Inference for Clustering and Anomaly Detection

Purvasha Chakravarti

Department of Statistics & Data Science

Larry Wasserman
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?

How many clusters are “really” there?

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))
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))
1. Clustering

How can we perform clustering that results in statistically significant clusters?
### Sections of the talk

<table>
<thead>
<tr>
<th>1. Clustering</th>
<th>2. Anomaly Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>How can we perform clustering that results in statistically significant clusters?</td>
<td>In high energy physics, how can we detect new signals in experimental data in a model-independent way?</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>1. Clustering</th>
<th>2. Anomaly Detection</th>
</tr>
</thead>
</table>

**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)

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.

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

   $C_k$: $k$th cluster and $X_k$: $k$th cluster mean.

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

4. Works well in HDLSS data.
Significant Clustering via SigClust: How it works!
Proposed by Liu, Hayes, Nobel and Marron (2008) (Liu et al., 2008)

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

   \[ H_0 : X_1, \ldots, X_n \sim \mathcal{N}(\mu, \Sigma) \text{ versus } \]
   \[ H_1 : X_1, \ldots, X_n \sim f(\cdot), \text{ which is a non-Gaussian distribution.} \]
Significant Clustering via SigClust: How it works!

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

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

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

2. Uses 2-means clustering and the Cluster Index for the test statistic.
   \[
   CI = \frac{\sum_{k=1}^{2} \sum_{j \in C_k} \|X_j - \overline{X}^k\|^2}{\sum_{j=1}^{n} \|X_j - \overline{X}\|^2},
   \]
   \[ C_k : k^{th} \text{ cluster and } \overline{X}^k : k^{th} \text{ cluster mean.} \]

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

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

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

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

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

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

4. Works well in HDLSS data.
Power of SigClust: Low power in some cases
Power of SigClust: Low power in some cases

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

\[ X_1, \ldots, X_n \sim \frac{1}{2} N(-\mu, \Sigma) + \frac{1}{2} N(\mu, \Sigma), \quad \mu = \left( \frac{a}{2}, 0, \ldots, 0 \right), \]
Power of SigClust: Low power in some cases

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

\[ X_1, \ldots, X_n \sim \frac{1}{2} N(-\mu, \Sigma) + \frac{1}{2} N(\mu, \Sigma), \mu = \left( \frac{a}{2}, 0, \ldots, 0 \right), \text{ and } \Sigma \text{ is diagonal} \]
\[ \sigma_1^2, \sigma_2^2 > \sigma_3^2 \geq \ldots \geq \sigma_d^2. \text{ Under some symmetry assumptions,} \]
Power of SigClust: Low power in some cases

<table>
<thead>
<tr>
<th>Theorem 1 (Chakravarti, Purvasha et al. (2019))</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1, \ldots, X_n \sim \frac{1}{2}N(-\mu, \Sigma) + \frac{1}{2}N(\mu, \Sigma)$, $\mu = \left(\frac{a}{2}, 0, \ldots, 0\right)$, and $\Sigma$ is diagonal $\sigma_1^2, \sigma_2^2 &gt; \sigma_3^2 \geq \ldots \geq \sigma_d^2$. Under some symmetry assumptions,</td>
</tr>
<tr>
<td>• if $\sigma_2^2 &gt; \frac{\pi}{2} \mathbb{E}[X_{i1}</td>
</tr>
<tr>
<td>$\frac{\pi}{2} \mathbb{E}[X_{i1}</td>
</tr>
</tbody>
</table>

Carnegie Mellon University
Power of SigClust: Low power in some cases

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

\[ X_1, \ldots, X_n \sim \frac{1}{2} N(-\mu, \Sigma) + \frac{1}{2} N(\mu, \Sigma), \quad \mu = \left( \frac{a}{2}, 0, \ldots, 0 \right), \text{ and } \Sigma \text{ is diagonal} \]

\[ \sigma^2_1, \sigma^2_2 > \sigma^2_3 \geq \ldots \geq \sigma^2_d. \]  

Under some symmetry assumptions,

- if \( \sigma^2_2 > \frac{\pi}{2} \mathbb{E}[X_{i1}|X_{i1} > 0]^2 \), then \( \lim_{n \to \infty} \text{Power}_n(a) < 1 \),

\[ \frac{\pi}{2} \mathbb{E}[X_{i1}|X_{i1} > 0]^2 \approx \sigma^2_1 + \frac{a^2}{4} \text{ for small } a. \]

---

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 $y \in \mathbb{R}^d$, 

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

where $\pi_k$ are the mixing proportions ($0 < \pi_k < 1, \sum_k \pi_k = 1$).
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 $y \in \mathbb{R}^d$,

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

where $\pi_k$ 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).
Proposed test: Relative Information Fit Test (RIFT)

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

\[ \hat{p}_1 \]

\[ D_1 \]

\[ \hat{p}_2 \]

\[ D_2 \]
Proposed test: Relative Information Fit Test (RIFT)

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

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

\[ \hat{p}_1, \hat{p}_2 \]

We test, conditioned on \( D_1 \), \( H_0 : \Gamma \leq 0 \) versus \( H_1 : \Gamma > 0 \).
Proposed test: Relative Information Fit Test (RIFT)

\[ \hat{\rho}_1, \hat{\rho}_2 \]

\[ \hat{\Gamma} = \frac{1}{n} \sum_{i \in D_2} R_i, \quad R_i = \log \left( \frac{\hat{\rho}_2(X_i)}{\hat{\rho}_1(X_i)} \right) \]

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

\[ \sqrt{n} \left( \hat{\Gamma} - \Gamma \right) / \tau \sim N(0, 1) \quad \Longrightarrow \quad \text{Reject } H_0 \text{ if } \hat{\Gamma} > \frac{z_{\alpha} \hat{\tau}}{\sqrt{n}}. \]
Power of RIFT 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 $P(\hat{\Gamma} > z_{\alpha} \hat{t} / \sqrt{n}) \to 1$ as $n \to \infty$. 
Power of RIFT converges to 1!

Power converges to 1!

\(\mathcal{P}_1: \text{Normals, } \mathcal{P}_2: \text{mixtures of two Normals.}\)

**Lemma 2**

Suppose that \( p \in \mathcal{P}_2 - \mathcal{P}_1. \) Then \( P(\hat{\tau} > z_\alpha \hat{\tau}/\sqrt{n}) \to 1 \) 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)!
Comparisons 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 \( n \to \infty \).
Comparisons 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
The ATLAS and the CMS experiments at the LHC

CMS experiment

ATLAS experiment
Events from the experiments
The Standard Model of particle physics

<table>
<thead>
<tr>
<th>three generations of matter (fermions)</th>
<th>interactions / force carriers (bosons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>mass charge spin</td>
<td>mass charge spin</td>
</tr>
<tr>
<td>$u$ (up)</td>
<td>$g$ (gluon)</td>
</tr>
<tr>
<td>$c$ (charm)</td>
<td>$H$ (higgs)</td>
</tr>
<tr>
<td>$t$ (top)</td>
<td>$\gamma$ (photon)</td>
</tr>
<tr>
<td>$d$ (down)</td>
<td>$\gamma$ (photon)</td>
</tr>
<tr>
<td>$s$ (strange)</td>
<td>$\gamma$ (photon)</td>
</tr>
<tr>
<td>$b$ (bottom)</td>
<td>$\gamma$ (photon)</td>
</tr>
<tr>
<td>$e$ (electron)</td>
<td>$Z$ (Z boson)</td>
</tr>
<tr>
<td>$\mu$ (muon)</td>
<td>$W$ (W boson)</td>
</tr>
<tr>
<td>$\tau$ (tau)</td>
<td>$W$ (W boson)</td>
</tr>
<tr>
<td>$\nu_e$ (electron neutrino)</td>
<td>$\nu_e$ (electron neutrino)</td>
</tr>
<tr>
<td>$\nu_\mu$ (muon neutrino)</td>
<td>$\nu_\mu$ (muon neutrino)</td>
</tr>
<tr>
<td>$\nu_\tau$ (tau neutrino)</td>
<td>$\nu_\tau$ (tau neutrino)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\approx 2.2$ MeV/c²</th>
<th>$\approx 1.28$ GeV/c²</th>
<th>$\approx 173.1$ GeV/c²</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{2}{3}$</td>
<td>$\frac{2}{3}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$1$</td>
</tr>
<tr>
<td>$\approx 4.7$ MeV/c²</td>
<td>$\approx 96$ MeV/c²</td>
<td>$\approx 4.18$ GeV/c²</td>
</tr>
<tr>
<td>$-\frac{1}{3}$</td>
<td>$-\frac{1}{3}$</td>
<td>$-\frac{1}{3}$</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$1$</td>
</tr>
<tr>
<td>$\approx 0.511$ MeV/c²</td>
<td>$\approx 105.66$ MeV/c²</td>
<td>$\approx 1.7768$ GeV/c²</td>
</tr>
<tr>
<td>$-1$</td>
<td>$-1$</td>
<td>$-1$</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$1$</td>
</tr>
<tr>
<td>$\approx 91.19$ GeV/c²</td>
<td>$\approx 80.39$ GeV/c²</td>
<td></td>
</tr>
<tr>
<td>$1$</td>
<td>$1$</td>
<td>$1$</td>
</tr>
</tbody>
</table>
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.

\[ q = (1 - \lambda) p_b + \lambda p_s \]

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.

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

\[ q = (1 - \lambda)p_b + \lambda p_s, \quad \text{No signal: } \lambda = 0. \]
Model-dependent supervised methods

Two sources of data are at hand:

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

  Background: \( X_1, \ldots, X_m \sim p_b \)
  
  Signal: \( Y_1, \ldots, Y_n \sim p_s \)

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

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

Train a classifier \( h \) to separate signal from background.
Model-dependent supervised methods

Two sources of data are at hand:

- **Background + signal** (Monte Carlo) sample - labelled observations
  
  Background: \( X_1, \ldots, X_m \sim p_b \)
  
  Signal: \( Y_1, \ldots, Y_n \sim p_s \)

- **Background + possible signal** (experimental) sample - unlabelled observations
  
  Experimental: \( W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s \)
Model-dependent supervised methods

Two sources of data are at hand:

- **Background + signal (Monte Carlo) sample - labelled observations**
  
  Background: \( X_1, \ldots, X_m \sim p_b \)
  
  Signal: \( Y_1, \ldots, Y_n \sim p_s \)

- **Background + possible signal (experimental) sample - unlabelled observations**
  
  Experimental: \( W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s \)

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.

\[ \lambda = 0 \text{ vs } H_1: 0 < \lambda < 1: \]

\[ L_q(\lambda) = \prod_i [(1 - \lambda) + \lambda \psi(W_i)] \]

\[ \psi = \frac{p_s}{p_b}. \]

The membership probabilities can be written as:

\[ h(z) = \hat{P}(Z \text{ is signal} | Z = z) = \frac{n p_s(z)}{n p_s(z) + m p_b(z)} = n \frac{\psi(z)}{n \psi(z) + m}. \]

We can estimate \( \hat{\psi}(z) = m h(z) \frac{n}{n - h(z)}. \)
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 \( H_1 : 0 < \lambda < 1 \):

\[
\frac{L_q(\lambda)}{L_q(0)} = \prod_i [(1 - \lambda) + \lambda \psi(W_i)], \quad \psi = \frac{p_s}{p_b}.
\]

The membership probabilities \( h \) can be written as:

\[
h(z) = \hat{P}(Z \text{ is signal } | Z = z) = \frac{n \psi(z)}{n \psi(z) + m}.
\]

We can estimate \( \hat{\psi}(z) = \frac{m h(z)}{n (1 - h(z))} \).
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 \( H_1 : 0 < \lambda < 1 \):

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

- The membership probabilities \( h \) can be written as:

\[
h(z) = \hat{P}(Z \text{ is signal} | Z = z) = \frac{np_s(z)}{np_s(z) + mp_b(z)} = \frac{n\psi(z)}{n\psi(z) + m}.
\]
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 $H_1 : 0 < \lambda < 1$:

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

- The membership probabilities $h$ can be written as:

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

- We can estimate

$$\hat{\psi}(z) = \frac{mh(z)}{n(1 - h(z))}.$$
Model-dependent supervised methods test statistics

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

$$\frac{L_q(\lambda)}{L_q(0)} = \prod_i [(1 - \lambda) + \lambda \psi(W_i)], \quad \psi = \frac{p_s}{p_b}.$$
Model-dependent supervised methods test statistics

• Likelihood Ratio on the \( W_i \)'s for \( H_0 : \lambda = 0 \) vs \( H_1 : 0 < \lambda < 1 \):

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

1 Likelihood Ratio Test Statistic:

\[
LRT = 2 \sum_i \log \left( (1 - \hat{\lambda}_{MLE}) + \hat{\lambda}_{MLE} \psi(W_i) \right)
\]
Likelihood Ratio on the $W_i$'s for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:

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

1. **Likelihood Ratio Test Statistic:**

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

2. **Score Test Statistic:**

$$S = \frac{1}{N} \sum_{i=1}^{N} \hat{\psi}(W_i).$$
Model-dependent supervised methods test statistics

- Likelihood Ratio on the $W_i$’s for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:
  \[
  \frac{L_q(\lambda)}{L_q(0)} = \prod_i [(1 - \lambda) + \lambda \psi(W_i)], \quad \psi = \frac{p_s}{p_b}.
  \]

1. Likelihood Ratio Test Statistic:
  \[
  LRT = 2 \sum_i \log \left((1 - \hat{\lambda}_{MLE}) + \hat{\lambda}_{MLE} \psi(W_i)\right)
  \]

2. Score Test Statistic:
  \[
  S = \frac{1}{N} \sum_{i=1}^{N} \hat{\psi}(W_i).
  \]

- 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**
  
  Background: \( X_1, \ldots, X_m \sim p_b \)

- **Background + possible signal (experimental) sample - unlabelled observations**
  
  Experimental: \( W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s \)
Solution: Model-independent methods

Two sources of data are at hand:

- **Background (Monte Carlo) sample - labelled observations**
  \[ X_1, \ldots, X_m \sim p_b \]

- **Background + possible signal (experimental) sample - unlabelled observations**
  \[ W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s \]

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

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

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

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

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
  
  Background: \( X_1, \ldots, X_m \sim p_b \)

- Background + possible signal (experimental) sample - unlabelled observations
  
  Experimental: \( W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s \)

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

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

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

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

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 $H_1 : 0 < \lambda < 1$:

$$\frac{L_q(\lambda)}{L_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.$$
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 $\hat{\tilde{\psi}}(z)$. 
Proposed test statistics

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

  \[
  \frac{L_q(\lambda)}{L_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.
  \]

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

- Likelihood Ratio Test Statistic:

  \[
  \text{LRT} = 2 \sum_i \log \tilde{\psi}(W_i).
  \]
Proposed test statistics

- Likelihood Ratio on the $W_i$’s for $H_0 : \lambda = 0$ vs $H_1 : 0 < \lambda < 1$:
  \[
  \frac{L_q(\lambda)}{L_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = q/p_b.
  \]

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

1. Likelihood Ratio Test Statistic:
   \[
   \text{LRT} = 2 \sum_i \log \tilde{\psi}(W_i).
   \]

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

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

$$\frac{L_q(\lambda)}{L_q(0)} = \prod_i \tilde{\psi}(W_i), \quad \tilde{\psi} = \frac{q}{p_b}.$$

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

1. Likelihood Ratio Test Statistic:

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

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.
Kaggle’s Higgs boson challenge

- Data provided by ATLAS.

- 15 variables.

---

1https://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.

---

1https://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.

---

1https://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, $\lambda$.

1https://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, $\lambda$.
- Compare power of the methods in detecting the Higgs boson.

---


Purvasha Chakravarti (CMU)
Power - simulations where the Higgs boson is detected

\( \lambda \) is the proportion of signal in the experimental data set.

100 simulations.

Model-dependent methods that have signal labels.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Signal Strength (( \lambda ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>Supervised LRT</td>
<td>Asymptotic</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Permutation</td>
<td>99</td>
</tr>
<tr>
<td>Supervised Score</td>
<td>Permutation</td>
<td>99</td>
</tr>
</tbody>
</table>
Power - simulations where the Higgs boson is detected

$\lambda$ is the proportion of signal in the experimental data set.

100 simulations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>0.15</th>
<th>0.1</th>
<th>0.07</th>
<th>0.05</th>
<th>0.01</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised LRT</td>
<td>Asymptotic</td>
<td>99</td>
<td>70</td>
<td>22</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Permutation</td>
<td>99</td>
<td>93</td>
<td>59</td>
<td>19</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Supervised Score</td>
<td>Permutation</td>
<td>99</td>
<td>94</td>
<td>80</td>
<td>51</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Semi-Supervised LRT</td>
<td>Asymptotic</td>
<td>99</td>
<td>63</td>
<td>16</td>
<td>20</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Permutation 1</td>
<td>99</td>
<td>60</td>
<td>17</td>
<td>19</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Semi-Supervised AUC</td>
<td>Asymptotic</td>
<td>96</td>
<td>63</td>
<td>17</td>
<td>17</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Permutation 1</td>
<td>97</td>
<td>62</td>
<td>18</td>
<td>16</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Permutation 2</td>
<td>100</td>
<td>74</td>
<td>38</td>
<td>23</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>NN Two-Sample</td>
<td>Permutation</td>
<td>74</td>
<td>33</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>
Density of the training data variables, $\lambda = 0.15$
Identifying the active subspace that explains the classifier

- Consider $\nabla_z \tilde{h}(z)$.
Identifying the active subspace that explains the classifier

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

- Perform Principal Component Analysis (PCA) or sparse PCA on $\nabla_z \tilde{h}(z)$. 
Identifying the active subspace that explains the classifier

- Consider \( \nabla_z \tilde{h}(z) \).

- Perform Principal Component Analysis (PCA) or sparse PCA on \( \nabla_z \tilde{h}(z) \).

- Let \( \mathbf{m}_1, \mathbf{m}_2, \ldots \) be the leading eigenvectors.
Identifying the active subspace that explains the classifier

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

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

\[
\nabla_z h(z) - \nabla_z h_j = \nabla_z \tilde{h}(W_j) \text{ using a local linear smoother on } \tilde{h}.
\]
Active subspace of $\tilde{h}(\cdot)$

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

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

- Perform Principal Component Analysis (PCA) or sparse PCA on $H = (\nabla_z h_1, \nabla_z h_2, \ldots, \nabla_z h_N)^T$. 

$E[\nabla_z \tilde{h}]$, $\hat{m}_1$, $\hat{m}_2$, ..., $\hat{m}_1$, $\hat{m}_2$, ..., $\nabla_z h_j = \frac{1}{N} \sum_{j=1}^{N} \nabla_z h_j$, $\hat{m}_1$, $\hat{m}_2$, ...
Active subspace of $\tilde{h}(\cdot)$

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

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

- Perform Principal Component Analysis (PCA) or sparse PCA on $H = (\nabla_z h_1, \nabla_z h_2, \ldots, \nabla_z h_N)^T$.

- Let $m_1, m_2, \ldots$ be the leading eigenvectors - $\hat{m}_1, \hat{m}_2, \ldots$. 
Active subspace of $\tilde{h}(\cdot)$

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

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

- Perform Principal Component Analysis (PCA) or sparse PCA on $H = (\nabla_z h_1, \nabla_z h_2, \ldots, \nabla_z h_N)^T$.

- Let $m_1, m_2, \ldots$ be the leading eigenvectors - $\hat{m}_1, \hat{m}_2, \ldots$.

- $\mathbb{E} \left[ \nabla_z \tilde{h} \right], m_1, m_2, \ldots - \nabla_z h_j = \frac{1}{N} \sum_{j=1}^{N} \nabla_z h_j, \hat{m}_1, \hat{m}_2, \ldots$. 

Purvasha Chakravarti (CMU)

Topics in Inference

April 27, 2020

31 / 35
Active subspace for $\tilde{h}(\cdot)$ when $\lambda = 0.15$

Mean Gradient

$$\left( \mathbb{E} \left[ \nabla_z \tilde{h} \right] \right)$$

First Eigenvector

$$(m_1)$$

Second Eigenvector

$$(m_2)$$
Active subspace for $\tilde{h}(\cdot)$ when $\lambda = 0.15$

The vectors capture the variable dependencies that influence the classifier.

**Mean Gradient**

\[
\left( \mathbb{E} \left[ \nabla_z \tilde{h} \right] \right)
\]

**First Eigenvector**

\[
(m_1)
\]

**Second Eigenvector**

\[
(m_2)
\]
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 committee members!

Questions?
References


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.**
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.**
  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.**
  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.**
  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.
Asymptotic normality of $\hat{\Gamma}$

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

**Theorem 3**

Assume each $\hat{\mu}_i \in A$, a compact set and the eigenvalues of $\hat{\Sigma}_i \in [c_1, c_2]$. Let $Z \sim N(0, \tau^2)$ where $\tau^2 = \mathbb{E}[(\tilde{R}_i - \Gamma)^2 | D_1]$. Then, under $H_0$

$$\sup_t \left| P(\sqrt{n}(\hat{\Gamma} - \Gamma) \leq t | D_1) - P(Z \leq t) \right| \leq \frac{C}{\sqrt{n}}$$

(1)

where $C$ is a constant that does not depend on $D_1$. 
Median RIFT (M-RIFT): A more robust test.

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

- Robustified version: \( \tilde{\Gamma} = \text{Median}_p[R] \), where \( R = \log \hat{\rho}_2(X)/\hat{\rho}_1(X) \).

- Sample median of \( R_1, \ldots, R_n \) is a consistent estimator, where \( R_i = \log \hat{\rho}_2(X_i)/\hat{\rho}_1(X_i) \).

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

- Reject $H_{0j}$ if
  \[ \max_s \hat{\Gamma}_{js} > \frac{z_{\alpha/m_j} \hat{\tau}_{js}}{\sqrt{n}} \]
  where $m_j = K_n - j$, $\hat{\Gamma}_{js} = \frac{1}{n} \sum_{i \in \mathcal{D}_2} R_i$, $R_i = \log \left( \frac{\hat{p}_s(X_i)}{\hat{p}_j(X_i)} \right)$ and
  \[ \hat{\tau}_{js}^2 = \frac{1}{n} \sum_{i \in \mathcal{D}_2} (R_i - \bar{R})^2. \]

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

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

**Lemma 4**

Under $H_{0j}$,

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

**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. \( \alpha = 0.05 \).
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
  \[ W_1, \ldots, W_N \sim q = (1 - \lambda)p_b + \lambda p_s. \]

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.
Confidence Intervals for AUC

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

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

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

- 100(1 – \( \alpha \))% confidence interval for AUC \( \theta \) is given by

\[
\hat{\theta} \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{\theta})},
\]

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

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

Value
Density

Class
-2.5
0.0
2.5
-2
0
2
4
5
6
7

Sublead eta
Sublead phi
All pt
Met_sumet
Lead pt
Lead eta
Sublead pt
Lept eta
Lept phi
Met
Met phi
Tau pt
Tau eta
Tau phi
Lept pt

Purvasha Chakravarti (CMU)
Topics in Inference
April 27, 2020 12 / 14
Hierarchical RIFT (H-RIFT)
Hierarchical RIFT (H-RIFT)

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]
Hierarchical RIFT (H-RIFT)

\[ \hat{p}_1 \ vs \ \hat{p}_2 \]
Hierarchical RIFT (H-RIFT)

\( \hat{p}_1 \ vs \ \hat{p}_2 \)

\( \hat{p}_1 \ vs \ \hat{p}_2 \)

\( \hat{p}_1 \ vs \ \hat{p}_2 \)

\( \hat{p}_1 \ vs \ \hat{p}_2 \)
Hierarchical RIFT (H-RIFT)

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

\[ \hat{p}_1 \text{ vs } \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_2 \text{ vs } \hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n} \]

\[ \ldots \]
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_2 \text{ vs } \hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_2 \text{ vs } \hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_3 \text{ vs } \hat{p}_4, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_3 \text{ vs } \hat{p}_4, \ldots, \hat{p}_{K_n} \]
Hierarchical RIFT (H-RIFT) vs Sequential RIFT (S-RIFT)

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_2 \text{ vs } \hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_3 \text{ vs } \hat{p}_4, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2 \]

\[ \hat{p}_1 \text{ vs } \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_2 \text{ vs } \hat{p}_3, \hat{p}_4, \ldots, \hat{p}_{K_n} \]

\[ \hat{p}_3 \text{ vs } \hat{p}_4, \ldots, \hat{p}_{K_n} \]