Statistical Significance of Clustering using Soft Thresholding

Hanwen Huang[1], Yufeng Liu[1,2,3,4], Ming Yuan [5], and J. S. Marron[1,2,3]

[1] *Department of Statistics and Operations Research*

[2] *Department of Biostatistics*

[3] *Lineberger Comprehensive Cancer Center*

[4] *Carolina Center for Genome Sciences*

University of North Carolina at Chapel Hill

Chapel Hill, North Carolina 27599

[5] *Department of Statistics*

University of Wisconsin-Madison

Madison, WI 53706

email:hanwenh.unc@gmail.com

yfliu@email.unc.edu

myuan@stat.wisc.edu

marron@email.unc.edu

Abstract

Clustering methods have led to a number of important discoveries in bioinformatics and beyond. A major challenge in their use is determining which clusters represent important underlying structure, as opposed to spurious sampling artifacts. This challenge is especially serious, and very few methods are available when the data are very high in dimension. Statistical Significance of Clustering (SigClust) is a recently developed cluster evaluation tool for high dimensional low sample size data. An important component of the SigClust approach is the very definition of a

single cluster as a subset of data sampled from a multivariate Gaussian distribution. The implementation of SigClust requires the estimation of the eigenvalues of the covariance matrix for the null multivariate Gaussian distribution. We show that the original eigenvalue estimation can lead to a test that suffers from severe inflation of type-I error, in the important case where there are huge single spikes in the eigenvalues. This paper addresses this critical challenge using a novel likelihood based soft thresholding approach to estimate these eigenvalues which leads to a much improved SigClust. These major improvements in SigClust performance are shown by both theoretical work and an extensive simulation study. Applications to some cancer genomic data further demonstrate the usefulness of these improvements.

Keywords: Clustering; Covariance Estimation; High Dimension; Invariance Principles; Unsupervised Learning.

1 Introduction

Clustering methods have been broadly applied in many fields including biomedical and genetic research. They aim to find data structure by identifying groups that are similar in some sense. Clustering is a common step in the exploratory analysis of data. Many clustering algorithms have been proposed in the literature (see Duda et al. (2000); Hastie et al. (2009) for comprehensive reviews). Clustering is an important example of unsupervised learning, in the sense that there are no class labels provided for the analysis. Clustering algorithms can give any desired number of clusters, which on some occasions have yielded important scientific discoveries, but can also easily be quite spurious. This motivates some natural cluster evaluation questions such as:

- how to assess the statistical significance of a clustering result?

- are clusters really there or are they mere artifacts of sampling fluctuations?

- how can the correct number of clusters for a given data set be estimated?

Several cluster evaluation methods have been developed. McShane et al. (2002) proposed a cluster hypothesis test for microarray data by assuming that important cluster structure in the data lies in the subspace of the first three principal components,

where the low dimensional methods can be used. Tibshirani and Walther (2005) proposed using resampling techniques to evaluate the prediction strength of different clusters. Suzuki and Shimodaira (2006) wrote an R package for assessing the significance of hierarchical clustering. Despite progress in this area, evaluating significance of clustering remains a serious challenge, especially for High Dimensional Low Sample Size (HDLSS) situations.

Numerous works in applying Gaussian mixture models to cluster analysis have appeared in the literature. Overviews can be found in Mclachlan and Peel (2000); Fraley and Raftery (2002). For more recent work in this area, see Pan and Shen (2007); Wang and Zhu (2008); Xie et al. (2008). Gaussian mixture models need estimation of the full parameters of the Gaussian components of the mixture models, which can be quite challenging when tackling HDLSS problems.

Liu et al. (2008) proposed a Monte Carlo based method called Statistical Significance of Clustering (SigClust) which was specifically designed to assess the significance of clustering results for HDLSS data. An important contribution of that paper included a careful examination of the question of "what is a cluster". With an eye firmly on the very challenging HDLSS case, the answer was taken to be "data generated from a single multivariate Gaussian distribution". This Gaussian definition of "cluster" has been previously used by Sarle and Kuo (1993); Mclachlan and Peel (2000). This was a specific choice, which made the HDLSS problem tractable, but entailed some important consequences. For example, it is possible that none of Cauchy, Uniform, nor even t distributed data sets will give a single cluster in this sense. While this may seem to be a strong assumption, it has allowed sensible real data analysis in otherwise very challenging HDLSS situations, with a strong record of usefulness in bioinformatics applications, see e.g. Chandriani et al. (2009); Verhaak et al. (2010). From this perspective SigClust formulates the problem as a hypothesis testing procedure with

H_0: the data are from a single Gaussian distribution

H_1: the data are not from a Gaussian distribution.

The test statistic used in SigClust is the 2-means cluster index which is defined as the ratio of the within cluster variation to the total variation. Because this statistic is location and rotation invariant, it is enough to work only with a Gaussian null distribution with

mean 0 and diagonal covariance matrix Λ. The null distribution of the test statistic can be approximated empirically using a direct Monte Carlo simulation procedure. The significance of a clustering result can be assessed by computing an appropriate p-value. Recently, Maitra et al. (2012) proposed a non-parametric bootstrap approach for assessing significance in the clustering of multidimensional datasets. They defined cluster as a subset of data sampled from a spherically symmetry, compact and unimodal distribution and a non-parametric version of the bootstrap was used to sample the null distribution. It is important to note that their method has not been developed to handle HDLSS situations yet.

SigClust has given useful and reasonable answers in many high dimensional applications (Milano et al. (2008); Chandriani et al. (2009); Verhaak et al. (2010)). However, SigClust was based on some approximations, with room for improvement. In order to simulate the null distribution of the test statistic, SigClust uses invariance principles to reduce the problem to just estimating a diagonal null covariance matrix. This is the same task as finding the underlying eigenvalues of the covariance matrix. Therefore, a key step in SigClust is the effective estimation of these eigenvalues. Currently a factor analysis model is used to reduce the covariance matrix eigenvalue estimation problem to the problem of estimating a low rank component that models biological effects together with a common background noise level. However, the empirical studies in Section 2.3 show that this method can be dramatically improved.

Recently, many sparse methods have been introduced to improve the estimation of the high dimensional covariance matrix, see e.g. Meinshausen and Bühlmann (2006); Yuan and Lin (2007); Friedman et al. (2008); Rothman et al. (2008); Fan et al. (2009); Witten and Tibshirani (2010); Yuan (2010); Cai et al. (2011); Danaher et al. (2011), among many others. A critical difference between SigClust and these sparse approaches is that SigClust only needs estimates of the d eigenvalues instead of the $d(d-1)/2$ parameters of the full covariance matrix. This is because the null distribution of the test statistic used in SigClust is determined by the eigenvalues rather than the entire covariance matrix. Nevertheless, these sparse proposals have motivated us to improve the estimation of the eigenvalues. The soft thresholding method proposed in this paper is closely related to sparse covariance matrix estimation methods.

The contributions in this paper start by showing that, when there is a single strong spike in the eigenvalues, the original SigClust (which in Section 2.2 is seen to be reasonably called *hard thresholded*) can be seriously anti-conservative. This motivates an appropriate soft thresholding variation, with much better SigClust performance in those contexts. However, in the case of small total spikes, the soft thresholding becomes anti-conservative, while the hard gives much better performance. This motivates a combined method, which is able to take advantage of the strengths of each method, to give an overall useful variation of SigClust. The combined method is seen to give vastly improved SigClust performance over a very wide range of settings, through detailed simulations and theoretical analysis.

The rest of the article is organized as follows. In Section 2, we first give a brief description of the SigClust procedure and the existing hard thresholding eigenvalue estimation approach. Then we carefully develop the new likelihood based soft thresholding and combined approaches. To compare the performances of different methods, numerical studies are given in Section 3 for simulated and in Section 4 for real data examples. We provide some discussion in Section 5 and collect proofs of the likelihood derivation in the supplementary material.

2   Methodology

In this section, we first briefly review the SigClust method in Section 2.1. In Section 2.2, we provide an alternative likelihood based derivation, based on hard thresholding, for the estimation of the covariance matrix eigenvalues used in the original SigClust paper. Then we introduce a new soft thresholding approach in Section 2.3. The relationship between the Gaussian 2-means cluster index and the eigenvalues of the covariance matrix is derived in Section 2.4. Section 2.5 introduces a combined SigClust p-value calculation method.

2.1   Review of the Original SigClust Method

Suppose that the original data set X, of dimension $d \times n$, has d variables and n observations. The null hypothesis of SigClust is that the data are from a single Gaussian distribution $N(\mu, \Sigma)$, where μ is a d-dimensional vector and Σ is a $d \times d$ covariance matrix. SigClust uses the cluster index as the test statistic which has the nice property of

being both location and rotation invariant. This leads to a dramatic reduction in the number of parameters to be estimated. In particular, during simulation, the mean μ can be taken to be $\mathbf{0}$. In a parallel way, rotation invariance provides a major reduction in the parametrization of Σ to a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$, using the eigenvalue decomposition $\Sigma = U\Lambda U^T$, where U is an orthogonal matrix (essentially a rotation matrix). A factor analysis model is used to estimate the d eigenvalues which are still a relatively large number of parameters compared with the sample size n for HDLSS data sets. Specifically, Λ is modeled as

$$\Lambda = \Lambda_0 + \sigma_N^2 I, \tag{1}$$

where the diagonal matrix Λ_0 represents the real biology and is typically low-dimensional, and σ_N^2 represents the level of background noise. First σ_N is estimated as

$$\hat{\sigma}_N = \frac{\mathrm{MAD}_{d \times n \text{ data set}}}{\mathrm{MAD}_{N(0,1)}}, \tag{2}$$

where MAD stands for the median absolute deviation from the median. Then Λ is estimated to be

$$\hat{\lambda}_j = \begin{cases} \tilde{\lambda}_j & \text{if } \tilde{\lambda}_j \geq \hat{\sigma}_N^2 \\ \hat{\sigma}_N^2 & \text{if } \tilde{\lambda}_j < \hat{\sigma}_N^2, \end{cases} \tag{3}$$

where $(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_d)$ are the eigenvalues of the sample covariance matrix.

The procedure for SigClust can be briefly summarized as follows:

Step 1. Calculate the cluster index for the original data set.

Step 2. Obtain estimates $(\hat{\lambda}_1, \cdots, \hat{\lambda}_d)$ for the eigenvalues $(\lambda_1, \cdots, \lambda_d)$ of Σ.

Step 3. Simulate data N_{sim} times with each data set consisting of n i.i.d. observations from the null distribution $(x_1, \cdots, x_d)$ with $x_j \sim N(0, \hat{\lambda}_j)$. Here N_{sim} is some large number.

Step 4. Calculate the corresponding cluster index on each simulated data set from Step 3 to obtain an empirical distribution of the cluster index based on the null hypothesis.

Step 5. Calculate a p-value for the original data set and draw a conclusion based on a prespecified test level. The p-value can be calculated using either empirical quantile or Gaussian quantile.

2.2 Likelihood Interpretation of the SigClust Method: Hard Thresholding

Now we first show that the solution (3) can also be obtained based on a more general likelihood consideration, which gives the method by Liu et al. (2008) a new interpretation.. In factor models, the covariance matrix can be written as

$$\Sigma = \Sigma_0 + \sigma_N^2 I \tag{4}$$

for some low rank positive semi-definite matrix Σ_0. Denote

$$C \equiv \Sigma^{-1} \equiv (\Sigma_0 + \sigma_N^2 I)^{-1} = \frac{1}{\sigma_N^2} I - W_0 \tag{5}$$

for some positive semi-definite matrix W_0, with rank(Σ_0)=rank(W_0).

To estimate Σ, we minimize the negative log-likelihood to yield the following semi-definite program

$$\text{argmin}_C \left[- \log |C| + \text{trace}(C\tilde{\Sigma}) \right], \tag{6}$$

$$\text{subject to} \qquad C = \tfrac{1}{\sigma_N^2} I - W_0, \ C, W_0 \succeq 0, \tag{7}$$

where $\tilde{\Sigma} = (1/n)(X - \bar{X})(X - \bar{X})^T$ and $A \succeq 0$ means that A is positive semi-definite. Here the use of the $\bar{X}$ term is to make the mean of each row 0 for X.

In factor models, we want to encourage a small number of factors which amounts to encouraging a small rank(Σ_0)=rank(W_0). The direct approach to enforcing low rank Σ_0 or W_0 is by adding an extra rank constraint:

$$\text{rank}(W_0) \leq l, \tag{8}$$

where l is a pre-specified tuning parameter. Denote the eigenvalue decomposition $W_0 = UDU^T$ and $\tilde{\Sigma} = \tilde{U}\tilde{\Lambda}\tilde{U}^T$, where $D = \text{diag}(d_1, \cdots, d_d)$ and $\tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_d)$. Then $C = U(\frac{1}{\sigma_N^2} I - D)U^T$.

Theorem 1. *The solution to (6),(7),(8) is given by $U = \tilde{U}$ and*

$$\hat{d}_k = \begin{cases} \frac{1}{\sigma_N^2} - \frac{1}{\tilde{\lambda}_k} & \text{if } k \leq l \text{ and } \tilde{\lambda}_k > \sigma_N^2 \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Proof of this Theorem and other proofs are given in the supplementary material. By Theorem 1, we get the eigenvalues of the covariance matrix which are identical to (3) with suitable choices of l, i.e. greater than or equal to the number of eigenvalues which are bigger than σ_N^2 given by (3). We call this estimation the hard thresholding approach, so this name applies to the estimation used in Liu et al. (2008) as described in Section 2.1 above.

2.3   Soft Thresholding Approach

As mentioned in Liu et al. (2008), a challenge for the hard thresholding method is to estimate the large eigenvalues in the HDLSS settings. This is illustrated in Figure 1 using a simple HDLSS example with $n = 50$ and $d = 1000$. The data are generated from a multivariate normal distribution with covariance matrix Λ, where Λ is diagonal with diagonal elements $(\underbrace{v, \cdots, v}_{w}, 1, \cdots, 1)$. We consider $v = 100$ and $w = 10$. In the HDLSS setting, it is well known that the sample estimators of the larger eigenvalues differ greatly from the corresponding true values (Baik and Silverstein, 2006). But from (3), the hard thresholding estimators are identical to the sample estimators on all eigenvalues beyond the background noise, and thus bigger than the corresponding true values. Now we propose a less aggressive thresholding scheme which can reduce the larger eigenvalues from the sample estimators.

Our approach is to add a smooth constraint instead of the rank constraint on W_0. Similar to the hard thresholding, instead of counting the nonzero eigenvalues, we add an extra constraint on the sum of the eigenvalues of W_0, which equals trace(W_0). Consequently, (8) becomes

$$\text{trace}(W_0) \leq M \tag{10}$$

for a tuning parameter $M \geq 0$. The constraint above is a convex envelop to rank(W_0) and therefore a convex relaxation to the constraint on rank(W_0). Clearly when $M = 0$, we force the covariance matrix to be the identity. When M increases, more and more factors enter the estimation. The constraint (10) is a nuclear norm constraint, which has been well-studied in the convex optimization literature, see e.g. Fazel (2002).

Interestingly, the solution of (6), (7), (10) can be given in a closed form as stated in Theorem 2 below.
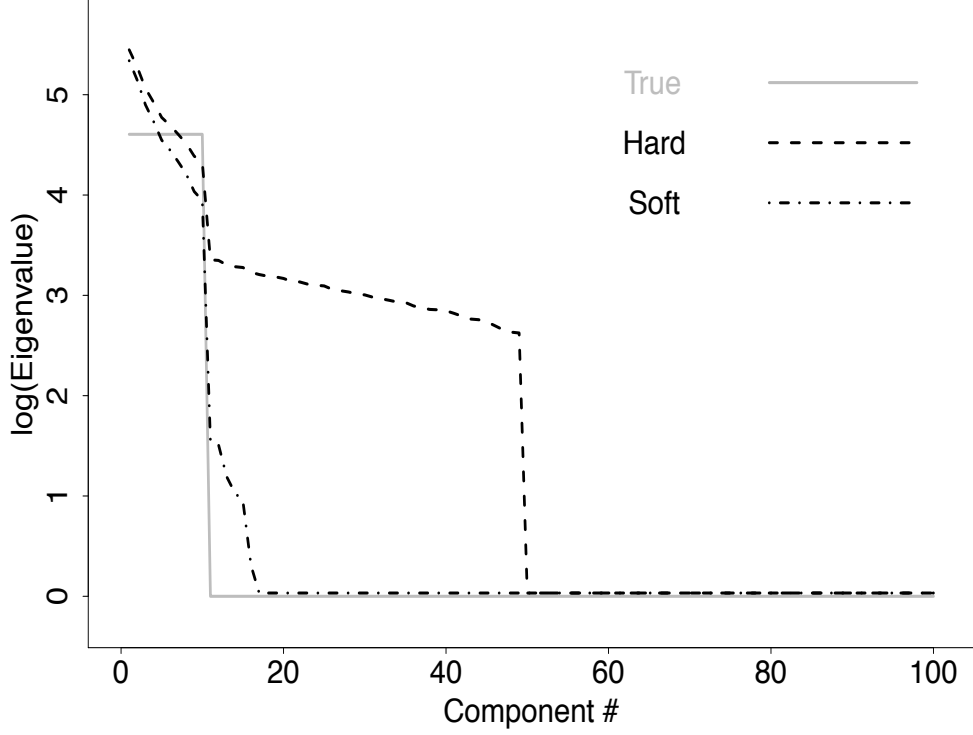
8

Figure 1: True and estimated covariance matrix eigenvalues based on hard- and soft-thresholding methods for a simulated data set with $d = 1000$ and $n = 50$. This shows that some eigenvalues are highly over-estimated by the hard thresholding method. The soft thresholding method gives major improvement for this example.

Theorem 2. *The solution to (6), (7), (10) is given by $U = \tilde{U}$ and*

$$\hat{d}_k = \left(\frac{1}{\sigma_N^2} - \frac{1}{\tilde{\lambda}_k - \tau} \right)_+ \tag{11}$$

where $\tau > 0$ is a constant such that

$$\sum_{k=1}^{d} \hat{d}_k = \sum_{k=1}^{d} \left(\frac{1}{\sigma_N^2} - \frac{1}{\tilde{\lambda}_k - \tau} \right)_+ = M. \tag{12}$$

Correspondingly, the eigenvalues of the estimated covariance matrix are given by

$$\hat{\lambda}_k = \begin{cases} \tilde{\lambda}_k - \tau & \text{if } \tilde{\lambda}_k > \tau + \sigma_N^2 \\ \sigma_N^2 & \text{if } \tilde{\lambda}_k \leq \tau + \sigma_N^2 \end{cases} = (\tilde{\lambda}_k - \tau - \sigma_N^2)_+ + \sigma_N^2. \tag{13}$$

To determine the optimal thresholding parameter M, we match the sum of the eigenvalues of the estimated covariance matrix with those of the sample covariance matrix:

$$\sum_{k=1}^{d} \{ (\tilde{\lambda}_k - \tau - \sigma_N^2)_+ + \sigma_N^2 \} = \sum_{k=1}^{d} \tilde{\lambda}_k, \tag{14}$$

9

where the right hand side is an unbiased estimate of trace(Σ).

The difference between the soft thresholding method (13) and the hard thresholding estimators (3) is that the large eigenvalues are subtracted by a constant τ. This improves the performance for the HDLSS example illustrated in Figure 1.

2.4   Theoretical Gaussian 2-means Cluster Index

Once the covariance matrix eigenvalues are estimated, we can proceed the SigClust analysis. Toward that end, we need to determine the null distribution of the 2-means cluster index. In this section, we will derive the relationship between the cluster index and eigenvalues theoretically.

Let $\mathbf{x} = (x_1, \cdots, x_d)$ be a d-dimensional random vector having a multivariate normal distribution of $\mathbf{x} \sim N(0, \Sigma)$ with mean 0 and covariance matrix $\Sigma = U\Lambda U^T$. Here U is an orthogonal matrix and Λ is a diagonal matrix, i.e., $\Lambda = diag(\lambda_1, \cdots, \lambda_d)$, where $\lambda_1 \geq \cdots \geq \lambda_d$. Define the theoretical total sum of squares as

$$\mathrm{T}SS = E||\mathbf{x}||^2 = \int ||\mathbf{x}||^2 \phi(\mathbf{x})d\mathbf{x}. \tag{15}$$

The theoretical within cluster sum of squares (WSS) is based on a theoretical analog of clusters, which is a partition of the entire feature space R^d into $\mathcal{S}_1$ and $\mathcal{S}_2$. Define $\boldsymbol{\mu}_1 = \int_{\mathbf{x} \in \mathcal{S}_1} \mathbf{x}\phi(\mathbf{x})d\mathbf{x}$ and $\boldsymbol{\mu}_2 = \int_{\mathbf{x} \in \mathcal{S}_2} \mathbf{x}\phi(\mathbf{x})d\mathbf{x}$. Then we have

$$\mathrm{W}SS = \int_{\mathbf{x} \in \mathcal{S}_1} ||\mathbf{x} - \boldsymbol{\mu}_1||^2 \phi(\mathbf{x})d\mathbf{x} + \int_{\mathbf{x} \in \mathcal{S}_2} ||\mathbf{x} - \boldsymbol{\mu}_2||^2 \phi(\mathbf{x})d\mathbf{x}. \tag{16}$$

The relationship between the theoretical cluster index (TCI) and the covariance matrix eigenvalues is stated by the following Theorem.

Theorem 3. *For the optimal choice of $\mathcal{S}_1$ and $\mathcal{S}_2$, i.e. the split is operated in such a way that the total WSS is minimized (this is the theoretical analog of 2-means clustering), the corresponding TCI is*

$$TCI = \frac{WSS}{TSS} = 1 - \frac{2}{\pi} \frac{\lambda_1}{\sum_{i=1}^d \lambda_i}. \tag{17}$$

Note that TCI is a population version of the cluster index which can help us to learn some insights about the behavior of cluster index in the population sense. Theorem 3 tells us that the optimal TCI is only determined by two quantities, the largest eigenvalue λ_1 and the total sum of eigenvalues $\sum_{i=1}^d \lambda_i$. In practice, different methods give different

estimations of these two quantities, and thus in turn lead to different influences on the SigClust p-values.

For the sample covariance estimation method, the estimated λ_1 is typically larger (i.e. biased upwards) than the true value, in HDLSS situations. Since the sum of the sample eigenvalues is a consistent estimate of the total variation, i.e. the denominator of (17), it follows that the resulting estimate of TCI will generally be smaller (biased downwards) than the true TCI in that case, giving a larger p-value and thus a conservative result.

For the hard thresholding method, estimation of λ_1 and $\sum \lambda_i$ are both biased. Let δ_1, and Δ denote the bias for λ_1 and $\sum_{i=1}^d \lambda_i$ respectively. Then the difference between the true TCI and the hard thresholding estimate is proportional to

$$E = \frac{\lambda_1 + \delta_1}{\sum_{i=1}^d \lambda_i + \Delta} - \frac{\lambda_1}{\sum_{i=1}^d \lambda_i} = \frac{\sum_{i=1}^d \lambda_i \delta_1 - \lambda_1 \Delta}{\sum_{i=1}^d \lambda_i (\sum_{i=1}^d \lambda_i + \Delta)}. \tag{18}$$

If $E < 0$, the result is anti-conservative.

For the soft thresholding method, the estimated denominator of (17) is unbiased whereas the estimated numerator is biased. Let the estimated numerator be $\lambda_1 + \delta_1 - \tau$, where τ is defined in (11). Clearly, if $\delta_1 < \tau$, the soft method is anti-conservative.

2.5 The combined SigClust p-value

Our theoretical results in Section 2.4 show that both the hard and soft thresholding methods can lead to anti-conservative results. However, a closer examination reveals that they can be complementary to each other. Let's consider a simple setting for a diagonal matrix Λ with diagonal elements $(\underbrace{v, \cdots, v}_{w}, 1, \cdots, 1)$. Assume $d > n$ and $w < n$. Accordingly to the random matrix theory (Baik and Silverstein, 2006), as $n = n(d) \to \infty$ such that $d/n \to \rho$, the eigenvalue estimates converge to

$$\hat{\lambda}_j \to \begin{cases} v + \frac{\rho v}{v-1} & 1 \leq j \leq w \text{ and } v > 1 + \sqrt{\rho} \\ (1 + \sqrt{\rho})^2 & j = w + 1 \\ (1 - \sqrt{\rho})^2 & j = n \\ 0 & j = n + 1, \cdots, d. \end{cases} \tag{19}$$

If $\lambda_1 = v \gg 1$, we have $\delta_1 \approx \rho$, $\Delta \approx d - n$. The numerator of (18) is

$$(wv + d - w)d/n - v(d - n) = (d - n - wd/n)\left(\frac{(d-w)/n}{1 - n/d - w/n} - v\right). \tag{20}$$

11

Therefore, for fixed n, d, and w, as v increases, the hard thresholding method tends to give an anti-conservative result. On the other hand, for large λ_1, the soft method tends to be conservative. This can be explained by the formulas in (2.4), since τ is mainly determined by the dimension and sample size which are fixed for a given setting. Larger total spikes will lead to larger δ_1 and move the soft method toward conservative. These observations drive the effects seen in the simulation studies in the next section.

This motivates us to propose a new p-value calculation method called the *combined* method which combines the best aspects of both the hard and soft approaches. This combination version of SigClust requires some modification of the algorithm given at the end of Section 2. To use the best features of both hard and soft thresholding, both sets of estimated eigenvalues are computed. At the data generation step 3, a single standard normal realization gives both hard and soft data cases, from multiplying by each set of square root eigenvalues. Let CI_{hard} and CI_{soft} denote the cluster indices computed from the hard and soft data cases respectively. The minimum of this pair is summarized across realizations to give the empirical distribution of the cluster index under the null for the combined method, i.e. $\mathrm{CI}_{combined} = \min(CI_{hard}, CI_{soft})$.

3 Simulation

In this section we investigate how the estimation of the covariance matrix eigenvalue affects the SigClust performance using extensive simulation studies. Five SigClust p-value computation methods are compared. The first four are based on the simulation results using the covariance matrices estimated from the true, sample, hard and soft thresholding approaches, which are referred to using those names. The fifth method is the combined method.

We have performed simulations for both low and high dimensional situations. Here we focus on high dimensional results, because our main contribution is in HDLSS settings. Three types of examples are generated here including situations under both null and alternative hypothesis. The sample size is $n = 100$ and dimension is $d = 1000$. We evaluate different methods based on the criterion of whether or not they can maximize the power while controlling the type-I error. In Section 3.1, we consider examples of data under the null hypothesis, i.e., having only one cluster generated by a single Gaussian

distribution. In each example we check the type-I error of SigClust by studying how often it incorrectly rejects the null hypothesis H_0. In Sections 3.2 and 3.3, we consider data from a collection of mixtures of two Gaussian distributions with different signal sizes and explore the power of SigClust in terms of how often it correctly rejects the null hypothesis. We summarize the simulation results in Section 3.4.

3.1 One Cluster

In order to evaluate the Type I error rates for different methods, data were generated under the null hypothesis, i.e. from a single multivariate Gaussian distribution with covariance matrix Λ which is diagonal with diagonal elements $(\underbrace{v, \cdots, v}_{w}, 1, \cdots, 1)$. We consider 31 combinations of v and w with $v = 1, \cdots, 1000$, and the corresponding $w = 1, \cdots, 100$, as shown in Table 1. The simulation procedure was repeated 100 times for each setting.

Table 1 summarizes the mean and 100 times the 0.05 and 0.1 quantiles of the p-values of the empirical distributions based on different methods under the various parameter settings. Theoretically the p-value follows the uniform $[0, 1]$ distribution since the data are generated from a single Gaussian distribution. As expected, the empirical distributions of p-values using the true method are relatively close to the uniform distribution. The sample method results in p-values whose means are always bigger than the expected ones. This is consistent with the theoretical results shown in Section 2.4. The 0.05 and 0.1 quantiles are almost all 0, so we conclude that the sample method is conservative in all of these settings. For settings of $v \geq 30$, i.e. for populations with a generally large first few eigenvalues (e.g. a strongly elongated distribution), with the exception of $(v, w) = (40, 25)$, results based on the hard method exhibits more small p-values than expected under the uniform distribution which implies that this approach is frequently anti-conservative. On the other hand, the hard method tends to be quite conservative, for relatively small values of v, e.g. for approximately more spherical Gaussian distributions. This can also be understood from equation (18).

For the soft method, in contrast to the hard one, the results of Table 1 show anti-conservatism in situations where the sum of the spikes, $v \times w$, is relatively small (less than 50), such as $(v, w) = (20, 1)$, $(v, w) = (10, 2)$ and $(10, 1)$. As either v or w grows, the soft method becomes conservative. For most of situations, either the hard or the soft methods will give conservative results. For this reason, the combined method leads to conservative

13

results in all of these settings except for $v = 30$, $w = 1$. We have included more simulated examples in the supplementary materials with different d, and the conclusions are very similar. For higher $d = 5000$, there was even less anti-conservatism. For lower d, there was a little more, but only for thin spike situations with $w = 1, 2$. As shown here the number of cases of anti-conservatism is far less than for either hard or soft thresholding alone. Like the sample method, the combined method effectively controls type-I error under the null hypothesis. But more importantly it can dramatically increase the power over the sample method under important alternative hypotheses as shown in the next section. Therefore, we recommend the combined method in practice, and this is what we use as the basis of comparison with existing methods in the power studies in the coming sections.

The means of the p-value populations give additional insights, and the results are mostly consistent with those from the quantiles. In particular, the theoretical means are generally close to the desired value of 0.5, the sample means tend to be larger, and the hard, soft and combined means fluctuate in a way that corresponds to their quantile behavior. An important point is that the combined means are generally substantially closer to 0.5 than is true for the hard ones.

3.2 Mixture of Two Gaussian Distributions with Signal in One Coordinate Direction

In this section, we compare the power properties of these various SigClust hypothesis tests. This is based on a mean mixture of two normal distributions, $.5N(0, \Lambda) + .5N(\mu, \Lambda)$, where $\mu = (a, 0, \cdots, 0)$ with $a = 0, 10, 20$ and $\Lambda = \mathrm{diag}(\underbrace{v, \cdots, v}_{w}, 1, \cdots, 1)$ a diagonal matrix. We choose $v = 10$ and $w = 10$ here. When $a = 0$, the distribution reduces to a single Gaussian distribution. The larger the a, the greater the signal. The theoretical null distribution is $N(0, \Lambda^*)$, where $\Lambda^* = \mathrm{diag}(\lambda_1 + 0.25a^2, \lambda_2, \cdots, \lambda_d)$. The empirical distributions of p-values are shown in Figure 2. As expected, the true method is very powerful under the alternative hypothesis and meanwhile can control the type-I error well under the null hypothesis ($a = 0$). Both the hard and combined methods give reasonable performance relative to the true method in this setting. On the other hand, the sample method is very conservative (more large p-values than expected from the uniform distribution) under the

Table 1: Summary table of empirical SigClust p-value distribution over 100 replications based on five methods under different settings in Simulation 3.1. The mean and the numbers of p-values which are less than 0.05 (denoted as P5) and 0.1 (denoted as P10) are reported (d=1000, n=100).

		True			Sample			Hard			Soft			Combined		
v	w	Mean	P5	P10	Mean	P5	P10	Mean	P5	P10	Mean	P5	P10	Mean	P5	P10
1000	1	0.47	5	8	0.52	0	1	0.00	100	100	0.47	1	1	0.46	1	2
200	5	0.38	4	11	0.82	0	0	0.01	95	100	0.69	0	0	0.69	0	0
100	10	0.35	9	16	0.95	0	0	0.08	35	69	0.82	0	0	0.82	0	0
40	25	0.31	8	18	1.00	0	0	0.55	0	0	0.96	0	0	0.96	0	0
20	50	0.26	9	17	1.00	0	0	0.96	0	0	1.00	0	0	1.00	0	0
10	100	0.22	15	26	1.00	0	0	1.00	0	0	1.00	0	0	1.00	0	0
200	1	0.49	6	11	0.62	0	0	0.00	100	100	0.41	0	0	0.41	0	0
100	1	0.48	4	12	0.78	0	0	0.00	100	100	0.42	0	0	0.42	0	0
50	1	0.51	5	8	0.93	0	0	0.00	98	100	0.30	0	4	0.30	0	6
40	1	0.53	7	10	0.96	0	0	0.01	94	98	0.28	2	5	0.28	2	5
30	1	0.51	7	11	0.99	0	0	0.06	58	83	0.20	6	23	0.22	5	19
20	1	0.54	3	6	1.00	0	0	0.49	1	4	0.14	20	45	0.52	0	2
10	1	0.47	4	10	1.00	0	0	1.00	0	0	0.05	70	90	1.00	0	0
50	10	0.37	10	18	0.98	0	0	0.06	57	82	0.77	0	0	0.77	0	0
40	10	0.37	2	14	0.99	0	0	0.07	45	81	0.76	0	0	0.76	0	0
30	10	0.37	7	13	1.00	0	0	0.12	18	47	0.72	0	0	0.73	0	0
20	10	0.37	7	12	1.00	0	0	0.38	4	5	0.66	0	0	0.70	0	0
10	10	0.33	6	16	1.00	0	0	0.99	0	0	0.43	0	2	0.99	0	0
50	5	0.35	8	12	0.96	0	0	0.01	99	100	0.57	0	0	0.57	0	0
40	5	0.40	6	11	0.98	0	0	0.02	94	100	0.56	0	0	0.56	0	0
30	5	0.38	11	15	0.99	0	0	0.05	68	89	0.49	0	0	0.50	0	0
20	5	0.39	3	15	1.00	0	0	0.29	2	12	0.39	0	0	0.48	0	0
10	5	0.37	10	15	1.00	0	0	0.99	0	0	0.19	5	24	0.99	0	0
50	2	0.48	6	11	0.95	0	0	0.00	100	100	0.42	0	0	0.42	0	0
40	2	0.41	12	20	0.96	0	0	0.01	97	99	0.35	0	2	0.35	0	3
30	2	0.47	6	10	0.99	0	0	0.04	70	87	0.31	0	3	0.31	0	4
20	2	0.45	4	12	1.00	0	0	0.34	2	6	0.18	3	23	0.40	2	3
10	2	0.45	5	12	1.00	0	0	1.00	0	0	0.06	54	80	1.00	0	0
5	1	0.23	17	31	1.00	0	0	1.00	0	0	0.18	32	49	1.00	0	0
3	1	0.16	27	46	1.00	0	0	1.00	0	0	0.24	15	29	1.00	0	0
1	1	0.19	19	33	1.00	0	0	1.00	0	0	0.28	10	20	1.00	0	0

null hypothesis and also lacks power under the alternative hypothesis when the signal is not big ($a = 10$). It gains some power (the curve bends toward the upper left corner) as the signal increases ($a = 20$).
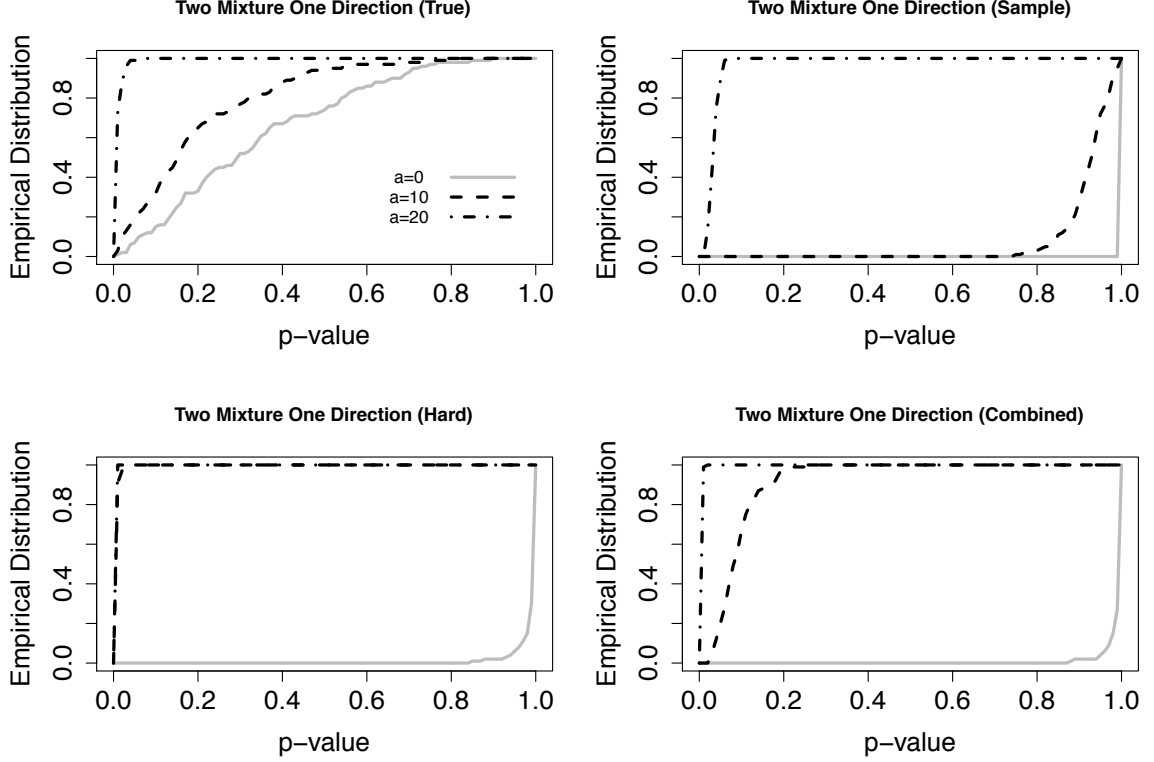

Figure 2: Empirical distributions of SigClust p-values for Simulation 3.2. This shows Sample is too conservative, while Combined is generally close to True in overall performance.

3.3 Mixture of Two Gaussian Distributions With Signal in All Coordinate Directions

In the previous subsection, the signal is only in the first coordinate direction. Now we consider power using another example with the signal in all coordinate directions. Similarly, we generate data from a mixture of two Gaussian distributions, $.5N(0, \Lambda) + .5N(\boldsymbol{\mu}, \Lambda)$, where $\boldsymbol{\mu} = (a, a, \cdots, a)$ with $a = 0, 0.4, 0.6$ and $\Lambda = \text{diag}(v, 1, \cdots, 1)$ with $v = 100$. This signal is very small in each direction, but can be large when all directions are combined together. The empirical distributions of p-values calculated from the 100 simulated datasets based on different methods are displayed in Figure 3. For $a = 0$ the

results are identical to the single cluster situation in Section 3.1 with $(v, w) = (100, 1)$. The hard thresholding method always yields smaller p-values than expected and thus is strongly anti-conservative under the null hypothesis and powerful under the alternative hypothesis. In contrast, the combined method is conservative under the null but becomes powerful as the signal increases. When the signal is big enough, e.g. $a = 0.6$, all methods can identify the significant clusters. For small signal situations, e.g. $a = 0.4$, the combined method is much more powerful than both the sample and true methods.
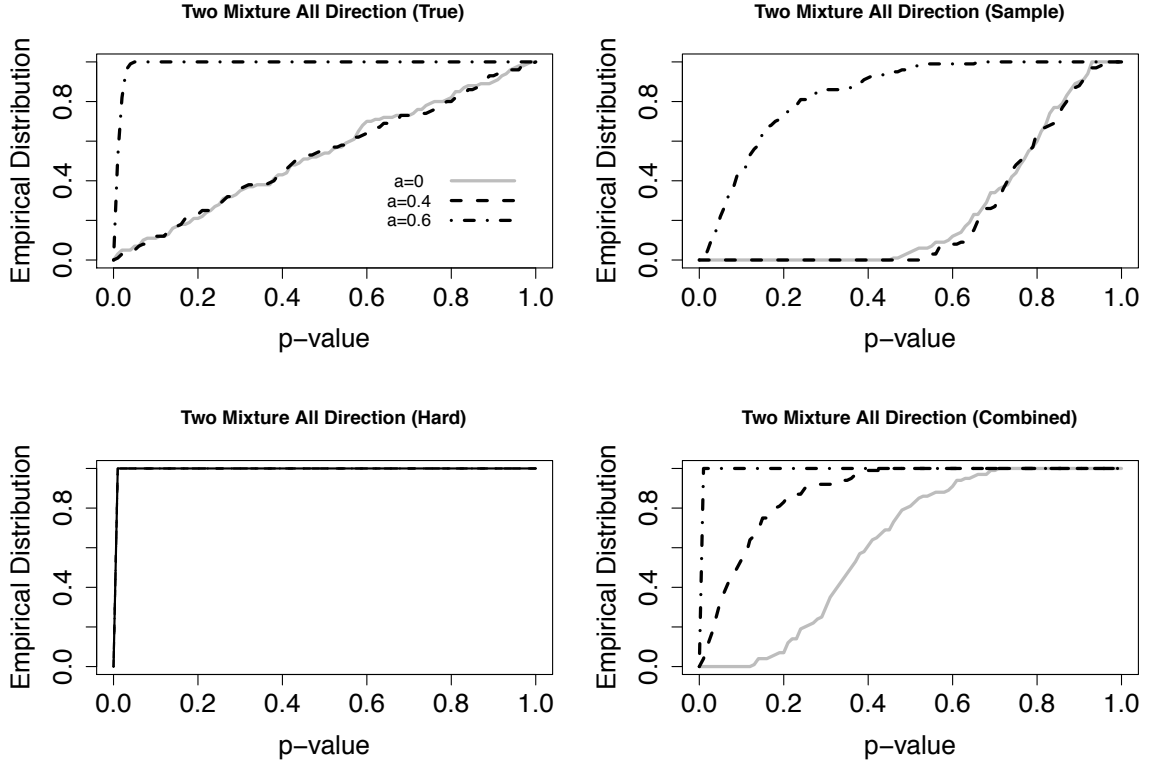

Figure 3: Empirical distributions of SigClust p-values for Simulation 3.3. The results indicate that Hard is strongly anti-conservative, while sample is too conservative. Overall best is the Combined method.

3.4 Simulation Summary

In summary, the sample method is strongly conservative and the hard method can be anti-conservative in many situations. The soft method is sometimes in-between and other times is more anti-conservative than the hard. Simulation results shown in Section 3.1 suggest that, under the null hypothesis, the performances of the hard and soft methods

vary from strongly conservative to strongly anti-conservative depending on the situations which are mainly characterized by the two quantities, v and $v \times w$. Fortunately, the two methods are frequently complementary to each other, i.e., the combined method yields conservative results in almost all settings. Simulation results from Sections 3.2 and 3.3 suggest that, under the alternative hypothesis, the hard method often has the largest power and the sample method has the smallest power. The combined method is appropriately in-between. If the signals are large enough, all methods can identify the significant clusters. However, in situations with relatively small signal, the sample method cannot distinguish the significant clusters. In practice, we recommend the combined method, i.e., small p-values from the combined method indicate the existence of distinct clusters.

4 Real Data

In this section, we apply our methods to some real cancer data sets. As mentioned in Verhaak et al. (2010), Glioblastoma Multiforme (GBM) is one of the most common forms of malignant brain cancer in adults. For the purposes of the current analysis, we selected a cohort of patients from The Cancer Genome Atlas Research Network (TCGA, 2010) with GBM cancer whose brain samples were assayed on three gene expression platforms (Affymetrix HuEx array, Affymetrix U133A array, and Agilent 244K array) and combined into a single unified data set. Four clinically relevant subtypes were identified using integrated genomic analysis in Verhaak et al. (2010), they are Proneural, Neural, Classical, and Mesenchymal. We first filter the genes using the ratio of the sample standard deviation and sample mean of each gene. After gene filtering, the data set contained 383 patients with 2727 genes. Among the 383 samples, there are 117 Mesenchymal samples, 69 Neural samples, 96 Proneural samples, and 101 Classical samples.

We apply SigClust to every possible pair-wise combination of subclasses and calculate the p-value based on the three different methods. Here the cluster-index is computed based on the given cluster label. Except the MES and CL pair, the p-values from all three methods are highly significant for all other pairs which implies that they are well separated from each other. For the MES and CL pair, the p-value is highly non-significant using the sample method (0.93) while it is highly significant for both the hard (1.49×10^{-8}) and

Table 2: SigClust p-values for each pair of subtypes for the BRCA data. The known cluster labels are used to calculate the cluster index. Here we use 0 to represent all the tiny p-values which are less than 10^{-10}.

	Basal.LumA	Basal.LumB	Basal.Her2	LumA.LumB	Her2.LumB	Her2.LumA
Sample	3.49×10^{-7}	1.06×10^{-4}	0.015	1	0.99	0.77
Hard	0	0	0	0.89	0.051	4.59×10^{-10}
Combined	0	0	8.86×10^{-7}	1	0.95	0.038

combined methods (8.16×10^{-5}). According to the conclusion drawn from our simulation studies in Section 3, the MES and CL are separated from each other as well.

The second real example we used is breast cancer data (BRCA) also from The Cancer Genome Atlas Research Network which include four subtypes: LumA, LumB, Her2 and Basal and have been extensively studied by microarray and hierarchical clustering analysis (Fan et al., 2006). The sample size is 348 and the number of genes used in the analysis after filtering is 4000. Among 348 samples, there are 154 LumA, 81 LumB, 42 Her2 and 66 Basal. The results of applying SigClust to each pair of subclasses are shown in Table 2. For pairs including Basal, the p-values from all three methods are significant which implies that Basal can be well separated from the rest. For the LumA and LumB pair, all methods report very high p-values, which suggests that they are actually one subtype which is consistent with the findings of Parker et al. (2009), which suggest that these are essentially a stretched Gaussian distribution (thus not flagged by SigClust), with an important clinical division within that distribution. For the Her2 and LumB pair, all three methods give a non-significant p-value although not as big as for the LumA and LumB pair, so there is no strong evidence for them to be separated (although hard thresholding appears to be close to a spuriously significant result). For the Her2 and LumA pair, the hard and combined methods give significant p-values, whereas the sample method fails to find this important difference. So this pair can be significantly separated as well. Note that the p-values listed in Table 2 are consistent with the scatter plot in Figure 4 where the projections of the data points onto the first four principal component (PC) directions are displayed. Clearly, Basal is well separated from the remaining data. LumA and LumB are close together and LumB and Her2 are closer than LumA and Her2.
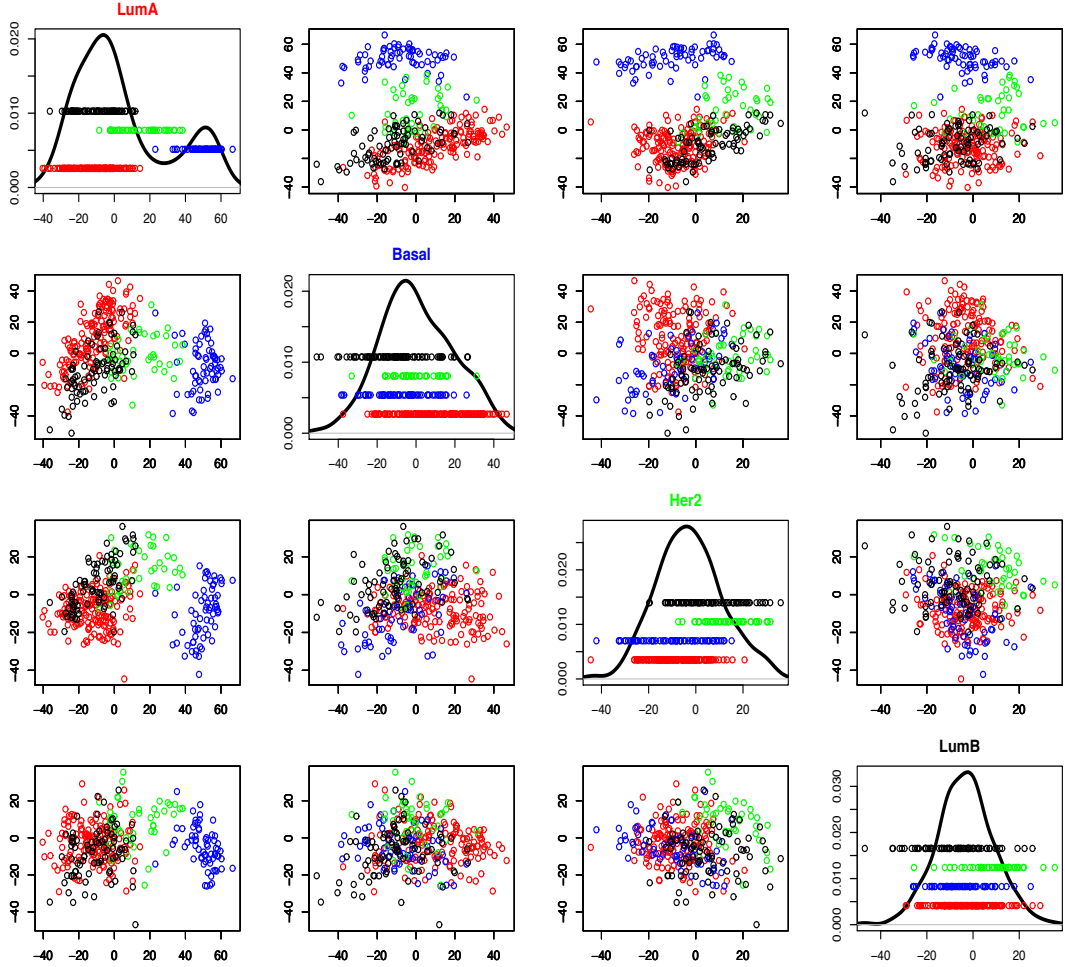
Figure 4: PCA projection scatter plot view of the BRCA data, showing 1D (diagonal) and 2D projections of the data onto PC directions. Groupings of colors indicate biological subtypes.

5 Discussion

Despite the work on covariance matrix estimation, accurate estimation of the eigenvalues of the covariance matrix remains challenging in high dimensions. In this paper, we developed a soft thresholding approach and examined its application to the SigClust method. We found that both the newly proposed soft thresholding approach and the hard thresholding approach used in the original SigClust paper can be derived under a likelihood based framework with two different regularizations (ℓ_0 regularization for hard and ℓ_1 for soft). Through extensive simulation, we compared the performance of the SigClust method based on different approaches in a wide variety of settings. We found that the sample method was always conservative, while both the hard and soft would sometimes

incorrectly reject the null. Fortunately, the latter occurrences were approximately complementary leading to a combined approach, which gave far fewer incorrect rejections. The combined approach was seen to have much better power properties than using simple sample covariance estimation. We recommend that our newly proposed combined method be used in practice because it has been shown to control the type-I error as well as the sample method under the null hypothesis, while gaining much more power under the alternative hypothesis.

SigClust is constructed on the basis of the definition of a single cluster as a Gaussian distribution. Thus a significant SigClust p-value indicates that the data do not come from a single Gaussian distribution. Note that other possible definitions of clusters, based on only unimodality, such as the uniform distribution, will be deliberately rejected by the current version of SigClust. Simulated examples are given in Huang et al. (2012) to illustrate these cases in which SigClust reports significant p-values for data sets generated from an uniform distribution on a two dimensional disk. Diagnostic tools to examine the applicability of SigClust appear in Huang et al. (2012). We recommend the typical application of the current SigClust approach be conducted together with diagnostics.

In terms of software, the R package for the current version of SigClust can be freely downloaded on the CRAN website: http://cran.r-project.org/web/packages/sigclust/index.html. Computation time depends on the number of simulated replications, and the size of the input data. In all cases here, we used 1000 replications, and it took around 1 minute for each simulated data set described in Section 3 and 10 minutes for both the GBM data and the BRCA data described in Section 4.

References

Baik, J. and J. W. Silverstein (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis 97*, 1382–1408.

Cai, T. T., L. W., and X. Luo (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*, 594–607.

Chandriani, S., E. Frengen, V. H. Cowling, S. A. Pendergrass, C. M. Perou, M. L. Whitfield, and M. D. Cole (2009). A core myc gene expression signature is prominent in

basal-like breast cancer but only partially overlaps the core serum response. *PLoS ONE 4*(8), e6693.

Danaher, P., P. Wang, and D. Witten (2011). The joint graphical lasso for inverse covariance estimation across multiple classes. arXiv:1111.0324.

Duda, R. O., P. E. Hart, and D. G. Stork (2000). *Pattern Classification*. Wiley-Interscience Publication.

Fan, C., D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. van't Veer, and C. M. Perou (2006). Concordance among gene-expressionbased predictors for breast cancer. *New England Journal of Medicine 355*(6), 560–569.

Fan, J., Y. Feng, and Y. Wu (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics 3*, 521–541.

Fazel, M. (2002). Matrix rank minimization and applications. Ph.D. Thesis, Stanford University.

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

Friedman, J. H., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*, 432–441.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (second ed.). Springer.

Huang, H., Y. Liu, D. N. Hayes, A. Nobel, J. S. Marron, and C. Hennig (2012). Significance testing in clustering. *Submitted*.

Liu, Y., D. N. Hayes, A. Nobel, and J. S. Marron (2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association 103*(483), 1281–1293.

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

Mclachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

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

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics 34*, 1436–1462.

Milano, A., S. A. Pendergrass, J. L. Sargent, L. K. George, T. H. McCalmont, M. K. Connolly, and M. L. Whitfield (2008). Molecular subsets in the gene expression signatures of scleroderma skin. *PLoS ONE 3*(7), e2696.

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

Parker, J. S., M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, and et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology 27*(8), 1160–1167.

Rothman, A., E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515.

Sarle, W. S. and A. H. Kuo (1993). The modeclus procedure. Technical Report P-256, Cary, NC: SAS Institute Inc.

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

TCGA (2010). The cancer genome atlas research network. http://cancergenome.nih.gov/wwd/pilot_program/research_network/cgcc.asp.

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

Verhaak, R. G., K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D.

James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes, and Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell 17*(1), 98–110.

Wang, S. and J. Zhu (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics 64*, 440–448.

Witten, D. and R. Tibshirani (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association 105*(490), 713–726.

Xie, B., W. Pan, and X. Shen (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics 2*, 168–212.

Yuan, M. (2010). Sparse inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research 11*, 2261–2286.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*(1), 19–35.