, as mentioned previously, is a central tool for statisticians that is used for high-dimensional data analysis. It is commonly used in situations such as cancer genomics, where researchers use expression profiles to discover potential diseases subtypes. However, high dimensionality creates the illusion of structure; even when data comes from a single Gaussian distribution, visualization such as PCA plots can appear to be clustered. This acts as the main motivator for the need of a formal statistical test to assess whether two apparent clusters represent genuine separation. SigClust addresses this problem by testing:

$$H_0 : \text{ The data come from a single Gaussian distribution}$$

versus

$$H_1 : \text{ The data consist of two or more distinct clusters.}$$

The test is built around the **cluster index**, which quantifies how well the data can be divided into two groups based on within-versus between-cluster variation. Small cluster indices indicate a strong separation, where as large cluster indices indicate clusters coming from a single Gaussian distribution. SigClust aims to estimate the null distribution of this index and produces a p-value which will indicate whether the observed separation is statistically meaningful. Note, the detailed statistical formulation of the test, which includes the covariance estimation, eigenvalue shrinkage, and theoretical considerations, will be presented in the methodology section of this report.

Our focus for this project is to:

1. Introduce the conceptual framework behind SigClust.
2. Run simulations that illustrate how the test behaves under simple clustering scenarios.
3. Highlight when visual intuition in high dimensions align or conflict with formal statistical testing.

To further relate back to the paper, we introduce and attempt the more advanced BRCA cancer subtype analysis later in the report. To add to the authenticity of replication, we mimicked the same computer settings as the original authors and use a computer with RAM 16GB.

2 Methodology

2.1 Statistical setup

2.2 The `sigclust()` Function

When working with high-dimensional simulations, it is important to first build intuition for what the function `sigclust()` is designed to detect. At its core, SigClust evaluates whether an observed two cluster partition reflects real separation, or whether the data could have plausibly come from one Gaussian distributions.

To visualize this idea, we begin with a simple two dimensional toy example, as shown in Figure 1. We generate two clusters that differ only by a shift in the x-coordinate, with signal strength controlled by parameter a. When $a = 0$, the two groups are completely overlapping and represent the true null setting, as in that they are samples from the same Gaussian distribution. As a increases, the separation gradually becomes more visual.

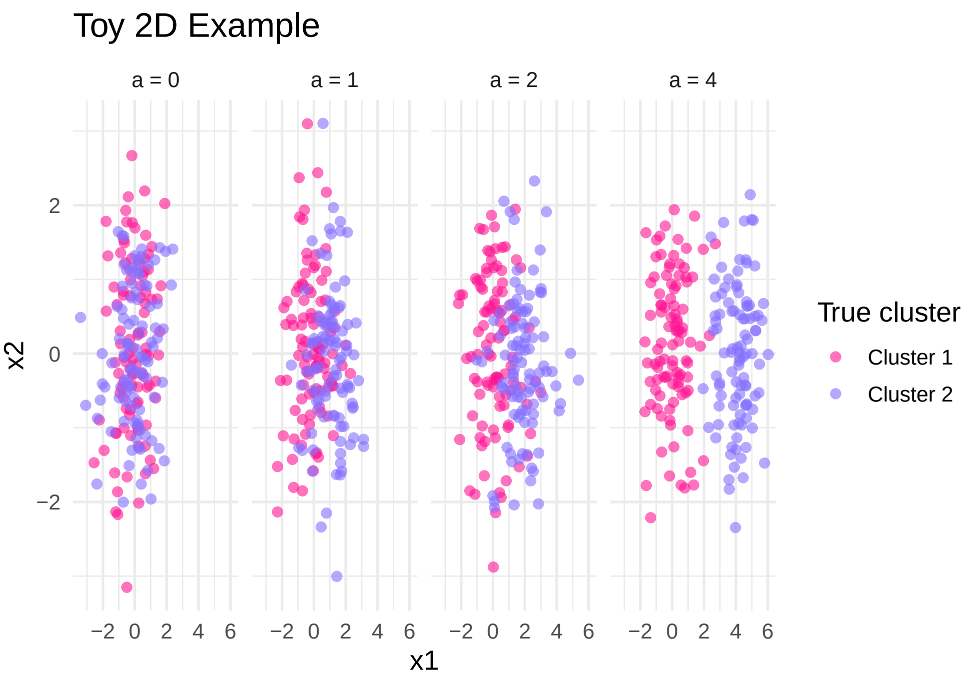

Figure 1: Toy 2D simulation illustrating increasing mean separation between two Gaussian subpopulations.

3 Implementation in Simulation Scenarios

To explore the SigClust hevaior in a controlled setting, we implemented simple simulations to generate high dimensional Gaussian data and evaluate three scenarios. We used the `sigclust()` function, something the authors created for this paper. We will explain the functionality of it below.

At its core, `sigclust()` implements the testing procedure described in Huang et al. (2015) and documented in the package reference manual. SigClust works by computing the clustering index, defied as the ratio of within cluster variation to total variation, and uses an embedded k-means clustering step. Specifically, when `labflag = 0`, SigClust will internally apply 2-means clustering to find the partition that minimized within-cluster sum of squares. When `labflag = 1`, the user supplies the labels but the same index is computed.

The observed cluster index is then compared to a simulated null distribution that is generated by repeatedly drawing samples from a single multivariate Gaussian model that is based on an estimated covariance, reclustering each simulated dataset with 2-means and recomputing the cluster index. This is essentially created a Monte Carlo p value, which is used to show the observed separation significance. All simulations share the same high-dimensional covariance structure. Following the motivation in Huang et al. (2015), we generate a *spiked* covariance matrix to emulate the strong eigenvalue structure observed in genomic datasets. Below, we detail how each scenario uses the `sigclust()` function.

3.1 Null: One Gaussian Distribution

Under the null hypothesis, the data comes from a single multivariate Gaussian distribution with a spiked covariance structure. We create a diagnonal covariance with ten very laruge eigenvalues, with variance = 50, and 990 noise dimensions, where variance = 1. This structure makes sure that the cluster behavior is non-trivial and is highly influenced by high dimensional noise. We also create a helper function that generates Gaussian samples under a specified mean vector. SigClust is applied using the `sigclust()` function with `labflag = 0`. The figure is one of the plot outputs of the function that is the most relevant to the discussion.

Figure 2 shows the SigClust diagnostic output when the data truly comes from one multivariate Gaussian

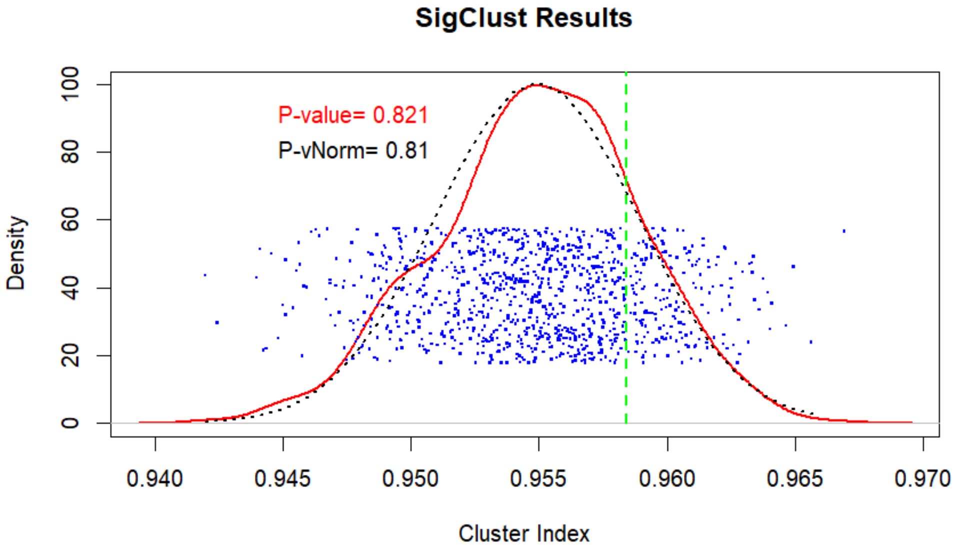

Figure 2: SigClust diagnostic plot for Null Hypothesis

distribution, or the null hypothesis. We notice there is a high p-value of approximately 0.82, which is far above the conventional significance threshold. This indicates that the observed cluster index is completely consistent with one expects when there is no real cluster structure.

The green vertical line sits inside the peak of the red density curve, suggesting that the clustering strength in our simulated data is similar to random fluctuations that could be caused by the Gaussian noise, and does not show a true two cluster separation.

The blue points represent the Monte Carlo simulated cluster indices from 1000 null Gaussian datasets. These form a dense and symmetric cloud around approximately 0.955, which is where the observed index lies if one cluster exists.

In summary, this shows that there is no evidence of two clusters, the observed cluster index looks like noise, and the SigClust test bheaves properly under the null and produces a large p-value.

3.2 Alternative One: One-Direction Strong Shift

For this simulation, we generate a very clear, two cluster signal and simulate from two Gaussian groups that differ only in the first coordinate with mean vectors: $\mu_1 = (0, 0, ..., 0)$ and $\mu_2 = (30, 0, ..., 0)$. The covariance structure is identical to the null case, but because the shift occurs in one informative direction, we are essentially creating a strong, sparse, signal that matches the one-direction signal referenced in the paper. We once again apply the same settings in the null case, making sure that two means clustering is implemented. The figure below is the final plot from the function.

We now observe that the cluster index, the green dashed line, is far to the left of the null density curve. This means that the within cluster variation is much smaller than expected under a single Gaussian model. This indicates that the estimated clusters are tighter and more separated than what would occur by chance. This behavior is further confirmed with a low p-value of approximately 0.032. Because this is below 0.05, this means SigClust rejects the null hypothesis of one Gaussian cluster.

The Monte Carlo simulated cluster indices (blue points) form a dense cloud around approximately 0.90, reflecting the null distribution. The observed index, however, lies far outside this cloud, demonstrating that the data are highly inconsistent with the one-cluster Gaussian model. This matches what we were expecting, a large directional shift results in two groups that are well separated.

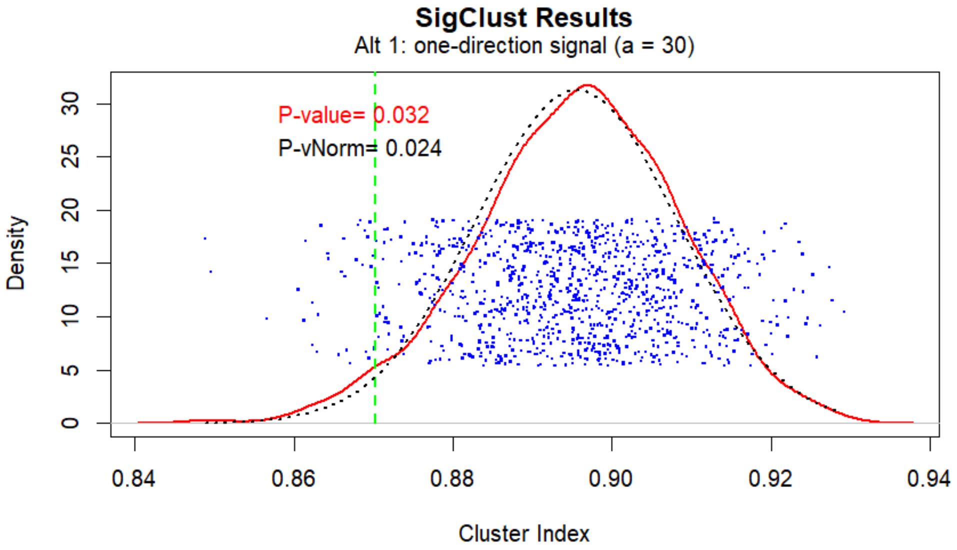

Figure 3: SigClust diagnostic plot for Alternative 1 (strong one-direction signal, a=30).

3.3 Alternative Two: Dense, Weak Shift

For the final simulation, we now create two groups with a small shift applied to every coordinate as opposed to a large shift in a single direction. We use the same covariate structure, but each coordinate of the second cluster is shifted by $a = 0.5$, which creates a dense, but very weak signal. This setup mimics situations where clusters differ in many genes, but each gene differs very slightly (a common pattern in high dimensional biological data). We once again apply the same settings in the null and alternate 1 case, making sure that two means clustering is implemented. The figure below is the final plot from the function.

Figure 4 shows the SigClust diagnostic output under the weak and diffuse signal. Unlike the first alternative, the observed cluster index lies a lot closer to the center of the null reference distribution. The green dashed line is only slightly on the left of the peak of the red density curve, which indicates that the separation created by the shift is extremely subtle. While it is stronger than pure nose, it is not so different.

The Monte Carlo remains centered around 0.955 and the observed cluster index falls near the lower tail of the distribution. We yield a p-value around 0.05, which is right on the conventional significance boundary. This suggests that SigClust barely detects the presence of two clusters, but the evidence is weaker than that of Alternative 1, which is to be expected.

This behavior is consistent with the theory that SigClust is more powerful with sparser and stronger signals. Dense and weak signals are more difficult to detect because the cluster index aggregates differences across all of the dimensions, and the small deviations are wiped out by high dimensional noise. SigClust is sensitive to strong but sparse structure, but less powerful for weak distributed signals, matching the patterns reported in the original paper.

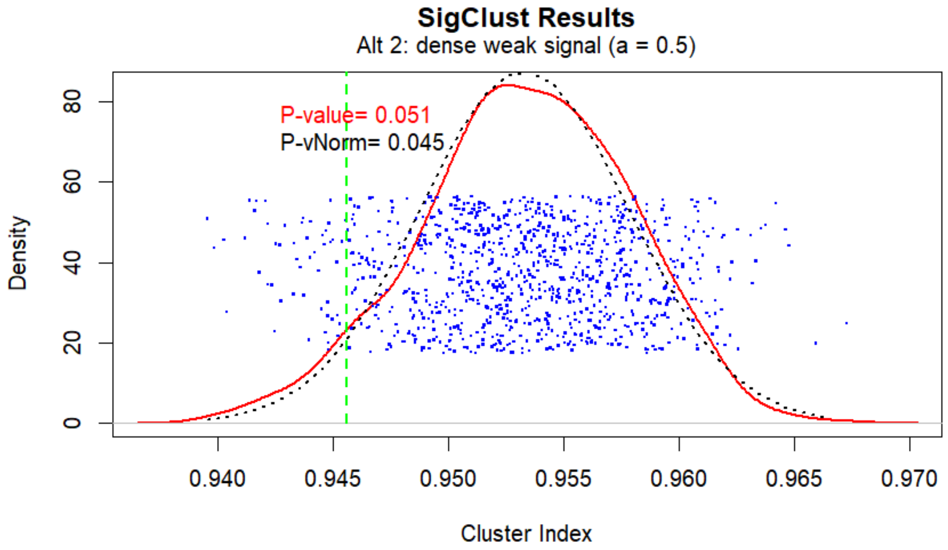

Figure 4: SigClust diagnostic plot for Alternative 2 (dense, weak signal, a=0.5)

4 Real Data Analysis

4.1 BRCA Simulation setup

4.2 Results

5 Conclusions

In this project, we explored how SigClust behaves under controlled, high dimensional simulation settings and recreated the paper's results. We observed substantial p-value variability across runs, with SigClust sometimes rejecting and sometimes failing to reject the null. This instability reflects a known phenomenon described in the original SigClust paper: power decreases dramatically for dense low-amplitude signals, especially in the presence of large eigenvalue spikes. Thus, our findings match the theoretical behavior reported by Huang et al. (2015).

For an implementation perspective, SigClust was computationally intensive. From the Monte Carlo procedure and eigenvalue shrinkage to the repeated k-means clustering, simulations became expensive and tuning `nsim` and the number of variables was mandatory.

Overall, this project does help clarify how SigClust behaves in both idealized null settings and more challenging alternatives. Future work perhaps could explore large Monte Carlo sample sizes, if SigClust works with more than two clusters, alternative covariance structures, and working on refining how it works with weaker signals.

6 References

7 References

Huang, H., Liu, Y., Yuan, M., & Marron, J. S. (2015). *Statistical Significance of Clustering Using SigClust.* Journal of Computational and Graphical Statistics, 24(3), 675–692. https://doi.org/10.1080/10618600.2014.951547

Huang, H., Liu, Y., Yuan, M., & Marron, J. S. (2015). *SigClust: Statistical Significance of Clustering.* R package documentation. https://cran.r-project.org/web/packages/sigclust/sigclust.pdf

8 Team member contribution

Team Member Contributions

Team Member	**Contributions**
Sharvee Joshi	Set up GitHub repository and project structure; implemented SigClust simulations (null, Alt 1, Alt 2); wrote R functions for covariance construction, data generation, and SigClust execution; created and interpreted diagnostic figures; authored the abstract and introduction; contributed to implementation narrative and debugging.
Makena Grigsby	Developed the methodology section, including theory behind SigClust and covariance estimation; summarized statistical framework; conducted BRCA data analysis; reproduced simulation components from the SigClust article; authored methodology and applied data-analysis sections.