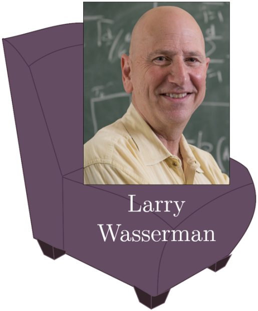
Gaussian Mixture Clustering Using Relative Tests of Fit

Purvasha Chakravarti



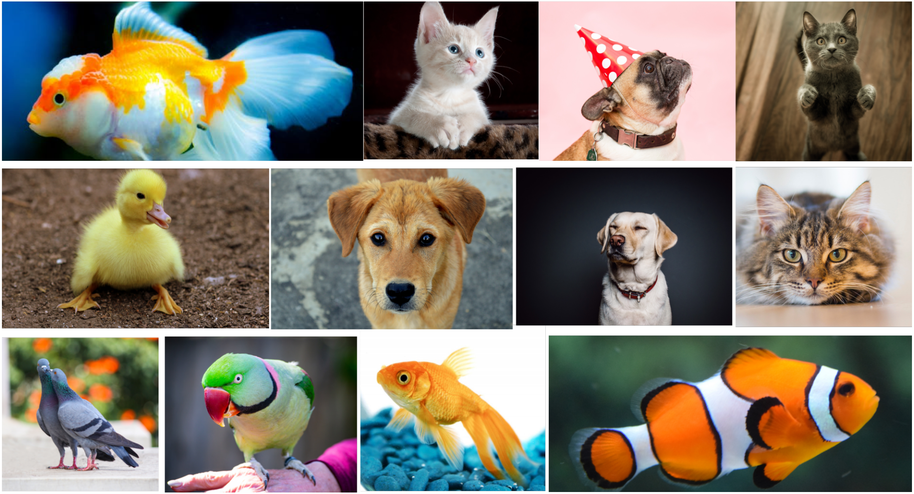
Siva Balakrishnan

Andrew Nobel

Rebecca Nugent

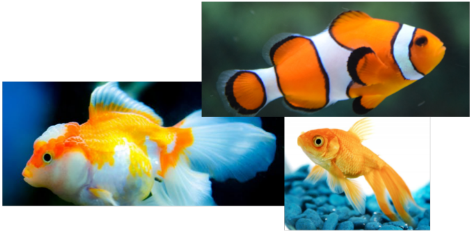
Alessandro Rinaldo

What is clustering?



What is clustering?

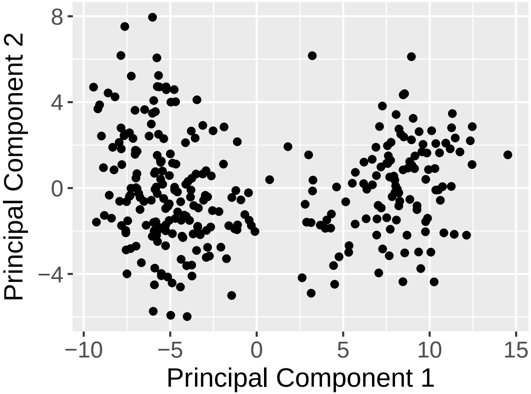
What is clustering?



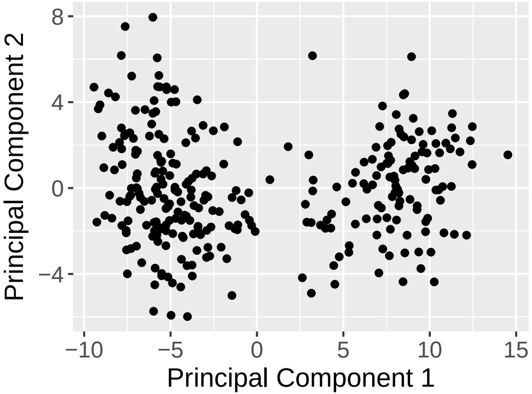
What is clustering?

"Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)."

How many clusters are “really” there?

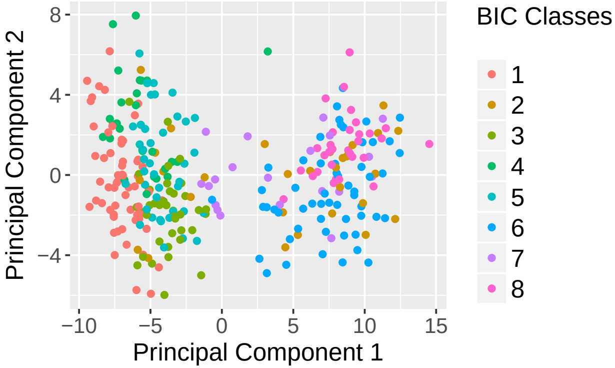


How many clusters are “really” there?



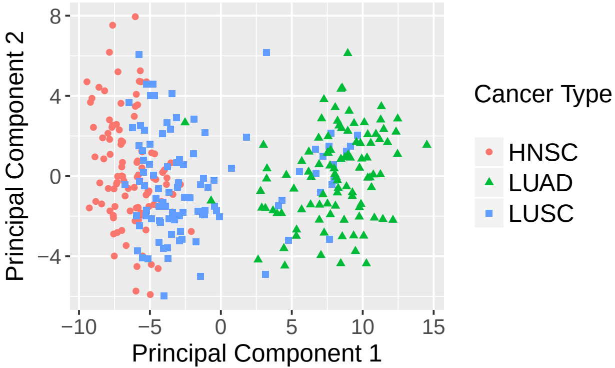
Popular answers: AIC, BIC, gap statistic (Tibshirani et al. (2001)), Hartigan index (Hartigan (1975)), the silhoutte statistic (Rousseeuw (1987)), Ghosh and Sen (1984), Milligan and Cooper (1985), Bock (1985), McLachlan and Peel (2000), Fraley and Raftery (2002), McLachlan and Peel (2004), McLachlan and Rathnayake (2014), ...

How many clusters are “really” there?

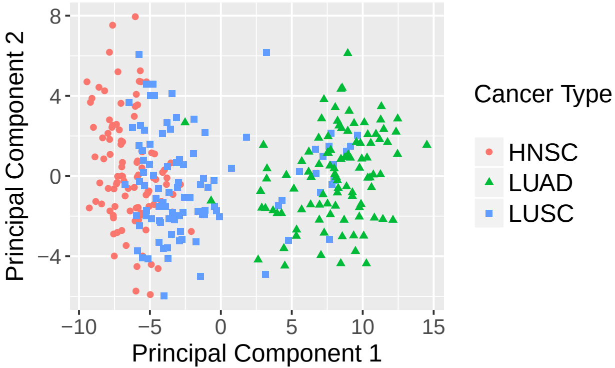


Popular answers: AIC, BIC, gap statistic (Tibshirani et al. (2001)), Hartigan index (Hartigan (1975)), the silhoutte statistic (Rousseeuw (1987)), Ghosh and Sen (1984), Milligan and Cooper (1985), Bock (1985), McLachlan and Peel (2000), Fraley and Raftery (2002), McLachlan and Peel (2004), McLachlan and Rathnayake (2014), ...

Eg: The Cancer Genome Atlas (TCGA) project.



Eg: The Cancer Genome Atlas (TCGA) project.



- RNA sequence data from 3 types of cancer (Network et al. (2012), Network et al. (2014)).
- Head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD).

Introduction: Gaussian Mixture Models.

- If $Y \in \mathbb{R}^d \sim p$ and p_k is the density of $N(\mu_k, \Sigma_k)$, then for $\mathbf{y} \in \mathbb{R}^d$,

$$p(\mathbf{y}|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{y}|\mu_k, \Sigma_k),$$

where π_k are the mixing proportions $(0 < \pi_k < 1, \sum_k \pi_k = 1)$.

Introduction: Gaussian Mixture Models.

- If $Y \in \mathbb{R}^d \sim p$ and p_k is the density of $N(\mu_k, \Sigma_k)$, then for $\mathbf{y} \in \mathbb{R}^d$,

$$p(\mathbf{y}|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{y}|\mu_k, \Sigma_k),$$

  where π_k are the mixing proportions $(0 < \pi_k < 1, \sum_k \pi_k = 1)$.

- Choosing K, requires some sort of testing or model selection.

Introduction: Gaussian Mixture Models.

- If $Y \in \mathbb{R}^d \sim p$ and p_k is the density of $N(\mu_k, \Sigma_k)$, then for $\mathbf{y} \in \mathbb{R}^d$,

$$p(\mathbf{y}|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{y}|\mu_k, \Sigma_k),$$

  where π_k are the mixing proportions ($0 < \pi_k < 1, \sum_k \pi_k = 1$).

- Choosing K, requires some sort of testing or model selection.

- Natural fix: Test "Gaussian" vs "a mixture of two Gaussians" using the likelihood ratio test.

Introduction: Gaussian Mixture Models.

- If $Y \in \mathbb{R}^d \sim p$ and p_k is the density of $N(\mu_k, \Sigma_k)$, then for $\mathbf{y} \in \mathbb{R}^d$,

$$p(\mathbf{y}|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{y}|\mu_k, \Sigma_k),$$

  where π_k are the mixing proportions ($0 < \pi_k < 1, \sum_k \pi_k = 1$).

- Choosing K, requires some sort of testing or model selection.

- Natural fix: Test "Gaussian" vs "a mixture of two Gaussians" using the likelihood ratio test.

- But usual regularity conditions fail for mixture models (Ghosh and Sen (1984); McLachlan and Rathnayake (2014); Dacunha-Castelle et al. (1999)).

SigClust: How it works!

Proposed by Liu, Hayes, Nobel and Marron (2008) (Liu et al., 2008)

SigClust: How it works!

Proposed by Liu, Hayes, Nobel and Marron (2008) (Liu et al., 2008)

1. If $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$.

$$H_0 : X_1, \ldots, X_n \sim N(\mu, \Sigma) \text{ versus}$$
$$H_1 : X_1, \ldots, X_n \sim f(\cdot), \quad \text{which is a non-Gaussian distribution.}$$

SigClust: How it works!

Proposed by Liu, Hayes, Nobel and Marron (2008) (Liu et al., 2008)

1. If $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$.

$$H_0 : X_1, \ldots, X_n \sim N(\mu, \Sigma) \text{ versus}$$
$$H_1 : X_1, \ldots, X_n \sim f(\cdot), \quad \text{which is a non-Gaussian distribution.}$$

2. Performs $2-$means clustering and uses Cluster Index as the test statistic.

$$CI = \frac{\sum_{k=1}^{2} \sum_{j \in C_k} ||X_j - \overline{X}^k||^2}{\sum_{j=1}^{n} ||X_j - \overline{X}||^2},$$

C_k: k^{th} cluster and $\overline{X}^k$: k^{th} cluster mean.

SigClust: How it works!

Proposed by Liu, Hayes, Nobel and Marron (2008) (Liu et al., 2008)

1. If $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$.

$$H_0 : X_1, \ldots, X_n \sim N(\mu, \Sigma) \text{ versus}$$
$$H_1 : X_1, \ldots, X_n \sim f(\cdot), \quad \text{which is a non-Gaussian distribution.}$$

2. Performs 2−means clustering and uses Cluster Index as the test statistic.

$$CI = \frac{\sum_{k=1}^{2} \sum_{j \in C_k} ||X_j - \overline{X}^k||^2}{\sum_{j=1}^{n} ||X_j - \overline{X}||^2},$$

C_k: k^{th} cluster and $\overline{X}^k$: k^{th} cluster mean.

3. Computes the distribution of the CI under H_0 and the p-value.

What does SigClust do?

Find CI_{data}

↓

Estimate $\hat{\Sigma}$

↓

$$Y_1^{(s)}, Y_2^{(s)}, \ldots, Y_m^{(s)} \sim N\left(0, \hat{\Sigma}\right), \quad s = 1, \ldots, N_{sim}$$

↓

Using $\{Y_i^{(s)}\}$ find CI_s, $s = 1, \ldots, N_{sim}$

↓

$$\text{p-value} = \frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} I\left\{CI_s < CI_{data}\right\}$$

What does SigClust do?

Find CI_{data}

↓

Estimate $\hat{\Sigma}$

↓

$Y_1^{(s)}, Y_2^{(s)}, \ldots, Y_m^{(s)} \sim N\left(0, \hat{\Sigma}\right), \quad s = 1, \ldots, N_{sim}$

↓

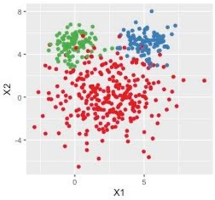
Using $\{Y_i^{(s)}\}$ find CI_s, $s = 1, \ldots, N_{sim}$

↓

$$\text{p-value} = \frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} I\left\{CI_s < CI_{data}\right\}$$

Note: Considers HDLSS data and estimates the covariance matrix in high dimensions under H_0. A difficult task!

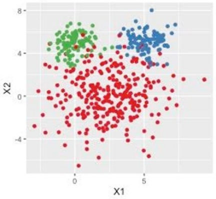
Hierarchical SigClust: Mickey Mouse Example



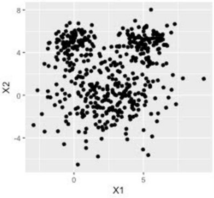
Original

$$X_1, X_2 \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

Hierarchical SigClust: Mickey Mouse Example



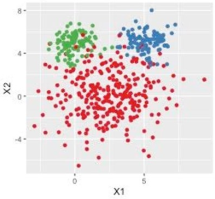
Original



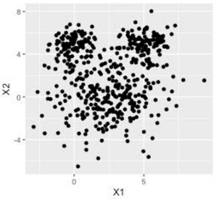
Split?

$$X_1, X_2 \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

Hierarchical SigClust: Mickey Mouse Example



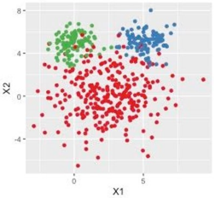
Original



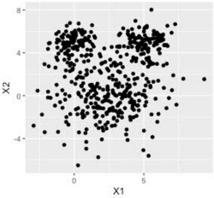
Yes

$$X_1, X_2. \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

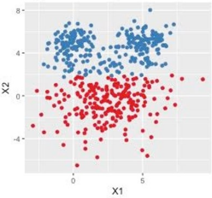
Hierarchical SigClust: Mickey Mouse Example



Original

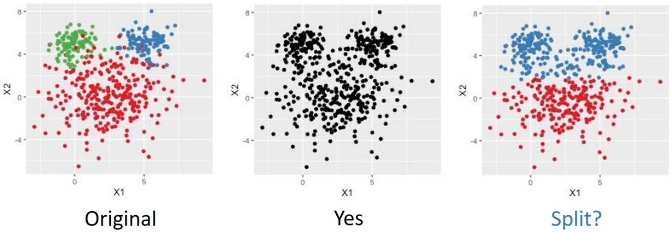


Yes



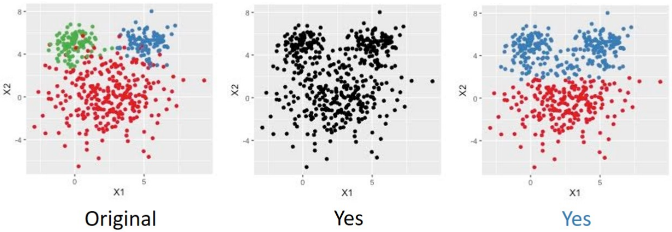
$$X_1, X_2. \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

Hierarchical SigClust: Mickey Mouse Example



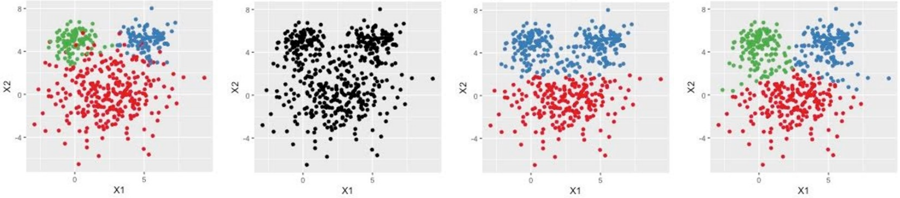
$$X_1, X_2 \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

Hierarchical SigClust: Mickey Mouse Example



$$X_1, X_2, \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

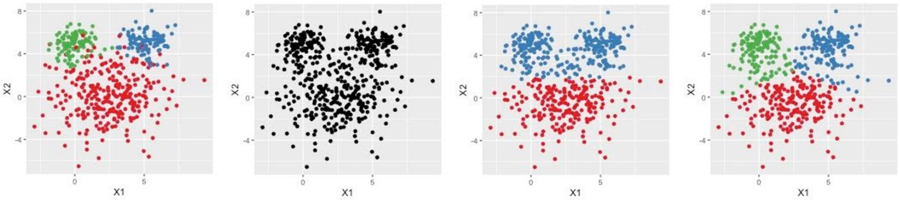
Hierarchical SigClust: Mickey Mouse Example



Original   Yes   Yes

$$X_1, X_2 \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

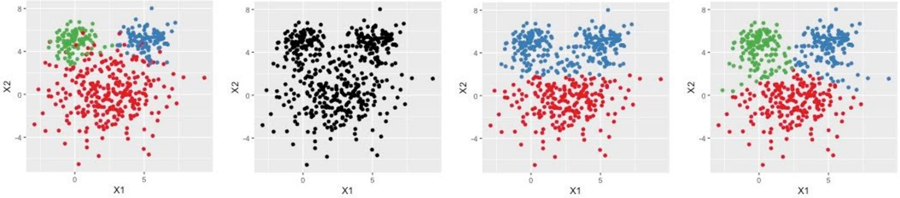
Hierarchical SigClust: Mickey Mouse Example



Original Yes Yes Split?

$$X_1, X_2. \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

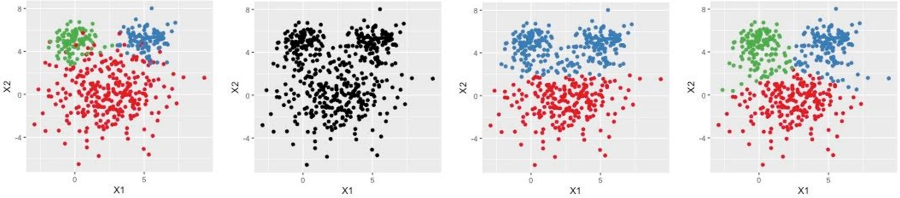
Hierarchical SigClust: Mickey Mouse Example



Original | Yes | Yes | No

$$X_1, X_2 \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

Hierarchical SigClust: Mickey Mouse Example



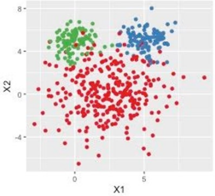
Original | Yes | Yes | Split?

$$X_1, X_2. \dots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

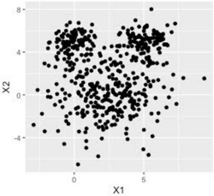
Hierarchical SigClust: Mickey Mouse Example

$$X_1, X_2, \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

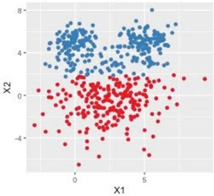
Hierarchical SigClust: Mickey Mouse Example



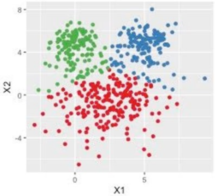
Original



Yes



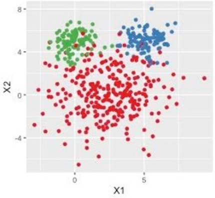
Yes



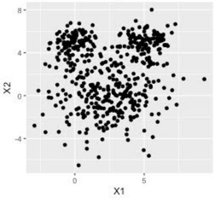
Split?

$$X_1, X_2 \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

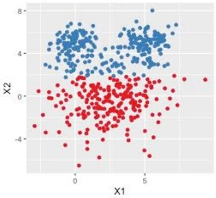
Hierarchical SigClust: Mickey Mouse Example



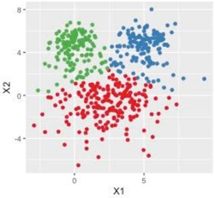
Original



Yes



Yes



No

$$X_1, X_2 \ldots, X_n \sim w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

Power of SigClust: Low power in some cases.

Theorem 1

$X_1, \ldots, X_n \sim \frac{1}{2}N(-\mu, \Sigma) + \frac{1}{2}N(\mu, \Sigma)$, $\mu = \left(\frac{a}{2}, 0, \ldots, 0\right)$,

Power of SigClust: Low power in some cases.

Theorem 1

$X_1, \ldots, X_n \sim \frac{1}{2}N(-\mu, \Sigma) + \frac{1}{2}N(\mu, \Sigma)$, $\mu = \left(\frac{a}{2}, 0, \ldots, 0\right)$, *and* Σ *is diagonal* $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$.

Power of SigClust: Low power in some cases.

Theorem 1

$X_1, \ldots, X_n \sim \frac{1}{2}N(-\mu, \Sigma) + \frac{1}{2}N(\mu, \Sigma)$, $\mu = \left(\frac{a}{2}, 0, \ldots, 0\right)$, *and* Σ *is diagonal* $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$.

- *If* $\sigma_2^2 < \frac{\pi}{2}\mathbb{E}[X_{i1}|X_{i1} > 0]^2$, *then* $\mathrm{Power}_n(a) \rightarrow 1$ *as* $n \rightarrow \infty$.

Power of SigClust: Low power in some cases.

Theorem 1

$X_1, \ldots, X_n \sim \frac{1}{2}N(-\mu, \Sigma) + \frac{1}{2}N(\mu, \Sigma)$, $\mu = \left(\frac{a}{2}, 0, \ldots, 0\right)$, *and* Σ *is diagonal* $\sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2$.

- *If* $\sigma_2^2 < \frac{\pi}{2}\mathbb{E}[X_{i1}|X_{i1} > 0]^2$, *then* $\mathrm{Power}_n(a) \to 1$ *as* $n \to \infty$.
- *If* $\sigma_2^2 > \frac{\pi}{2}\mathbb{E}[X_{i1}|X_{i1} > 0]^2$, *then* $\lim_{n\to\infty} \mathrm{Power}_n(a) < 1$,

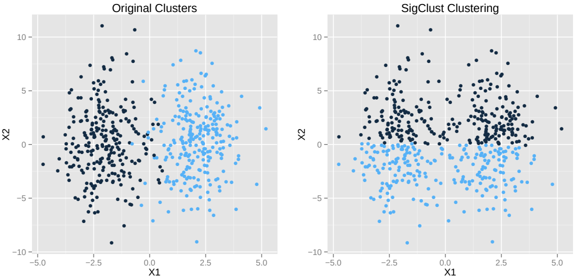
where $\frac{\pi}{2}\mathbb{E}[X_{i1}|X_{i1} > 0]^2 \approx \sigma_1^2 + \frac{a^2}{4}$ for small a.

SigClust fails to detect clusters!

$$X_1, \ldots, X_n \sim \frac{1}{2}N(-\mu, \Sigma) + \frac{1}{2}N(\mu, \Sigma),$$

where $\mu = \left(\frac{a}{2}, 0, \ldots, 0\right) \in \mathbb{R}^2$ and Σ is diagonal with entries σ_1^2 and σ_2^2.

If $\frac{\pi}{2}\mathbb{E}[X_{i1}|X_{i1} > 0]^2 < \sigma_2^2$, k-means optimal split, splits horizontally!



Proposed Test: Relative Information Fit Test (RIFT)

- Test if a mixture of Normals **fits** the data significantly better than a single Normal.

Proposed Test: Relative Information Fit Test (RIFT)

- Test if a mixture of Normals **fits** the data significantly better than a single Normal.
- Randomly split data into D_1 (Estimating) and D_2 (Testing).

Proposed Test: Relative Information Fit Test (RIFT)

- Test if a mixture of Normals **fits** the data significantly better than a single Normal.
- Randomly split data into D_1 (Estimating) and D_2 (Testing).
- Using D_1, fit a Normal $\hat{p}_1$ and a mixture of two Normals $\hat{p}_2$.

Proposed Test: Relative Information Fit Test (RIFT)

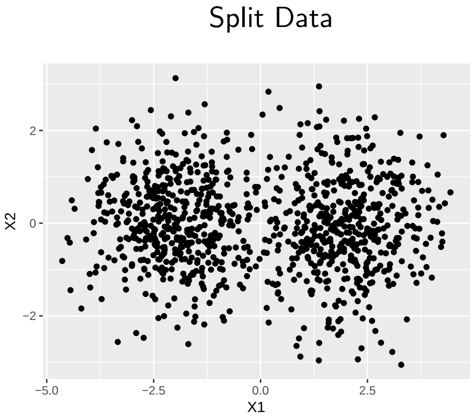
- Test if a mixture of Normals **fits** the data significantly better than a single Normal.
- Randomly split data into D_1 (Estimating) and D_2 (Testing).
- Using D_1, fit a Normal $\hat{p}_1$ and a mixture of two Normals $\hat{p}_2$.
- $\Gamma = K(p, \hat{p}_1) - K(p, \hat{p}_2)$, where K is the Kullback-Leibler distance and p is the true density.

Proposed Test: Relative Information Fit Test (RIFT)

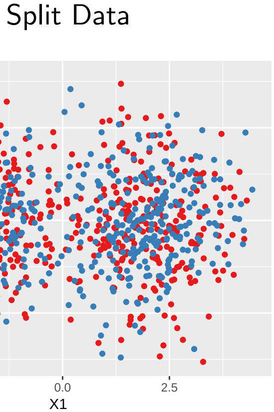
- Test if a mixture of Normals **fits** the data significantly better than a single Normal.
- Randomly split data into D_1 (Estimating) and D_2 (Testing).
- Using D_1, fit a Normal $\hat{p}_1$ and a mixture of two Normals $\hat{p}_2$.
- $\Gamma = K(p, \hat{p}_1) - K(p, \hat{p}_2)$, where K is the Kullback-Leibler distance and p is the true density.
- We test, conditional on D_1, using D_2

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

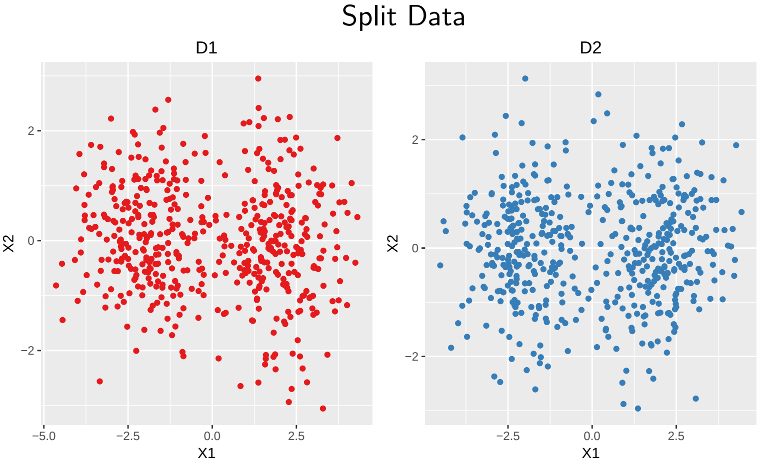
Relative Information Fit Test (RIFT): How it works!



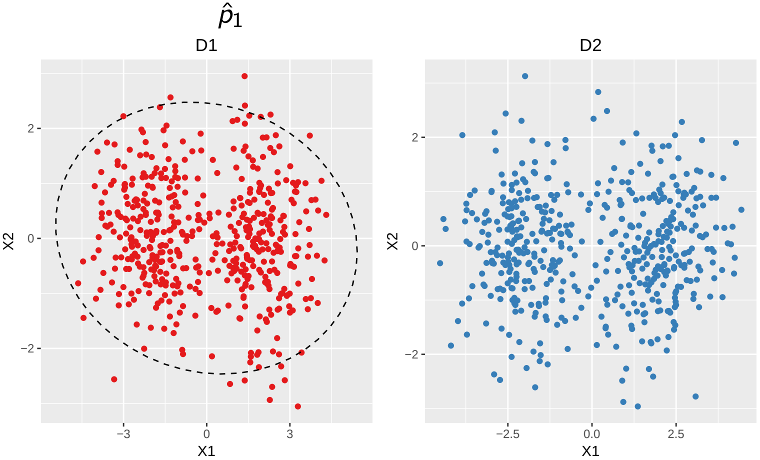
Relative Information Fit Test (RIFT): How it works!



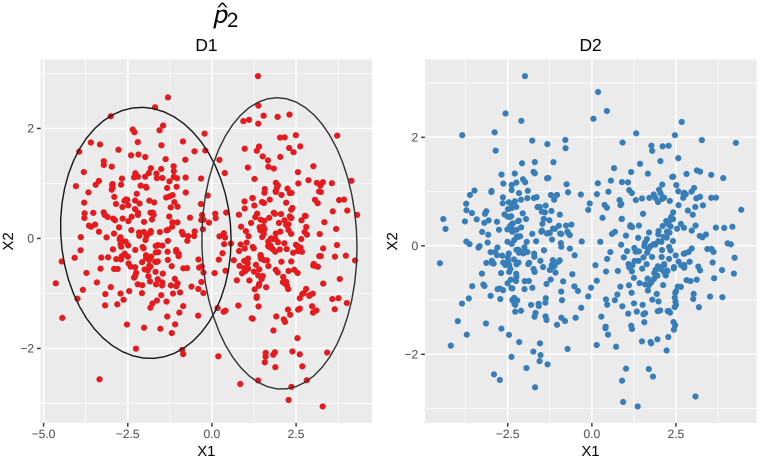
Relative Information Fit Test (RIFT): How it works!



Relative Information Fit Test (RIFT): How it works!

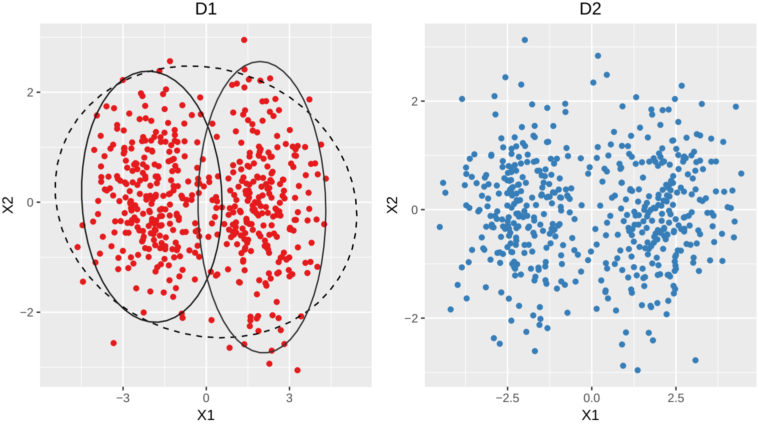


Relative Information Fit Test (RIFT): How it works!



Relative Information Fit Test (RIFT): How it works!

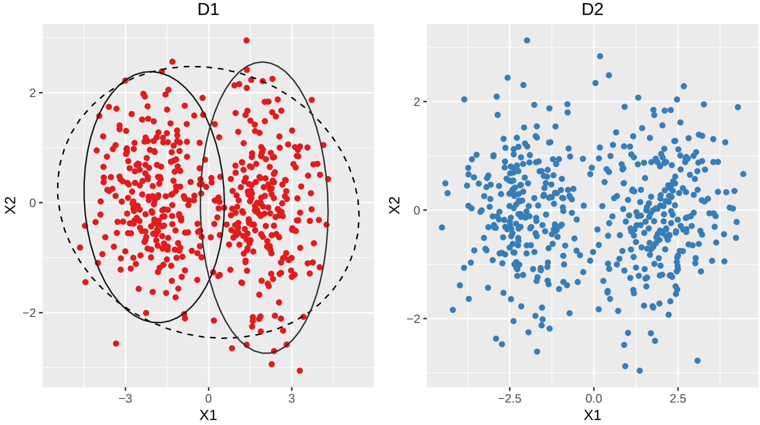
$\hat{p}_1, \ \hat{p}_2$



Conditioned on D_1, $H_0 : \Gamma \leq 0$ versus $H_1 : \Gamma > 0$.

Relative Information Fit Test (RIFT): How it works!

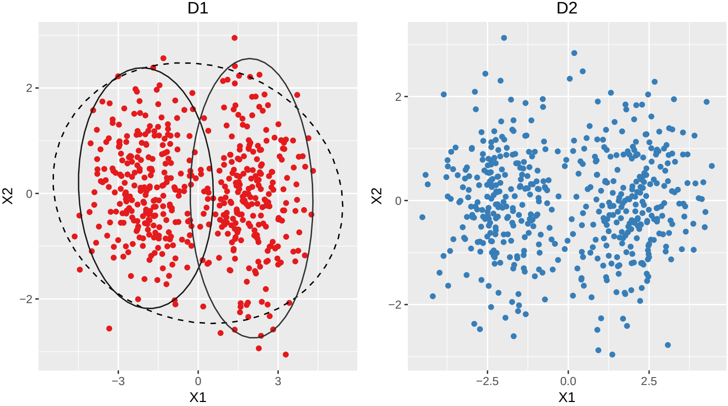
$\hat{p}_1, \ \hat{p}_2$ $\qquad$ $\hat{\Gamma} = \frac{1}{n} \sum_{i \in D_2} R_i, \ \ R_i = \log\left(\frac{\hat{p}_2(X_i)}{\hat{p}_1(X_i)}\right)$



Conditioned on D_1, $H_0 : \Gamma \leq 0$ versus $H_1 : \Gamma > 0$.

Relative Information Fit Test (RIFT): How it works!

$\hat{p}_1, \ \hat{p}_2 \qquad \hat{\Gamma} = \frac{1}{n} \sum_{i \in D_2} R_i, \ \ R_i = \log \left(\frac{\hat{p}_2(X_i)}{\hat{p}_1(X_i)} \right)$



Conditioned on D_1, $H_0 : \Gamma \leq 0$ versus $H_1 : \Gamma > 0$.

$$\sqrt{n}\left(\hat{\Gamma} - \Gamma\right) \rightsquigarrow N(0, \tau^2) \implies \text{Reject } H_0 \text{ if } \hat{\Gamma} > \frac{z_\alpha \hat{\tau}}{\sqrt{n}}.$$

Asymptotic Normality of $\hat{\Gamma}$

- Let $\hat{p}_1 = N(\hat{\mu}_0, \hat{\Sigma}_0)$ and $\hat{p}_2 = \hat{\alpha} N(\hat{\mu}_1, \hat{\Sigma}_1) + (1 - \hat{\alpha}) N(\hat{\mu}_2, \hat{\Sigma}_2)$.

Asymptotic Normality of $\hat{\Gamma}$

- Let $\hat{p}_1 = N(\hat{\mu}_0, \hat{\Sigma}_0)$ and $\hat{p}_2 = \hat{\alpha} N(\hat{\mu}_1, \hat{\Sigma}_1) + (1 - \hat{\alpha}) N(\hat{\mu}_2, \hat{\Sigma}_2)$.

Theorem 2

Assume each $\hat{\mu}_i \in \mathcal{A}$, a compact set and the eigenvalues of $\hat{\Sigma}_i \in [c_1, c_2]$.

Asymptotic Normality of $\hat{\Gamma}$

- Let $\hat{p}_1 = N(\hat{\mu}_0, \hat{\Sigma}_0)$ and $\hat{p}_2 = \hat{\alpha} N(\hat{\mu}_1, \hat{\Sigma}_1) + (1 - \hat{\alpha}) N(\hat{\mu}_2, \hat{\Sigma}_2)$.

Theorem 2

Assume each $\hat{\mu}_i \in \mathcal{A}$, a compact set and the eigenvalues of $\hat{\Sigma}_i \in [c_1, c_2]$. Let $Z \sim N(0, \tau^2)$ where $\tau^2 = \mathbb{E}[(\tilde{R}_i - \Gamma)^2 | \mathcal{D}_1]$. Then, under H_0

$$\sup_t \left| P(\sqrt{n}(\hat{\Gamma} - \Gamma) \leq t \,|\, \mathcal{D}_1) - P(Z \leq t) \right| \leq \frac{C}{\sqrt{n}} \tag{1}$$

Asymptotic Normality of $\hat{\Gamma}$

- Let $\hat{p}_1 = N(\hat{\mu}_0, \hat{\Sigma}_0)$ and $\hat{p}_2 = \hat{\alpha} N(\hat{\mu}_1, \hat{\Sigma}_1) + (1 - \hat{\alpha}) N(\hat{\mu}_2, \hat{\Sigma}_2)$.

Theorem 2

Assume each $\hat{\mu}_i \in \mathcal{A}$, a compact set and the eigenvalues of $\hat{\Sigma}_i \in [c_1, c_2]$. Let $Z \sim N(0, \tau^2)$ where $\tau^2 = \mathbb{E}[(\tilde{R}_i - \Gamma)^2 | \mathcal{D}_1]$. Then, under H_0

$$\sup_t \left| P(\sqrt{n}(\hat{\Gamma} - \Gamma) \leq t \mid \mathcal{D}_1) - P(Z \leq t) \right| \leq \frac{C}{\sqrt{n}} \tag{1}$$

where C is a constant that does not depend on $\mathcal{D}_1$.

Power of RIFT converges to 1

Power converges to 1!

$\mathcal{P}_1$: Normals, $\mathcal{P}_2$: mixtures of two Normals.

Lemma 3

Suppose that $p \in \mathcal{P}_2 - \mathcal{P}_1$. *Then* $P(\hat{\Gamma} > z_\alpha \hat{\tau}/\sqrt{n}) \to 1$ *as* $n \to \infty$.

Aside: A Test for Mixtures

$\mathcal{P}_1$: Normals, $\mathcal{P}_2$: mixtures of two Normals. $\hat{p}_1 \in \mathcal{P}_1, \hat{p}_2 \in \mathcal{P}_2$.

Earlier:

$$H_0 : K(p, \hat{p}_1) - K(p, \hat{p}_2) \leq 0 \quad \text{vs} \quad H_1 : K(p, \hat{p}_1) - K(p, \hat{p}_2) > 0.$$

Now:

$$H_0 : p \in \mathcal{P}_1 \quad \text{vs} \quad H_1 : p \in \mathcal{P}_2,$$

Aside: A Test for Mixtures

$\mathcal{P}_1$: Normals, $\mathcal{P}_2$: mixtures of two Normals. $\hat{p}_1 \in \mathcal{P}_1, \hat{p}_2 \in \mathcal{P}_2$.

Earlier:

$$H_0 : K(p, \hat{p}_1) - K(p, \hat{p}_2) \leq 0 \quad \text{vs} \quad H_1 : K(p, \hat{p}_1) - K(p, \hat{p}_2) > 0.$$

Now:

$$H_0 : p \in \mathcal{P}_1 \quad \text{vs} \quad H_1 : p \in \mathcal{P}_2,$$

Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \ \forall \ p \in \mathcal{P}_1$, where $\Delta > 0$ small constant (Ghosh and Sen (1984) separation idea).

Previous test, that rejects H_0 when $\hat{\Gamma} > z_\alpha \hat{\tau}/\sqrt{n}$ is a valid level α test!

Aside: A Test for Mixtures

$\mathcal{P}_1$: Normals, $\mathcal{P}_2$: mixtures of two Normals. $\hat{p}_1 \in \mathcal{P}_1, \hat{p}_2 \in \mathcal{P}_2$.

Earlier:

$$H_0 : K(p, \hat{p}_1) - K(p, \hat{p}_2) \leq 0 \quad \text{vs} \quad H_1 : K(p, \hat{p}_1) - K(p, \hat{p}_2) > 0.$$

Now:

$$H_0 : p \in \mathcal{P}_1 \quad \text{vs} \quad H_1 : p \in \mathcal{P}_2,$$

Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \ \forall \ p \in \mathcal{P}_1$, where $\Delta > 0$ small constant (Ghosh and Sen (1984) separation idea).

Theorem 4

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

Aside: A Test for Mixtures

$\mathcal{P}_1$: Normals, $\mathcal{P}_2$: mixtures of two Normals. $\hat{p}_1 \in \mathcal{P}_1, \hat{p}_2 \in \mathcal{P}_2$.

Earlier:

$$H_0 : K(p, \hat{p}_1) - K(p, \hat{p}_2) \leq 0 \quad \text{vs} \quad H_1 : K(p, \hat{p}_1) - K(p, \hat{p}_2) > 0.$$

Now:

$$H_0 : p \in \mathcal{P}_1 \quad \text{vs} \quad H_1 : p \in \mathcal{P}_2,$$

Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \ \forall \ p \in \mathcal{P}_1$, where $\Delta > 0$ small constant (Ghosh and Sen (1984) separation idea).

Theorem 4

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

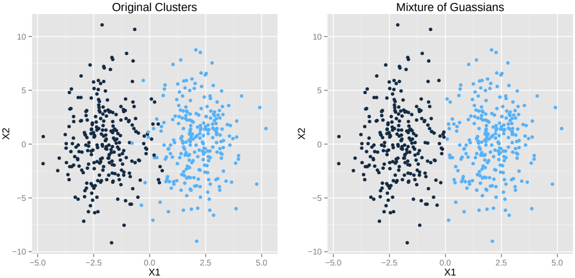
Note: Simpler test compared to existing tests. Eg: Gassiat (2002) Gassiat (2002), Dacunha-Castelle et al. (1999) Dacunha-Castelle et al. (1999), Chen (2017) Chen (2017), ...

RIFT works in the previous example

$$X_1, \ldots, X_n \sim \frac{1}{2}N(-\mu, \Sigma) + \frac{1}{2}N(\mu, \Sigma),$$

where $\mu = \left(\frac{a}{2}, 0, \ldots, 0\right) \in \mathbb{R}^d$.

$\frac{\pi}{2}\mathbb{E}[X_{i1}|X_{i1} > 0]^2 < \sigma_2^2, d = 2.$



Median RIFT (M-RIFT): A more robust test.

- $\Gamma = \mathbb{E}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.

Median RIFT (M-RIFT): A more robust test.

- $\Gamma = \mathbb{E}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.
- Robustified version: $\tilde{\Gamma} = \mathrm{Median}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.

Median RIFT (M-RIFT): A more robust test.

- $\Gamma = \mathbb{E}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.
- Robustified version: $\tilde{\Gamma} = \mathrm{Median}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.
- Sample median of $R_1, \ldots, R_n$ is a consistent estimator, where $R_i = \log \hat{p}_2(X_i)/\hat{p}_1(X_i)$.

Median RIFT (M-RIFT): A more robust test.

- $\Gamma = \mathbb{E}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.
- Robustified version: $\tilde{\Gamma} = \mathrm{Median}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.
- Sample median of $R_1, \ldots, R_n$ is a consistent estimator, where $R_i = \log \hat{p}_2(X_i)/\hat{p}_1(X_i)$.
- Test $H_0 : \tilde{\Gamma} \leq 0$ versus $H_1 : \tilde{\Gamma} > 0$ using the sign test.

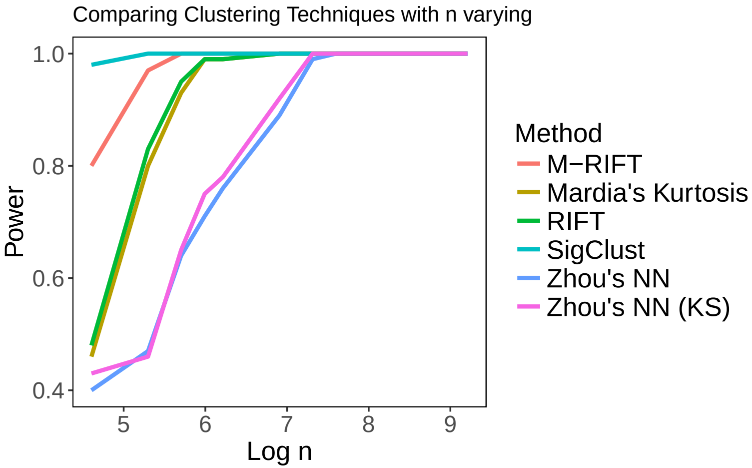
Median RIFT (M-RIFT): A more robust test.

- $\Gamma = \mathbb{E}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.
- Robustified version: $\tilde{\Gamma} = \mathrm{Median}_p[R]$, where $R = \log \hat{p}_2(X)/\hat{p}_1(X)$.
- Sample median of $R_1, \ldots, R_n$ is a consistent estimator, where $R_i = \log \hat{p}_2(X_i)/\hat{p}_1(X_i)$.
- Test $H_0 : \tilde{\Gamma} \leq 0$ versus $H_1 : \tilde{\Gamma} > 0$ using the sign test.
- Replace KL distance with its median version. Gives an exact test!

Comparisions for 2 Normals

$$X_1, \ldots, X_n \sim 0.5N(\mu, I_d) + 0.5N(-\mu, I_d) \text{ where } \mu = (a, 0, \ldots, 0)$$

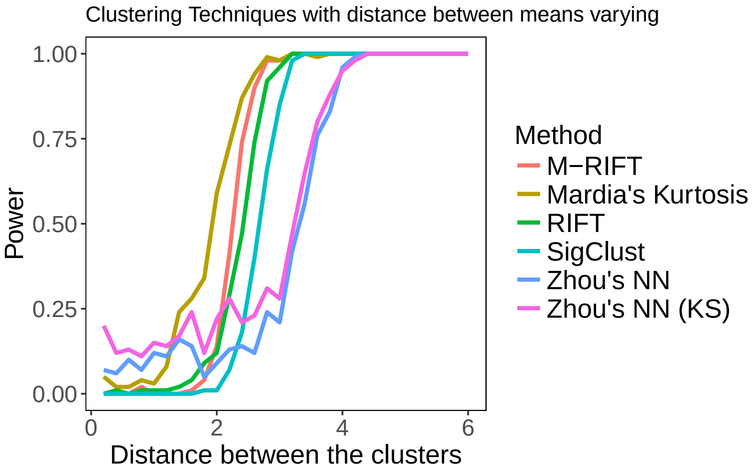
SigClust performs better than RIFTs.



Comparisions for 2 Normals

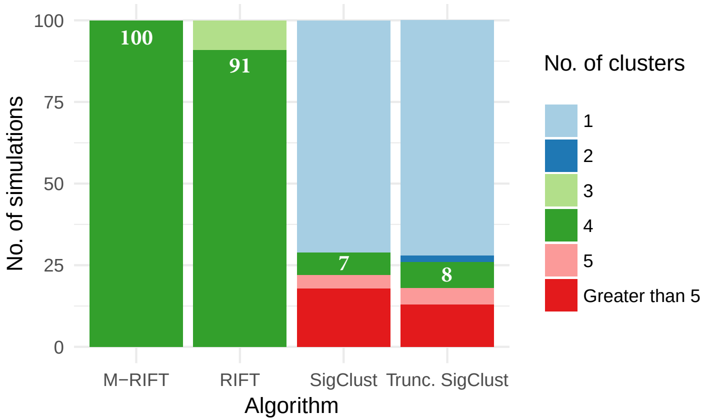
$$X_1, \ldots, X_n \sim 0.5N(\mu, I_d) + 0.5N(-\mu, I_d) \text{ where } \mu = (a, 0, \ldots, 0)$$

RIFTs perform better than SigClust.



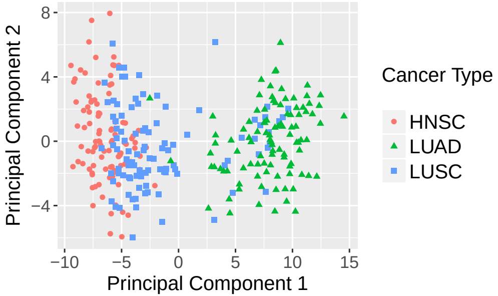
4 Normals: Hierarchical SigClust and RIFT

- $X_1, \ldots, X_n \sim$ 4 Normals at vertices of a regular tetrahedron with side $\delta = 5$ in $\mathbb{R}^3$.
- 50 samples from each. 100 simulations. $\alpha = 0.05$.



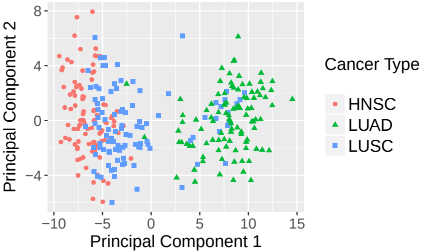
TCGA project: Multi-Cancer Gene Expression Dataset

- RNA sequence data from 3 types of cancer (Network et al. (2012), Network et al. (2014)).
- Head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD).
- 300 samples: 100 from each of HNSC, LUSC and LUAD.



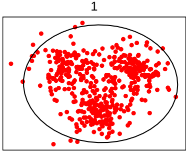
TCGA project: Multi-Cancer Gene Expression Dataset

1. RIFTs: 3 clusters.
2. SigClust: 9 clusters.
3. AIC: 12, BIC: 8.

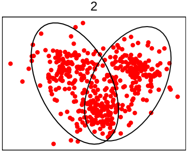


Hierarchical RIFT (H-RIFT)

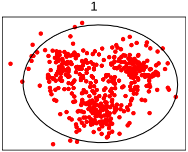
Hierarchical RIFT (H-RIFT)



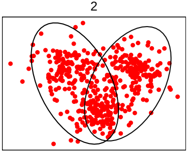
$\hat{p}_1$ vs $\hat{p}_2$



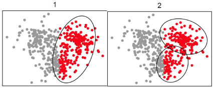
Hierarchical RIFT (H-RIFT)



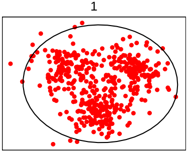
$\hat{p}_1$ vs $\hat{p}_2$



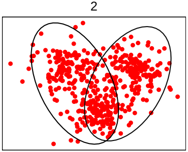
$\hat{p}_1$ vs $\hat{p}_2$



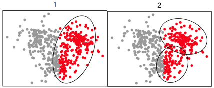
Hierarchical RIFT (H-RIFT)



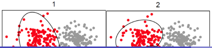
$\hat{p}_1$ vs $\hat{p}_2$



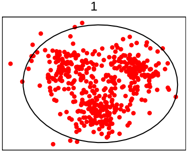
$\hat{p}_1$ vs $\hat{p}_2$



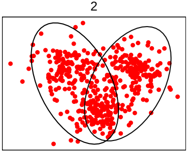
$\hat{p}_1$ vs $\hat{p}_2$



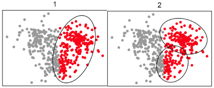
Hierarchical RIFT (H-RIFT)



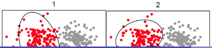
$\hat{p}_1$ vs $\hat{p}_2$



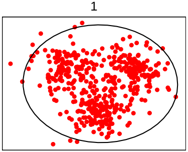
$\hat{p}_1$ vs $\hat{p}_2$



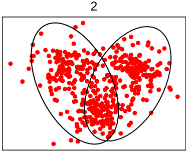
$\hat{p}_1$ vs $\hat{p}_2$



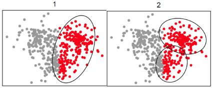
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)



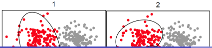
$\hat{p}_1$ vs $\hat{p}_2$



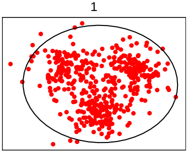
$\hat{p}_1$ vs $\hat{p}_2$



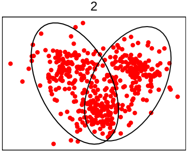
$\hat{p}_1$ vs $\hat{p}_2$



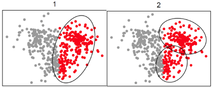
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)



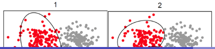
$\hat{p}_1$ vs $\hat{p}_2$



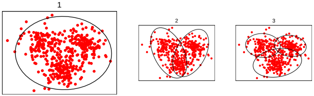
$\hat{p}_1$ vs $\hat{p}_2$



$\hat{p}_1$ vs $\hat{p}_2$

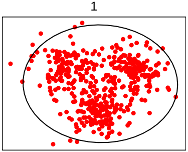


$\hat{p}_1$ vs $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n}$

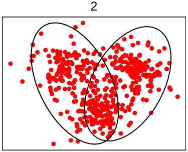


...

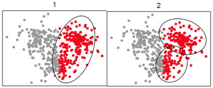
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)



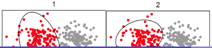
$\hat{p}_1$ vs $\hat{p}_2$



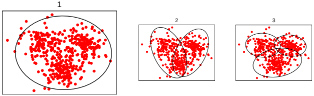
$\hat{p}_1$ vs $\hat{p}_2$



$\hat{p}_1$ vs $\hat{p}_2$

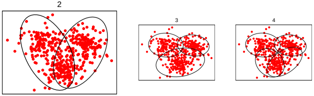


$\hat{p}_1$ vs $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n}$



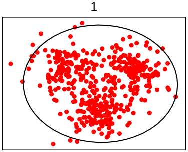
$\ldots$

$\hat{p}_2$ vs $\hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n}$

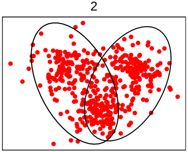


$\ldots$

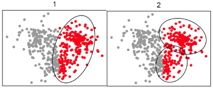
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)



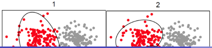
$\hat{p}_1$ vs $\hat{p}_2$



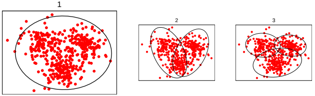
$\hat{p}_1$ vs $\hat{p}_2$



$\hat{p}_1$ vs $\hat{p}_2$

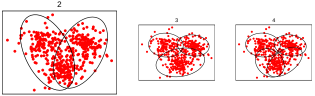


$\hat{p}_1$ vs $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n}$



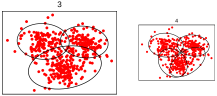
. . .

$\hat{p}_2$ vs $\hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n}$



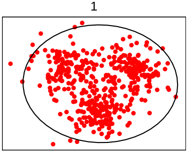
. . .

$\hat{p}_3$ vs $\hat{p}_4, \ldots, \hat{p}_{K_n}$

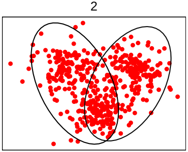


. . .

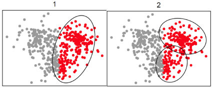
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)



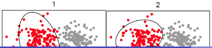
$\hat{p}_1$ vs $\hat{p}_2$



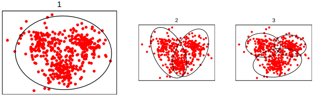
$\hat{p}_1$ vs $\hat{p}_2$



$\hat{p}_1$ vs $\hat{p}_2$

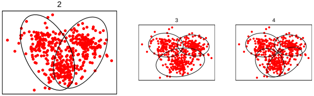


$\hat{p}_1$ vs $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n}$



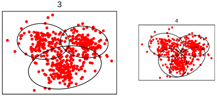
$\ldots$

$\hat{p}_2$ vs $\hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n}$

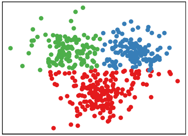


$\ldots$

$\hat{p}_3$ vs $\hat{p}_4, \ldots, \hat{p}_{K_n}$



$\ldots$



Validity of S-RIFT

Unlike AIC or BIC, provides a valid, asymptotic, type I error control.

Lemma 5

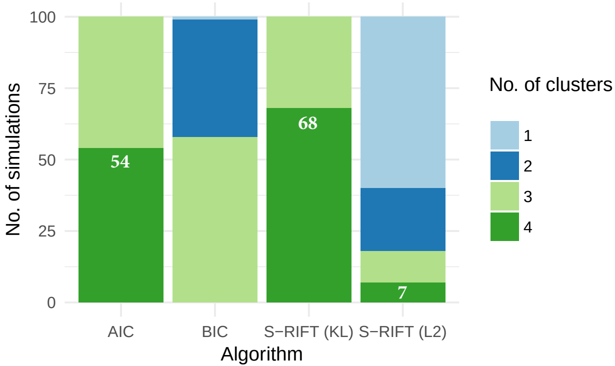
Under H_{0j},

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

Note: Can be used with L_2 distance or Median version of KL distance.

4 Normals: Comparing S-RIFT to AIC and BIC

- $X_1, \ldots, X_n \sim$ 4 Normals at vertices of a regular tetrahedron with side $\delta = 6$ in $\mathbb{R}^{10}$.
- 100 samples from each. 100 simulations. $\alpha = 0.05$.



Summary

- RIFTs - simple and easy tests to detect significant clusters.

Summary

- RIFTs - simple and easy tests to detect significant clusters.
- It can be applied both hierarchically and sequentially, while asymptotically controlling for type I error.

Summary

- RIFTs - simple and easy tests to detect significant clusters.
- It can be applied both hierarchically and sequentially, while asymptotically controlling for type I error.
- For very close clusters or if variance in other directions is higher - RIFTs perform better than SigClust.

Summary

- RIFTs - simple and easy tests to detect significant clusters.
- It can be applied both hierarchically and sequentially, while asymptotically controlling for type I error.
- For very close clusters or if variance in other directions is higher - RIFTs perform better than SigClust.
- HDLSS - SigClust performs better.

Summary

- RIFTs - simple and easy tests to detect significant clusters.
- It can be applied both hierarchically and sequentially, while asymptotically controlling for type I error.
- For very close clusters or if variance in other directions is higher - RIFTs perform better than SigClust.
- HDLSS - SigClust performs better.
- In a hierarchical setting, RIFTs perform better.

Future Work

- Apply the Ghosh-Sen separation idea in practice.
  - Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \; \forall \; p \in \mathcal{P}_1$.

Future Work

- Apply the Ghosh-Sen separation idea in practice.
  - Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \ \forall \ p \in \mathcal{P}_1$.
- Improve the performance of RIFTs in higher dimensions.

Future Work

- Apply the Ghosh-Sen separation idea in practice.
  - Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \; \forall \; p \in \mathcal{P}_1$.
- Improve the performance of RIFTs in higher dimensions.
- Analyze performance of hierarchical RIFTs theoretically.

Future Work

- Apply the Ghosh-Sen separation idea in practice.
  - Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \; \forall \; p \in \mathcal{P}_1$.
- Improve the performance of RIFTs in higher dimensions.
- Analyze performance of hierarchical RIFTs theoretically.
- Study why S-RIFT with L_2 distance performs poorly.

Future Work

- Apply the Ghosh-Sen separation idea in practice.
  - Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \; \forall \; p \in \mathcal{P}_1$.
- Improve the performance of RIFTs in higher dimensions.
- Analyze performance of hierarchical RIFTs theoretically.
- Study why S-RIFT with L_2 distance performs poorly.
- Explore power of RIFT for higher dimensions ($d \to \infty$).

Future Work

- Apply the Ghosh-Sen separation idea in practice.
  - Constrain $\hat{p}_2$ s.t. $K(p, \hat{p}_2) > \Delta \; \forall \; p \in \mathcal{P}_1$.
- Improve the performance of RIFTs in higher dimensions.
- Analyze performance of hierarchical RIFTs theoretically.
- Study why S-RIFT with L_2 distance performs poorly.
- Explore power of RIFT for higher dimensions ($d \to \infty$).
- Find the minimax testing rate.

Future Work - Timeline

1. Spring 2019
   - Apply the Ghosh-Sen separation idea in practice.
   - Analyze performance of hierarchical RIFTs theoretically.
   - Study why S-RIFT with L_2 distance performs poorly.

Future Work - Timeline

1. Spring 2019
    - Apply the Ghosh-Sen separation idea in practice.
    - Analyze performance of hierarchical RIFTs theoretically.
    - Study why S-RIFT with L_2 distance performs poorly.

2. Summer 2019
    - Improve the performance of RIFTs in higher dimensions.
    - Explore power of RIFT for higher dimensions ($d \to \infty$).
    - Find the minimax testing rate.

Future Work - Timeline

1. Spring 2019
   - Apply the Ghosh-Sen separation idea in practice.
   - Analyze performance of hierarchical RIFTs theoretically.
   - Study why S-RIFT with L_2 distance performs poorly.

2. Summer 2019
   - Improve the performance of RIFTs in higher dimensions.
   - Explore power of RIFT for higher dimensions ($d \to \infty$).
   - Find the minimax testing rate.

3. Fall 2019
   - Work on an anomaly detection algorithm for CERN with Mikael Kuusela.
   - Writing Thesis.
   - Job applications.

Future Work - Timeline

1. Spring 2019
   - Apply the Ghosh-Sen separation idea in practice.
   - Analyze performance of hierarchical RIFTs theoretically.
   - Study why S-RIFT with L_2 distance performs poorly.

2. Summer 2019
   - Improve the performance of RIFTs in higher dimensions.
   - Explore power of RIFT for higher dimensions ($d \to \infty$).
   - Find the minimax testing rate.

3. Fall 2019
   - Work on an anomaly detection algorithm for CERN with Mikael Kuusela.
   - Writing Thesis.
   - Job applications.

4. Spring 2020
   - Defend.

References

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

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

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.

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.

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*.

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

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.

McLachlan, G. and Peel, D. (2000). Finite mixture models, willey series in probability and statistics.

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.

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.

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.

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.

Thank you!

Asymptotic Normality

- Replace $R_i \to \tilde{R}_i = R_i + \delta Z_i$, $Z_1, \ldots, Z_n \sim N(0,1)$, $\delta = 10^{-5}$ (say).
- Let $\hat{p}_1 = N(\hat{\mu}_0, \hat{\Sigma}_0)$ and $\hat{p}_2 = \hat{\alpha} N(\hat{\mu}_1, \hat{\Sigma}_1) + (1-\hat{\alpha}) N(\hat{\mu}_2, \hat{\Sigma}_2)$.

Theorem 6

Assume each $\hat{\mu}_i \in \mathcal{A}$, a compact set and the eigenvalues of $\hat{\Sigma}_i \in [c_1, c_2]$. Let $Z \sim N(0, \tau^2)$ where $\tau^2 = \mathbb{E}[(\tilde{R}_i - \Gamma)^2 | \mathcal{D}_1]$. Then, under H_0

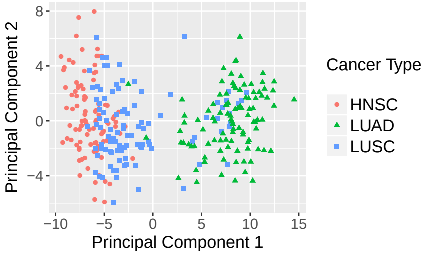
$$\sup_t \left| P(\sqrt{n}(\hat{\Gamma} - \Gamma) \le t \,|\, \mathcal{D}_1) - P(Z \le t) \right| \le \frac{C}{\sqrt{n}} \tag{2}$$

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.
Since C does not depend on $\mathcal{D}_1$ we also have,

$$\sup_t \left| P(\sqrt{n}(\hat{\Gamma} - \Gamma) \le t) - P(Z \le t) \right| \le \frac{C}{\sqrt{n}}. \tag{3}$$

TCGA project: Multi-Cancer Gene Expression Dataset

1. RIFTs: 3 clusters.
2. SigClust: 9 clusters.
3. AIC: 12, BIC: 8.



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

True	SigClust's 1st 3 Classes		
	HNSC	LUSC	LUAD
HNSC	90	10	0
LUSC	4	74	22
LUAD	0	1	99

Sequential RIFT (S-RIFT)

- Using $\mathcal{D}_1$, fit a mixture of k Normals for $k = 1, 2, \ldots, K_n$, $K_n = \sqrt{n}$ (say).

- Using $\mathcal{D}_2$, for $j = 1, 2, \ldots,$ we test

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

- Reject H_{0j} if

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

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

- $\hat{k}$ is the first value of j for which H_{0j} is not rejected. $\hat{p}_{\hat{k}}$ defines the clusters.

Same location, changing proportion.

$$X_1, \ldots, X_n \sim \pi N(\mathbf{0}, I_d) + (1-\pi) N(\mathbf{0}, 5\ I_d)$$

Mardia's Kurtosis performs the best! M-RIFT has low power when $\pi < 5$.