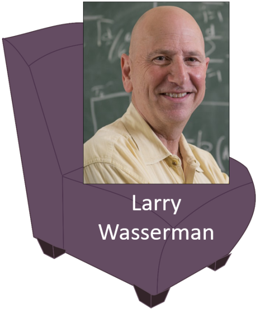
Inference for Clustering and Anomaly Detection

Purvasha Chakravarti

Department of Statistics & Data Science



Siva Balakrishnan

Mikael Kuusela

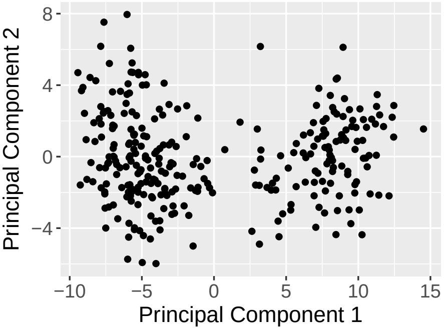
Andrew Nobel

Rebecca Nugent

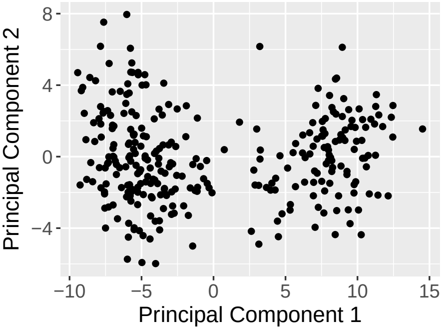
Alessandro Rinaldo

Carnegie Mellon University

How many clusters are "really" there?

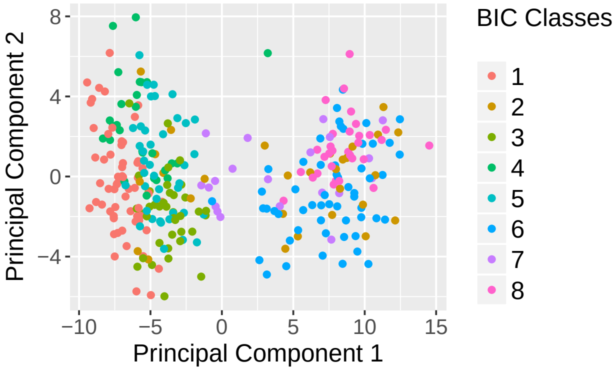


How many clusters are “really” there?



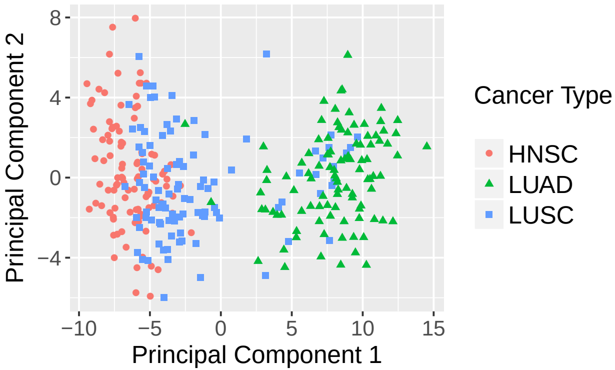
Popular answers: AIC, BIC, gap statistic (Tibshirani et al. (2001)), Hartigan index (Hartigan (1975)), the silhoutte statistic (Rousseeuw (1987)), Ghosh and Sen (1984), Milligan and Cooper (1985), Bock (1985), McLachlan and Peel (2000), Fraley and Raftery (2002), McLachlan and Peel (2004), McLachlan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

How many clusters are “really” there?



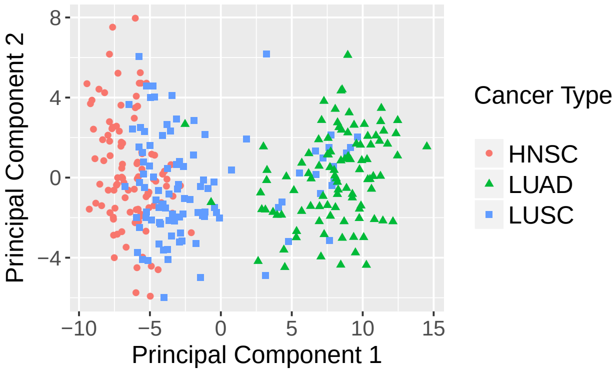
Popular answers: AIC, BIC, gap statistic (Tibshirani et al. (2001)), Hartigan index (Hartigan (1975)), the silhoutte statistic (Rousseeuw (1987)), Ghosh and Sen (1984), Milligan and Cooper (1985), Bock (1985), McLachlan and Peel (2000), Fraley and Raftery (2002), McLachlan and Peel (2004), McLachlan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

Eg: The Cancer Genome Atlas (TCGA) project



Carnegie Mellon University

Eg: The Cancer Genome Atlas (TCGA) project



RNA sequence data: Head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). (Network et al. (2012), Network et al. (2014))

Sections of the talk

1. Clustering

How can we perform clustering that results in statistically significant clusters?

Carnegie Mellon University

Sections of the talk

1. Clustering

How can we perform clustering that results in statistically significant clusters?

2. Anomaly Detection

In high energy physics, how can we detect new signals in experimental data in a model-independent way?

Sections of the talk

1. Clustering

Gaussian Mixture Clustering Using Relative Tests of Fit

Joint work with:

Sivaraman Balakrishnan and

Larry Wasserman

2. Anomaly Detection

In high energy physics, how can we detect new signals in experimental data in a model-independent way?

Sections of the talk

1. Clustering

Gaussian Mixture Clustering Using Relative Tests of Fit

Joint work with:

Sivaraman Balakrishnan and Larry Wasserman

2. Anomaly Detection

Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests

Joint work with:

Mikael Kuusela and Larry Wasserman

Significant Clustering via SigClust: How it works!

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

Significant Clustering via SigClust: How it works!

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

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

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

Significant Clustering via 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. Uses 2-means clustering and the Cluster Index for 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.

Significant Clustering via 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. Uses 2-means clustering and the Cluster Index for 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.

Significant Clustering via SigClust: How it works!

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

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

   $H_0 : X_1, \ldots, X_n \sim N(\mu, \Sigma)$ versus

   $H_1 : X_1, \ldots, X_n \sim f(\cdot)$, which is a non-Gaussian distribution.

2. Uses 2-means clustering and the Cluster Index for 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.

4. Works well in HDLSS data.

Carnegie Mellon University

Power of SigClust: Low power in some cases

Power of SigClust: Low power in some cases

Theorem 1 (Chakravarti, Purvasha et al. (2019))

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

Power of SigClust: Low power in some cases

Theorem 1 (Chakravarti, Purvasha et al. (2019))

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

Carnegie Mellon University

Power of SigClust: Low power in some cases

Theorem 1 (Chakravarti, Purvasha et al. (2019))

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

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

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

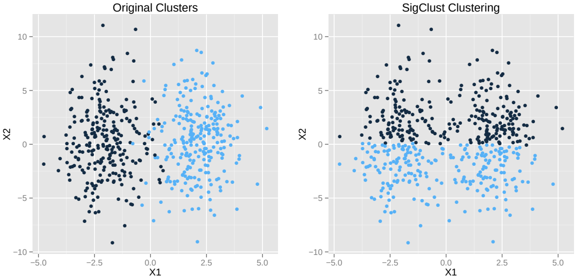
Power of SigClust: Low power in some cases

Theorem 1 (Chakravarti, Purvasha et al. (2019))

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

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

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



k-means optimal split, splits horizontally!

Proposed test: Relative Information Fit Test (RIFT)

1. 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 ($0 < \pi_k < 1, \sum_k \pi_k = 1$).

Carnegie Mellon University

Proposed test: Relative Information Fit Test (RIFT)

1. 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 $\mathbf{y} \in \mathbb{R}^d$,

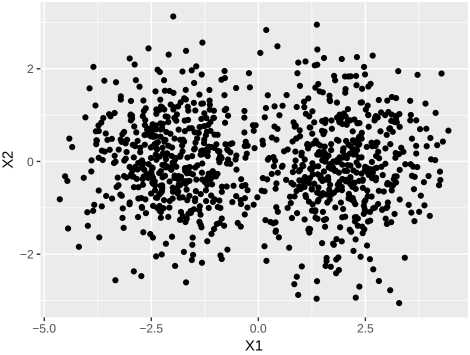
$$p(\mathbf{y}|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{y}|\mu_k, \Sigma_k),$$

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

2. Test if a mixture of two Gaussians **fits** the data significantly better than a single Gaussian.

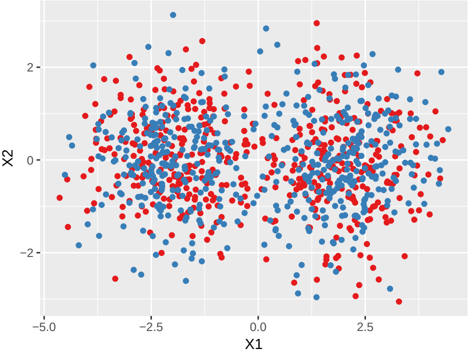
Proposed test: Relative Information Fit Test (RIFT)

Randomly split data into D_1 (Estimating) and D_2 (Testing).



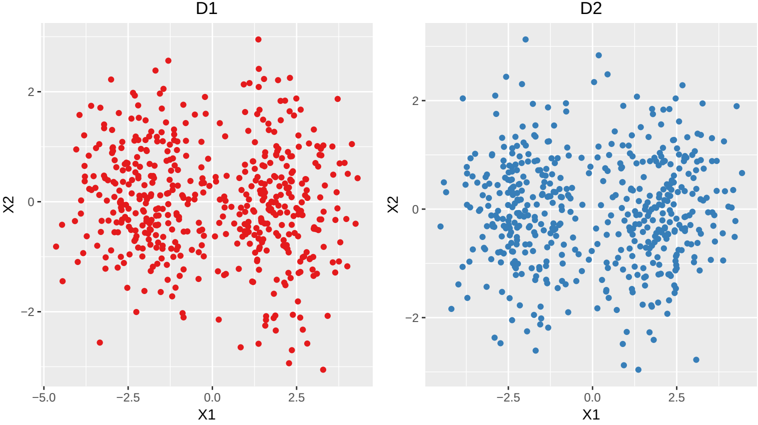
Proposed test: Relative Information Fit Test (RIFT)

Randomly split data into D_1 (Estimating) and D_2 (Testing).



Proposed test: Relative Information Fit Test (RIFT)

Randomly split data into D_1 (Estimating) and D_2 (Testing).

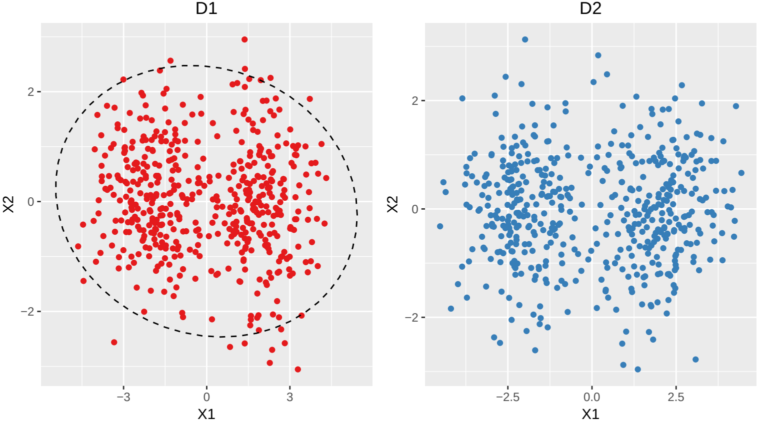


Carnegie Mellon University

Proposed test: Relative Information Fit Test (RIFT)

Using D_1, fit a Normal $\hat{p}_1$ and a mixture of two Normals $\hat{p}_2$.

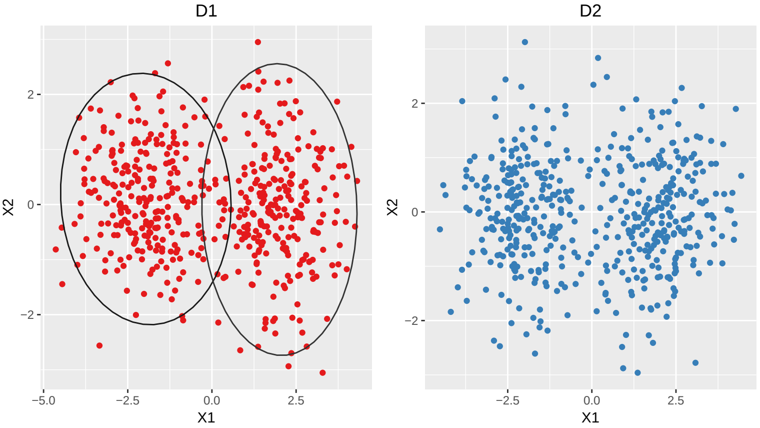
$\hat{p}_1$



Proposed test: Relative Information Fit Test (RIFT)

Using D_1, fit a Normal $\hat{p}_1$ and a mixture of two Normals $\hat{p}_2$.

$\hat{p}_2$

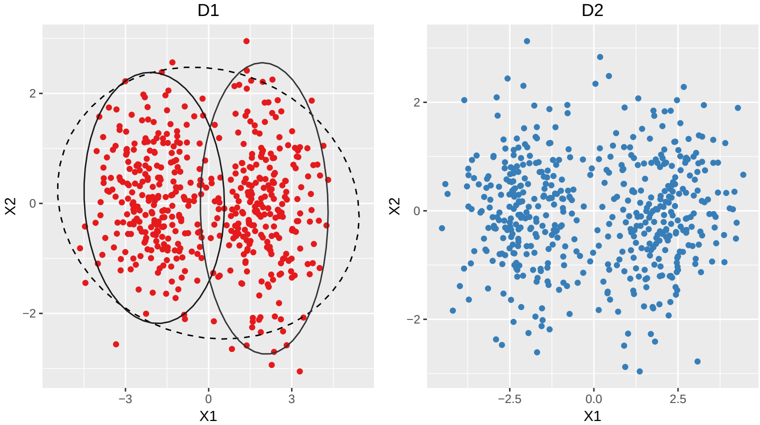


Carnegie Mellon University

Proposed test: Relative Information Fit Test (RIFT)

$\Gamma = K(p, \hat{p}_1) - K(p, \hat{p}_2)$, where K is the KL distance, p is the true density.

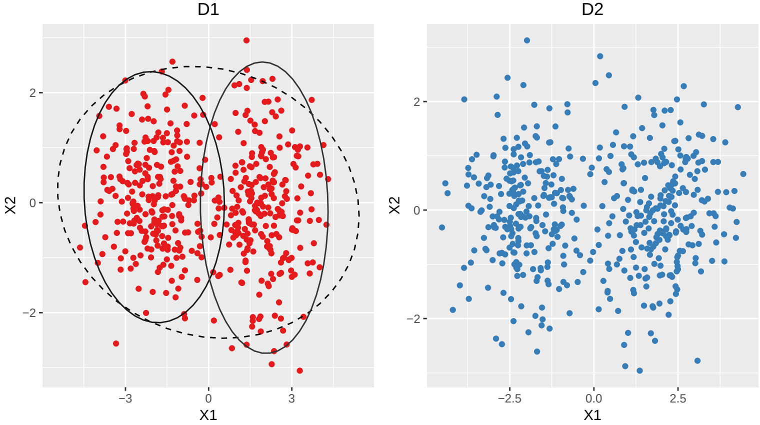
$\hat{p}_1, \ \hat{p}_2$



We test, conditioned on D_1, $H_0 : \Gamma \leq 0$ versus $H_1 : \Gamma > 0$.

Proposed test: Relative Information Fit Test (RIFT)

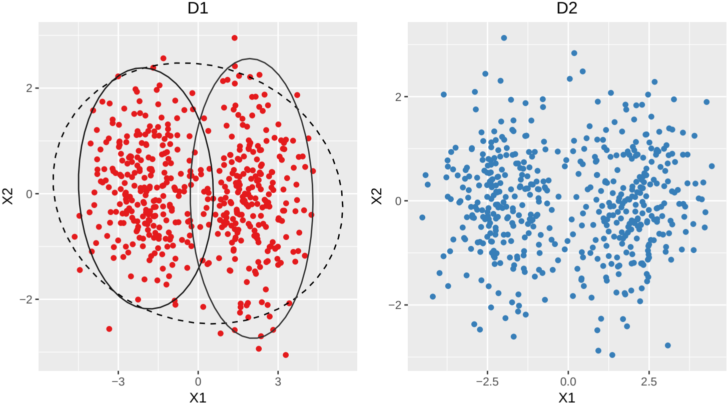
$\hat{p}_1,\ \hat{p}_2$ $\qquad$ $\hat{\Gamma} = \frac{1}{n}\sum_{i \in D_2} R_i,\ \ R_i = \log\left(\frac{\hat{p}_2(X_i)}{\hat{p}_1(X_i)}\right)$



We test, conditioned on D_1, $H_0 : \Gamma \leq 0$ versus $H_1 : \Gamma > 0$.

Proposed test: Relative Information Fit Test (RIFT)

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



We test, conditioned on D_1, $H_0 : \Gamma \leq 0$ versus $H_1 : \Gamma > 0$.

$$\sqrt{n}\left(\hat{\Gamma} - \Gamma\right)/\tau \rightsquigarrow N(0,1) \implies \text{Reject } H_0 \text{ if } \hat{\Gamma} > \frac{z_\alpha \hat{\tau}}{\sqrt{n}}.$$

Carnegie Mellon University

Power of RIFT converges to 1!

Power converges to 1!

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

Lemma 2

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

Carnegie Mellon University

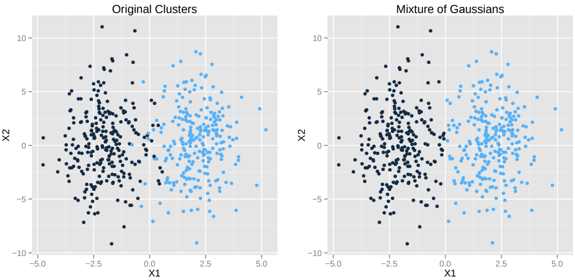
Power of RIFT converges to 1!

Power converges to 1!

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

Lemma 2

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



Power of RIFT converges to 1!

Power converges to 1!

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

Lemma 2

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

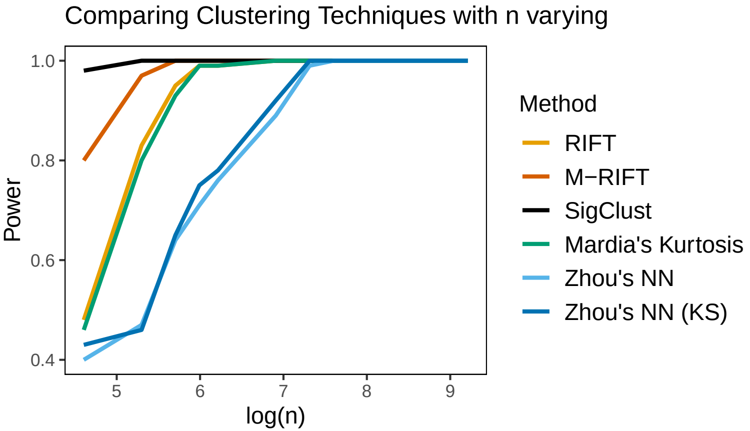
RIFT can be applied both hierarchically and sequentially to detect more than two clusters with asymptotic error control!

RIFT also has a more robust version - Median RIFT (M-RIFT)!

Comparisions for 2 Normals: SigClust performs better

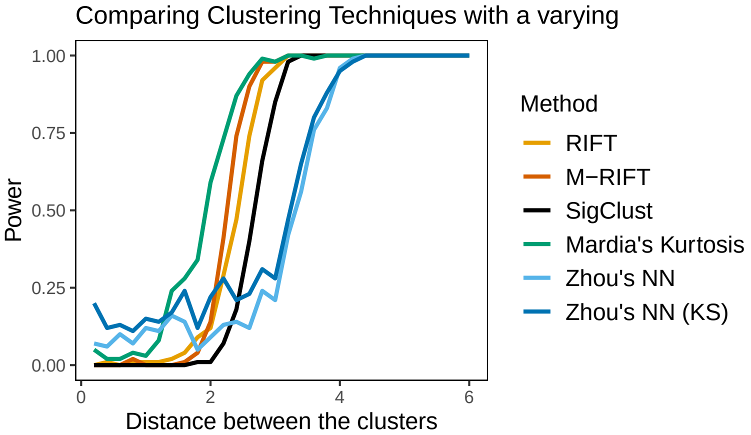
$$X_1, \ldots, X_n \sim \frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d) \text{ where } \mu = (a, 0, \ldots, 0)$$

Example where SigClust's power converges to 1 as $n \to \infty$.



Comparisions for 2 Normals: RIFTs perform better

$$X_1, \ldots, X_n \sim \frac{1}{2} N(\mu, I_d) + \frac{1}{2} N(-\mu, I_d) \text{ where } \mu = (a, 0, \ldots, 0)$$



Overview of Contributions

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

Overview of Contributions

- RIFTs - simple and easy tests to detect significant clusters.
- RIFTs don't make any model assumptions on the clusters.

Overview of Contributions

- RIFTs - simple and easy tests to detect significant clusters.
- RIFTs don't make any model assumptions on the clusters.
- They can be applied hierarchically as well as sequentially, while asymptotically controlling for type I error.

Overview of Contributions

- RIFTs - simple and easy tests to detect significant clusters.
- RIFTs don't make any model assumptions on the clusters.
- They can be applied hierarchically as well as 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.

Overview of Contributions

- RIFTs - simple and easy tests to detect significant clusters.
- RIFTs don't make any model assumptions on the clusters.
- They can be applied hierarchically as well as 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.

Overview of Contributions

- RIFTs - simple and easy tests to detect significant clusters.
- RIFTs don't make any model assumptions on the clusters.
- They can be applied hierarchically as well as 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.

Sections of the talk

1. Clustering

Gaussian Mixture Clustering Using Relative Tests of Fit

Joint work with:

Sivaraman Balakrishnan and

Larry Wasserman

2. Anomaly Detection

Model-Independent Detection of New Physics Signals Using Semi-Supervised Classifier Tests

Joint work with:

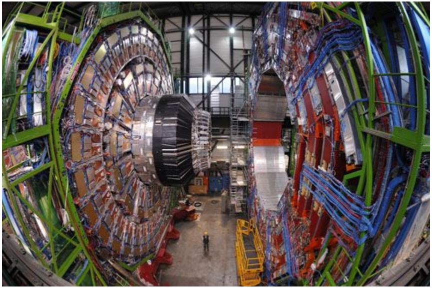
Mikael Kuusela and Larry Wasserman

CERN and the Large Hadron Collider

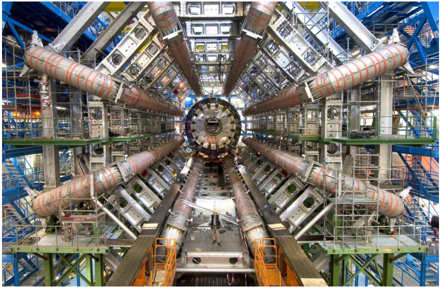
Carnegie Mellon University

The ATLAS and the CMS experiments at the LHC

CMS experiment



ATLAS experiment



Carnegie Mellon University

Events from the experiments

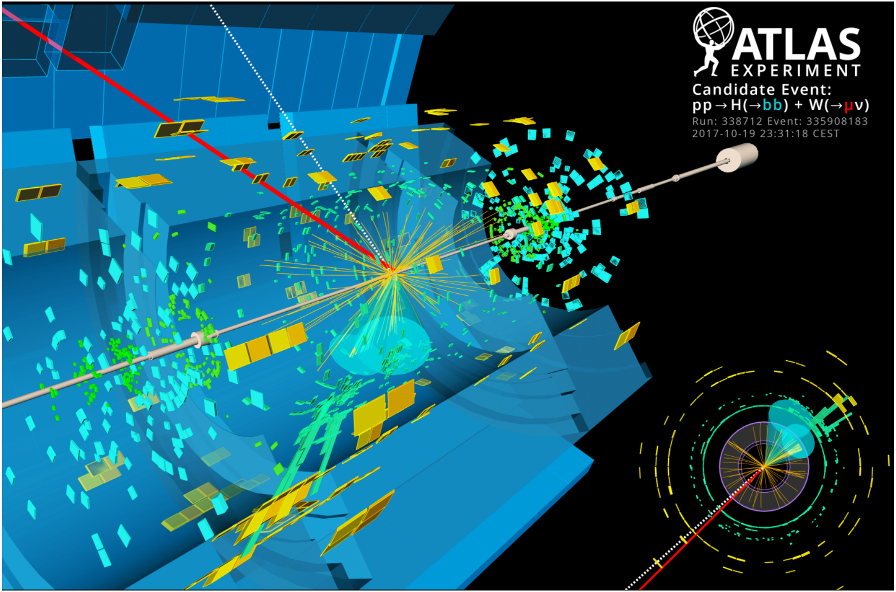
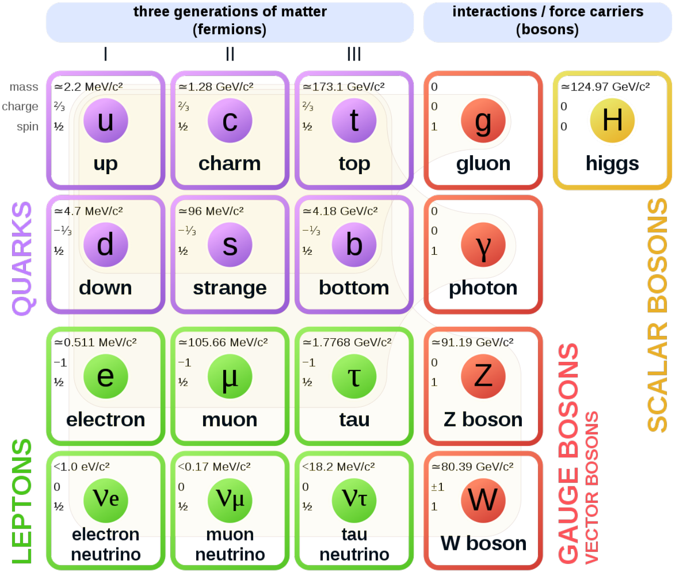


Image credit: CERN

The Standard Model of particle physics



Experimental data

Experimental data are generated from one of the two processes:

Background - refers to the known physics (SM).

Signal - represents an unknown possible particle or interaction not accounted for in the SM.

Experimental data

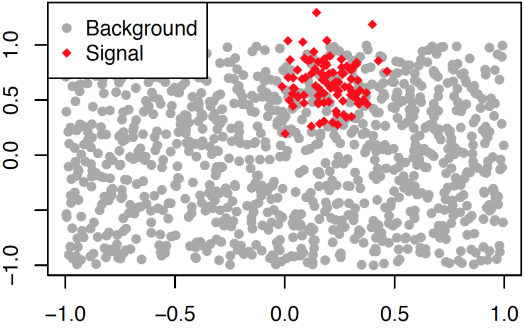
Experimental data are generated from one of the two processes:
Background - refers to the known physics (SM).
Signal - represents an unknown possible particle or interaction not accounted for in the SM.

$$q = (1 - \lambda)p_b + \lambda p_s, \quad \text{No signal: } \lambda = 0.$$

Experimental data

Experimental data are generated from one of the two processes:
Background - refers to the known physics (SM).
Signal - represents an unknown possible particle or interaction not accounted for in the SM.

$$q = (1-\lambda)p_b + \lambda p_s, \quad \text{No signal: } \lambda = 0.$$



Two-dimensional toy example.

Carnegie Mellon University

Model-dependent supervised methods

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\begin{aligned} \text{Background:} &\quad X_1, \ldots, X_m \sim p_b \\ \text{Signal:} &\quad Y_1, \ldots, Y_n \sim p_s \end{aligned}$$

Model-dependent supervised methods

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\begin{aligned} \text{Background:} \quad & X_1, \ldots, X_m \sim p_b \\ \text{Signal:} \quad & Y_1, \ldots, Y_n \sim p_s \end{aligned}$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s$$

Model-dependent supervised methods

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_m \sim p_b$$
$$\text{Signal:} \quad Y_1, \ldots, Y_n \sim p_s$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_N \sim q = (1-\lambda)p_b + \lambda p_s$$

Test $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$.

Train a classifier (h) to separate signal from background.

Model-dependent likelihood ratio using supervised classifier

- Classifier (h) separates signal from background.

Model-dependent likelihood ratio using supervised classifier

- Classifier (h) separates signal from background.
- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i [(1 - \lambda) + \lambda \psi(W_i)], \quad \psi = p_s / p_b.$$

Carnegie Mellon University

Model-dependent likelihood ratio using supervised classifier

- Classifier (h) separates signal from background.

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i [(1-\lambda) + \lambda\psi(W_i)], \quad \psi = p_s/p_b.$$

- The membership probabilities h can be written as:

$$h(z) = \widehat{\mathbb{P}}(Z \text{ is signal} | Z = z) = \frac{np_s(z)}{np_s(z) + mp_b(z)} = \frac{n\psi(z)}{n\psi(z) + m}.$$

Carnegie Mellon University

Model-dependent likelihood ratio using supervised classifier

- Classifier (h) separates signal from background.

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i [(1-\lambda) + \lambda\psi(W_i)], \quad \psi = p_s/p_b.$$

- The membership probabilities h can be written as:

$$h(z) = \widehat{\mathbb{P}}(Z \text{ is signal} | Z = z) = \frac{np_s(z)}{np_s(z) + mp_b(z)} = \frac{n\psi(z)}{n\psi(z) + m}.$$

- We can estimate

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

Carnegie Mellon University

Model-dependent supervised methods test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i [(1-\lambda) + \lambda\psi(W_i)], \quad \psi = p_s/p_b.$$

Model-dependent supervised methods test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i [(1-\lambda) + \lambda\psi(W_i)], \quad \psi = p_s/p_b.$$

1. Likelihood Ratio Test Statistic:

$$\mathrm{LRT} = 2\sum_i \log\left((1-\hat{\lambda}_{\mathrm{MLE}}) + \hat{\lambda}_{\mathrm{MLE}}\hat{\psi}(W_i)\right)$$

Carnegie Mellon University

Model-dependent supervised methods test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i [(1-\lambda) + \lambda\psi(W_i)], \quad \psi = p_s/p_b.$$

1. Likelihood Ratio Test Statistic:

$$\text{LRT} = 2\sum_i \log\left((1-\hat{\lambda}_{\text{MLE}}) + \hat{\lambda}_{\text{MLE}}\hat{\psi}(W_i)\right)$$

2. Score Test Statistic:

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

Model-dependent supervised methods test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i [(1-\lambda) + \lambda\psi(W_i)], \quad \psi = p_s/p_b.$$

1. Likelihood Ratio Test Statistic:

$$\text{LRT} = 2\sum_i \log\left((1-\hat{\lambda}_{\text{MLE}}) + \hat{\lambda}_{\text{MLE}}\hat{\psi}(W_i)\right)$$

2. Score Test Statistic:

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

- Asymptotic method for first, permutation and bootstrap methods for both.

Motivation for model-independent methods

- What if none of the current proposed models are right for the New Physics (NP) signals?

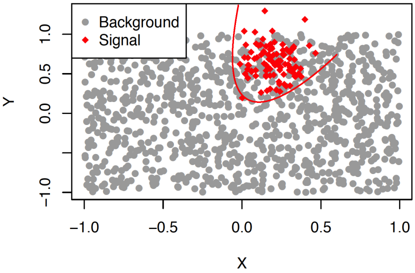
- How to look for NP when one is not totally sure what to look for?

Motivation for model-independent methods

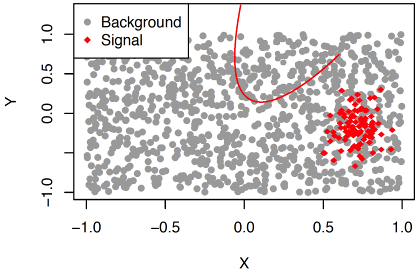
- What if none of the current proposed models are right for the New Physics (NP) signals?

- How to look for NP when one is not totally sure what to look for?

Classifier decision boundary



Actual NP signal



Solution: Model-independent methods

Two sources of data are at hand:

- Background (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_m \sim p_b$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_N \sim q = (1-\lambda)p_b + \lambda p_s$$

Solution: Model-independent methods

Two sources of data are at hand:

- Background (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_m \sim p_b$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s$$

Kuusela et al. (2012) and Vatanen et al. (2012) use Gaussian Mixture Models.

Solution: Model-independent methods

Two sources of data are at hand:

- Background (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_m \sim p_b$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_N \sim q = (1-\lambda)p_b + \lambda p_s$$

Kuusela et al. (2012) and Vatanen et al. (2012) use Gaussian Mixture Models.

We use a classifier to detect the signal through rigorous inference.

Proposed model-independent semi-supervised methods

Two sources of data are at hand:

- Background (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_m \sim p_b$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_N \sim q = (1-\lambda)p_b + \lambda p_s$$

Train a classifier ($\tilde{h}$) to separate experimental from background.

Proposed model-independent semi-supervised methods

Two sources of data are at hand:

- Background (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_m \sim p_b$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_N \sim q = (1-\lambda)p_b + \lambda p_s$$

Train a classifier ($\tilde{h}$) to separate experimental from background.

Note:
1. We don't use labelled signal observations.
2. We used Random Forest as a classifier.

Proposed test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.$$

Carnegie Mellon University

Proposed test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.$$

- Classifier $\tilde{h}$ that separates experimental from background, gives $\widehat{\tilde{\psi}}(z)$.

Carnegie Mellon University

Proposed test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.$$

- Classifier $\tilde{h}$ that separates experimental from background, gives $\widehat{\tilde{\psi}}(z)$.

  1. Likelihood Ratio Test Statistic:

$$\text{LRT} = 2\sum_i \log \widehat{\tilde{\psi}}(W_i).$$

Carnegie Mellon University

Proposed test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.$$

- Classifier $\tilde{h}$ that separates experimental from background, gives $\widehat{\tilde{\psi}}(z)$.

1. Likelihood Ratio Test Statistic:

$$\text{LRT} = 2 \sum_i \log \widehat{\tilde{\psi}}(W_i).$$

2. Area Under the Curve Test (AUC) Statistic: $\hat{\theta}$
   Test $H_0 : \theta = 0.5$ versus $H_1 : 0.5 < \theta < 1$.

Carnegie Mellon University

Proposed test statistics

- Likelihood Ratio on the W_i's for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

$$\frac{\mathcal{L}_q(\lambda)}{\mathcal{L}_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.$$

- Classifier $\tilde{h}$ that separates experimental from background, gives $\widehat{\tilde{\psi}}(z)$.

1. Likelihood Ratio Test Statistic:

$$\text{LRT} = 2 \sum_i \log \widehat{\tilde{\psi}}(W_i).$$

2. Area Under the Curve Test (AUC) Statistic: $\hat{\theta}$
   Test $H_0 : \theta = 0.5$ versus $H_1 : 0.5 < \theta < 1$.

- Asymptotic, permutation and bootstrap methods for both.

Kaggle's Higgs boson challenge

- Data provided by ATLAS.

[1] https://www.kaggle.com/c/higgs-boson

Kaggle's Higgs boson challenge

- Data provided by ATLAS.
- 15 variables.

[1] https://www.kaggle.com/c/higgs-boson

Kaggle's Higgs boson challenge

- Data provided by ATLAS.
- 15 variables.
- Transverse momentum and energy as well as angles of resulting particles and jets of particles in a collision event.

[1] https://www.kaggle.com/c/higgs-boson

Kaggle's Higgs boson challenge

- Data provided by ATLAS.
- 15 variables.
- Transverse momentum and energy as well as angles of resulting particles and jets of particles in a collision event.
- $24,645$ background events and $25,734$ signal events.

[1] https://www.kaggle.com/c/higgs-boson

Kaggle's Higgs boson challenge

- Data provided by ATLAS.
- 15 variables.
- Transverse momentum and energy as well as angles of resulting particles and jets of particles in a collision event.
- $24,645$ background events and $25,734$ signal events.
- Create experimental data in 100 simulations with varying signal strength, λ.

[1] https://www.kaggle.com/c/higgs-boson

Kaggle's Higgs boson challenge

- Data provided by ATLAS.
- 15 variables.
- Transverse momentum and energy as well as angles of resulting particles and jets of particles in a collision event.
- $24,645$ background events and $25,734$ signal events.
- Create experimental data in 100 simulations with varying signal strength, λ.
- Compare power of the methods in detecting the Higgs boson.

[1] https://www.kaggle.com/c/higgs-boson

Power - simulations where the Higgs boson is detected

λ is the proportion of signal in the experimental data set.

100 simulations.

Model-dependent methods that have signal labels.

	Model	Method	Signal Strength (λ)					
			0.15	0.1	0.07	0.05	0.01	0
Signal Labels	Supervised LRT	Asymptotic	99	70	22	5	0	0
		Permutation	99	93	59	19	1	0
	Supervised Score	Permutation	99	94	80	51	13	7

Carnegie Mellon University

Power - simulations where the Higgs boson is detected

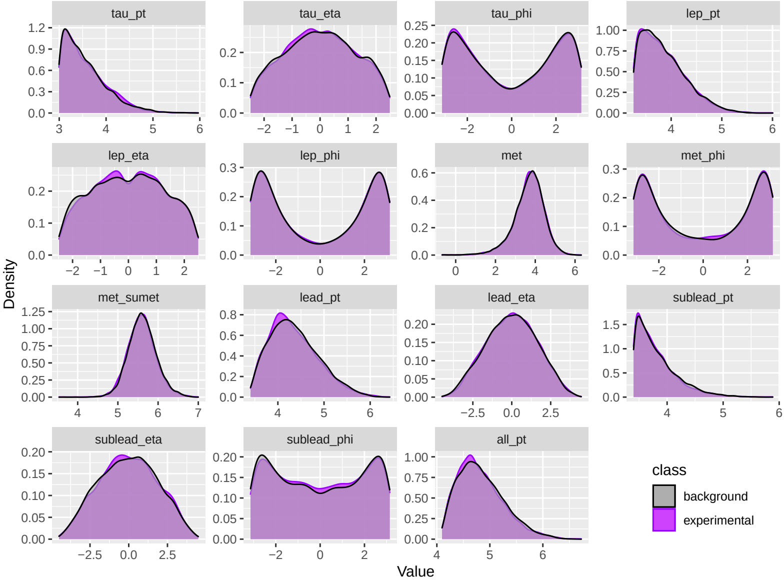
λ is the proportion of signal in the experimental data set.

100 simulations.

	Model	Method	Signal Strength (λ)					
			0.15	0.1	0.07	0.05	0.01	0
Signal Labels	Supervised LRT	Asymptotic	99	70	22	5	0	0
		Permutation	99	93	59	19	1	0
	Supervised Score	Permutation	99	94	80	51	13	7
NO Signal Labels	Semi-Supervised LRT	Asymptotic	99	63	16	20	5	7
		Permutation 1	99	60	17	19	5	8
	Semi-Supervised AUC	Asymptotic	96	63	17	17	6	8
		Permutation 1	97	62	18	16	6	8
		Permutation 2	100	74	38	23	4	6
	NN Two-Sample	Permutation	74	33	10	10	8	5

Carnegie Mellon University

Density of the training data variables, $\lambda = 0.15$



Carnegie Mellon University

Identifying the active subspace that explains the classifier

- Consider $\nabla_{\mathbf{z}} \tilde{h}(\mathbf{z})$.

Identifying the active subspace that explains the classifier

- Consider $\nabla_{\mathbf{z}} \tilde{h}(\mathbf{z})$.
- Perform Principal Component Analysis (PCA) or sparse PCA on $\nabla_{\mathbf{z}} \tilde{h}(\mathbf{z})$.

Carnegie Mellon University

Identifying the active subspace that explains the classifier

- Consider $\nabla_{\mathbf{z}} \tilde{h}(\mathbf{z})$.
- Perform Principal Component Analysis (PCA) or sparse PCA on $\nabla_{\mathbf{z}} \tilde{h}(\mathbf{z})$.
- Let $\mathbf{m}_1, \mathbf{m}_2, \ldots$ be the leading eigenvectors.

Identifying the active subspace that explains the classifier

- Consider $\nabla_{\mathbf{z}} \tilde{h}(\mathbf{z})$.
- Perform Principal Component Analysis (PCA) or sparse PCA on $\nabla_{\mathbf{z}} \tilde{h}(\mathbf{z})$.
- Let $\mathbf{m}_1, \mathbf{m}_2, \ldots$ be the leading eigenvectors.
- Then $\mathbb{E}\left[\nabla_{\mathbf{z}} \tilde{h}\right], \mathbf{m}_1, \mathbf{m}_2, \ldots$ best captures the variation in the classifier $\tilde{h}$ (Constantine, 2015).

Active subspace of $\tilde{h}(\cdot)$

For experimental data $W_1, \ldots, W_N$,

- $\nabla_{\mathbf{z}} h(\mathbf{z})$ - $\nabla_{\mathbf{z}} h_j = \widehat{\nabla_{\mathbf{z}} \tilde{h}(W_j)}$ using a local linear smoother on $\tilde{h}$.

Carnegie Mellon University

Active subspace of $\tilde{h}(\cdot)$

For experimental data $W_1, \ldots, W_N$,

- $\nabla_{\mathbf{z}} h(\mathbf{z})$ - $\nabla_{\mathbf{z}} h_j = \widehat{\nabla_{\mathbf{z}} \tilde{h}(W_j)}$ using a local linear smoother on $\tilde{h}$.
- Perform Principal Component Analysis (PCA) or sparse PCA on $H = (\nabla_{\mathbf{z}} h_1, \nabla_{\mathbf{z}} h_2, \ldots, \nabla_{\mathbf{z}} h_N)^T$.

Active subspace of $\tilde{h}(\cdot)$

For experimental data $W_1, \ldots, W_N$,

- $\nabla_{\mathbf{z}} h(\mathbf{z})$ - $\nabla_{\mathbf{z}} h_j = \widehat{\nabla_{\mathbf{z}} \tilde{h}(W_j)}$ using a local linear smoother on $\tilde{h}$.
- Perform Principal Component Analysis (PCA) or sparse PCA on $H = (\nabla_{\mathbf{z}} h_1, \nabla_{\mathbf{z}} h_2, \ldots, \nabla_{\mathbf{z}} h_N)^T$.
- Let $\mathbf{m}_1, \mathbf{m}_2, \ldots$ be the leading eigenvectors - $\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, \ldots$.

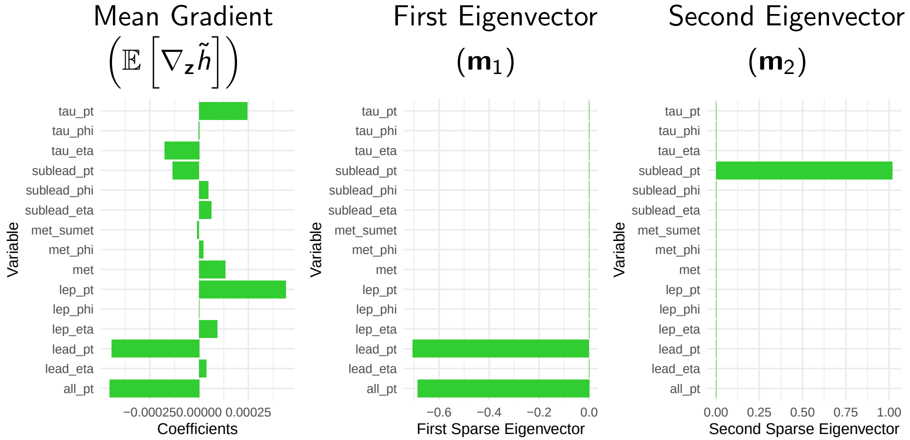
Active subspace of $\tilde{h}(\cdot)$

For experimental data $W_1, \ldots, W_N$,

- $\nabla_{\mathbf{z}} h(\mathbf{z})$ - $\nabla_{\mathbf{z}} h_j = \widehat{\nabla_{\mathbf{z}} \tilde{h}(W_j)}$ using a local linear smoother on $\tilde{h}$.
- Perform Principal Component Analysis (PCA) or sparse PCA on $H = (\nabla_{\mathbf{z}} h_1, \nabla_{\mathbf{z}} h_2, \ldots, \nabla_{\mathbf{z}} h_N)^T$.
- Let $\mathbf{m}_1, \mathbf{m}_2, \ldots$ be the leading eigenvectors - $\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, \ldots$.
- $\mathbb{E}\left[\nabla_{\mathbf{z}} \tilde{h}\right], \mathbf{m}_1, \mathbf{m}_2, \ldots$ - $\overline{\nabla_{\mathbf{z}} h_j} = \frac{1}{N} \sum_{j=1}^{N} \nabla_{\mathbf{z}} h_j, \ \hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, \ldots$.

Carnegie Mellon University

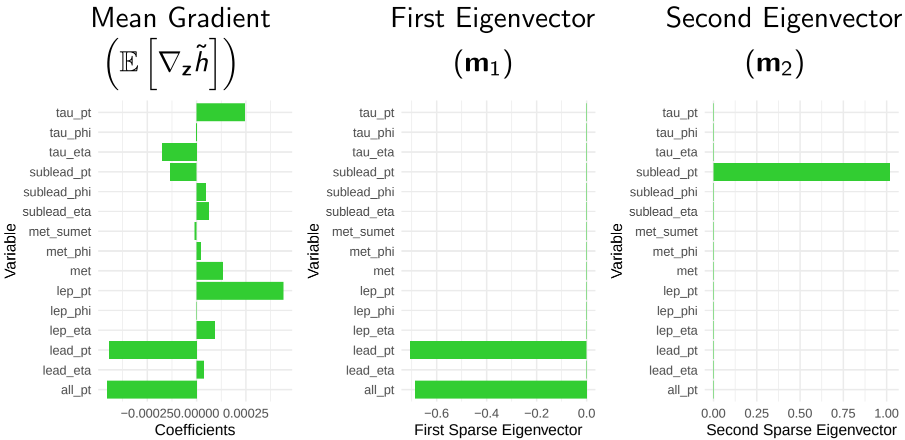
Active subspace for $\tilde{h}(\cdot)$ when $\lambda = 0.15$



Carnegie Mellon University

Active subspace for $\tilde{h}(\cdot)$ when $\lambda = 0.15$

The vectors capture the variable dependencies that influence the classifier.



Carnegie Mellon University

Overview of Contributions

- Propose semi-supervised classifiers that separate experimental data from the background.

Overview of Contributions

- Propose semi-supervised classifiers that separate experimental data from the background.
- Detect signal in a model-independent way through rigorous inference.

Overview of Contributions

- Propose semi-supervised classifiers that separate experimental data from the background.
- Detect signal in a model-independent way through rigorous inference.
- Use LRT and AUC statistics to perform the test.

Overview of Contributions

- Propose semi-supervised classifiers that separate experimental data from the background.
- Detect signal in a model-independent way through rigorous inference.
- Use LRT and AUC statistics to perform the test.
- Propose active subspace methods to explain the classifier.

Thank you CMU Statistics & Data Science
and commitee members!

References

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

Chakravarti, Purvasha, Balakrishnan, S., and Wasserman, L. (2019). Gaussian mixture clustering using relative tests of fit. *arXiv preprint arXiv:1910.02566*.

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

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.

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.

Kuusela, M., Vatanen, T., Malmi, E., Raiko, T., Aaltonen, T., and Nagai, Y. (2012). Semi-supervised anomaly detection–towards model-independent searches of new physics. In *Journal of Physics: Conference Series*, volume 368, page 012032. IOP Publishing.

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.

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

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.

Vatanen, T., Kuusela, M., Malmi, E., Raiko, T., Aaltonen, T., and Nagai, Y. (2012). Semi-supervised detection of collective anomalies with an application in high energy particle physics. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Future Work

- **High-dimensional Clustering.**
  1(a). Clustering after dimension reduction.
  1(b). Better ways of fitting high-dimensional mixture of Gaussians.

Future Work

- **High-dimensional Clustering.**

  1(a). Clustering after dimension reduction.

  1(b). Better ways of fitting high-dimensional mixture of Gaussians.

  2. Consistency of proposed hierarchical clustering algorithms.

Future Work

- **High-dimensional Clustering.**
  1(a). Clustering after dimension reduction.
  1(b). Better ways of fitting high-dimensional mixture of Gaussians.
  2. Consistency of proposed hierarchical clustering algorithms.

- **Semi-Supervised Anomaly Detection in Particle Physics.**
  1. Compare methods for mis-specified signal models.

Future Work

- **High-dimensional Clustering.**
  1(a). Clustering after dimension reduction.
  1(b). Better ways of fitting high-dimensional mixture of Gaussians.
  2. Consistency of proposed hierarchical clustering algorithms.

- **Semi-Supervised Anomaly Detection in Particle Physics.**
  1. Compare methods for mis-specified signal models.
  2. Explore other interpretability methods like Shaply values.

Future Work

- **High-dimensional Clustering.**
  1(a). Clustering after dimension reduction.
  1(b). Better ways of fitting high-dimensional mixture of Gaussians.
  2. Consistency of proposed hierarchical clustering algorithms.

- **Semi-Supervised Anomaly Detection in Particle Physics.**
  1. Compare methods for mis-specified signal models.
  2. Explore other interpretability methods like Shaply values.

- **Relative Fit Methods.** Compare different distance measures when comparing fits of densities.

Future Work

- **High-dimensional Clustering.**
  1(a). Clustering after dimension reduction.
  1(b). Better ways of fitting high-dimensional mixture of Gaussians.
  2. Consistency of proposed hierarchical clustering algorithms.

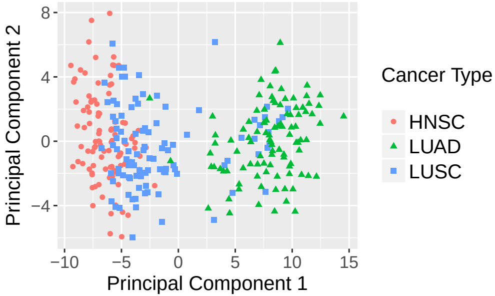
- **Semi-Supervised Anomaly Detection in Particle Physics.**
  1. Compare methods for mis-specified signal models.
  2. Explore other interpretability methods like Shaply values.

- **Relative Fit Methods.** Compare different distance measures when comparing fits of densities.

- **Interdisciplinary Collaborations.**

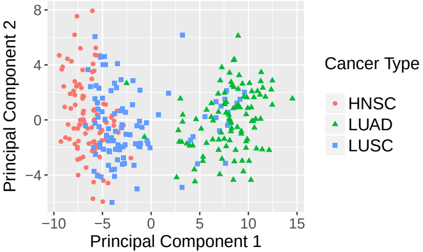
TCGA project: Multi-Cancer Gene Expression Dataset

- 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.



Carnegie Mellon University

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 3

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

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

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

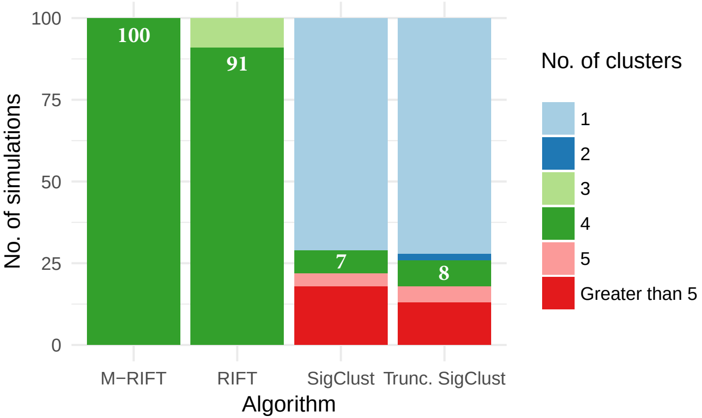
Carnegie Mellon University

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!

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. $\alpha = 0.05$.



Hierarchical RIFT has Type I error control but hierarchical SigClust does not!

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.

Validity of S-RIFT

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

Lemma 4

Under H_{0j},

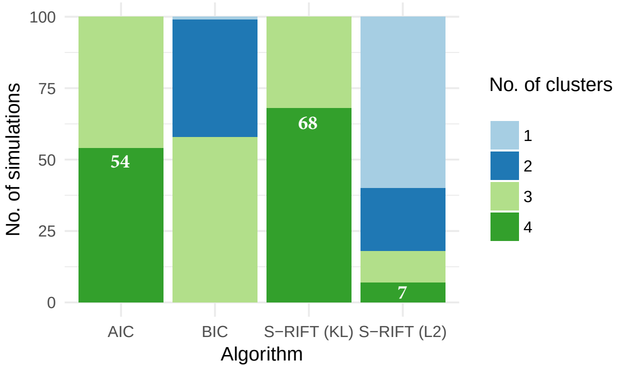
$$\limsup_{n\to\infty} P(\text{rejecting } H_{0j}) \le \alpha.$$

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. $\alpha = 0.05$.



Model-independent Method using Gaussian Mixture Models (GMMs)

Two sources of data are at hand:

- Background (Monte Carlo) sample - labelled observations

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

- Background + possible signal (experimental) sample - unlabelled observations

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

$$q(w|\theta_{sb}) = (1-\lambda)p_b(w|\theta_b) + \lambda p_s(\mathbf{y}|\theta_s),$$

where $\theta_{sb} = (\theta_s, \theta_b, \lambda)$ and both the distribution of the anomaly p_s and the distribution of the background p_b are modeled by mixtures of Gaussian components.

Test for $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$ using likelihood ratio test.

Carnegie Mellon University

Confidence Intervals for AUC

- Newcombe's Wald Method (Newcombe, 2006) gives

$$\widehat{V(\hat{\theta})} = \frac{\hat{\theta}(1-\hat{\theta})}{(n-1)(m-1)}\left[2M - 1 - \frac{3M-3}{(2-\hat{\theta})(1+\hat{\theta})}\right],$$

where $M = \frac{n+m}{2}$.

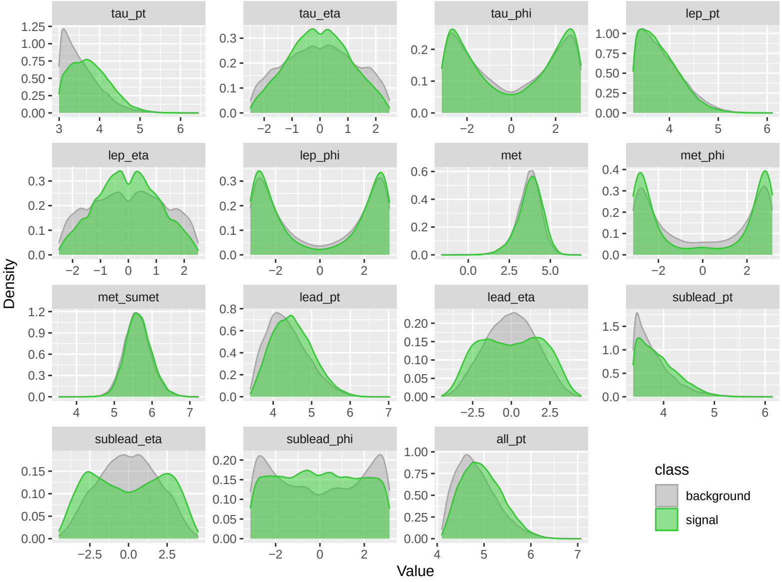
- $100(1-\alpha)\%$ confidence interval for AUC θ is given by

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{V(\hat{\theta})}},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of $\mathrm{N}(0,1)$.

- Test by rejecting $H_0 : \theta = 0.5$ if 0.5 is not in the $100(1-\alpha)\%$ CI.

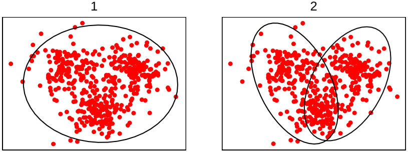
Density of the variables



Hierarchical RIFT (H-RIFT)

Hierarchical RIFT (H-RIFT)

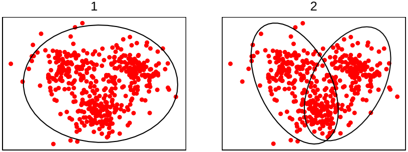
$\hat{p}_1$ vs $\hat{p}_2$



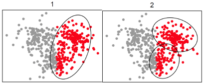
Carnegie Mellon University

Hierarchical RIFT (H-RIFT)

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

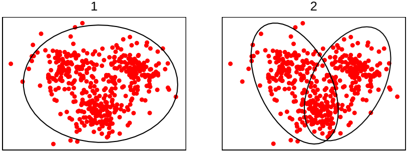


$\hat{p}_1$ vs $\hat{p}_2$

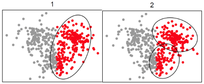


Hierarchical RIFT (H-RIFT)

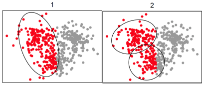
$\hat{p}_1$ vs $\hat{p}_2$



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

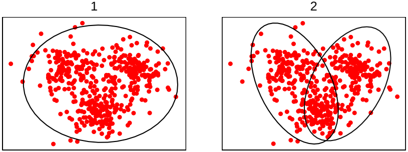


$\hat{p}_1$ vs $\hat{p}_2$

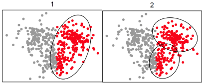


Hierarchical RIFT (H-RIFT)

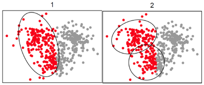
$\hat{p}_1$ vs $\hat{p}_2$

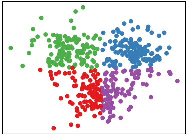


$\hat{p}_1$ vs $\hat{p}_2$



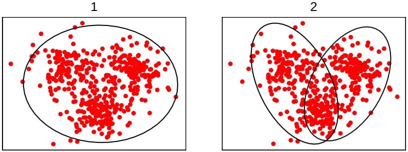
$\hat{p}_1$ vs $\hat{p}_2$



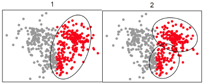


Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

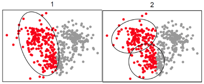
$\hat{p}_1$ vs $\hat{p}_2$

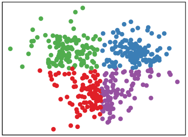


$\hat{p}_1$ vs $\hat{p}_2$



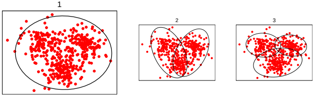
$\hat{p}_1$ vs $\hat{p}_2$





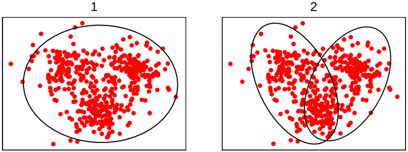
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

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

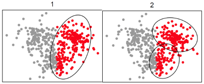


. . .

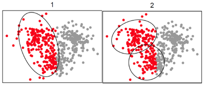
$\hat{p}_1$ vs $\hat{p}_2$

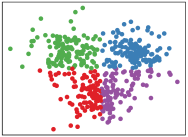


$\hat{p}_1$ vs $\hat{p}_2$



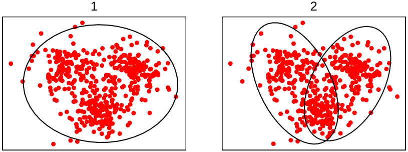
$\hat{p}_1$ vs $\hat{p}_2$



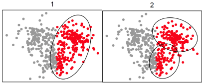


Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

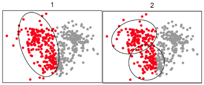
$\hat{p}_1$ vs $\hat{p}_2$

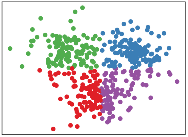


$\hat{p}_1$ vs $\hat{p}_2$

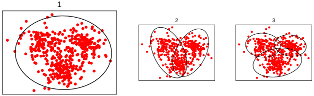


$\hat{p}_1$ vs $\hat{p}_2$



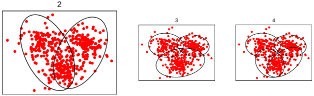


$\hat{p}_1$ vs $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n}$



. . .

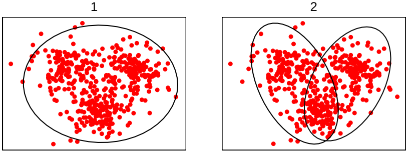
$\hat{p}_2$ vs $\hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n}$



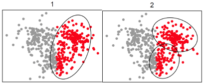
. . .

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

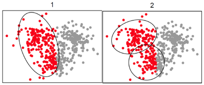
$\hat{p}_1$ vs $\hat{p}_2$

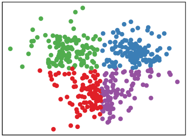


$\hat{p}_1$ vs $\hat{p}_2$

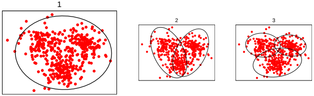


$\hat{p}_1$ vs $\hat{p}_2$



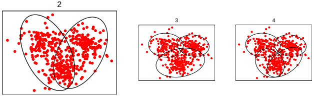


$\hat{p}_1$ vs $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n}$



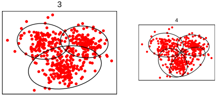
$\ldots$

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



$\ldots$

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

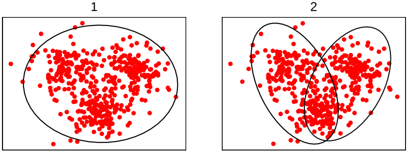


$\ldots$

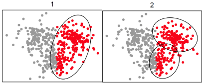
Carnegie Mellon University

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

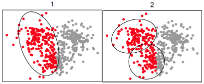
$\hat{p}_1$ vs $\hat{p}_2$

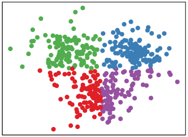


$\hat{p}_1$ vs $\hat{p}_2$

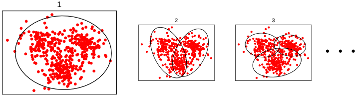


$\hat{p}_1$ vs $\hat{p}_2$

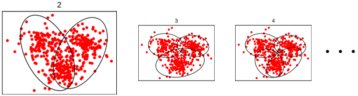




$\hat{p}_1$ vs $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n}$



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



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