Detecting Distributional Differences in Labeled Sequences of Tropical Cyclone Satellite Imagery

Ann B. Lee
Department of Statistics & Data Science
Carnegie Mellon University
(Joint work with Trey McNeely, Galen Vincent, Kim Wood, and Rafael Izbicki)
DETECTING DISTRIBUTIONAL DIFFERENCES IN LABELED SEQUENCE DATA WITH APPLICATION TO TROPICAL CYCLONE SATELLITE IMAGERY

By Trey McNeely$^{1,a}$, Galen Vincent$^{1,b}$, Kimberly M. Wood$^{2,d}$, Rafael Izbicki$^{3,e}$ and Ann B. Lee$^{1,c}$

$^{1}$Department of Statistics and Data Science, Carnegie Mellon University, $^{a}$treymcneely@gmail.com,
$^{b}$galenbvincent@gmail.com, $^{c}$annlee@andrew.cmu.edu
$^{2}$Department of Geosciences, Mississippi State University, $^{d}$kimberly.wood@msstate.edu
$^{3}$Department of Statistics, Federal University of São Carlos, $^{e}$rafaelizbicki@gmail.com
Both Methodology and Applied Relevance

• Motivating application:
  • To identify spatiotemporal patterns in tropical cyclone (TC) satellite imagery that lead up to an upcoming rapid intensity change event.

• Requires new methodology:
  • For detecting distributional differences between sequences of images
    \( S_{<t} = \{X_{t-T}, X_{t-T+1}, \ldots, X_t\} \) preceding an event \( (Y_t=1) \) vs non-event \( (Y_t=0) \).
  • The problem is difficult because the data are high-dimensional.
  • The data \( \{(S_{<t}, Y_t)\}_{t \geq 0} \) are also not IID because of strong temporal dependence.
Tropical Cyclones (TCs) are Rapidly Rotating Systems
Develop over Warm Tropical Waters

- Because TCs develop far from land-based observing networks, geostationary satellite imagery (GOES) is critical to monitor these storms.
HURRICANE STRUCTURE
IN THE NORTHERN HEMISPHERE

Outflow cirrus shield

Warm rising air

Cold falling air

Eye wall

Eye

Rain bands

Storm rotation COUNTERCLOCKWISE

Left: Edouard 2014 (95 kt; Category 2); Right: Nicole 2016 (47 kt; TS)
Spatio-Temporal Information in IR Imagery Underutilized

Trajectory Forecasts vs. TC Short-term Intensity Forecasts (24-hr)

“RI event” = increase in wind speed of ≥25 knots in 24 hrs
Two Databases
TC location & intensity

1. **HURDAT2**
   - Hurricane best-track data
   - 6-hr resolution (1979-2020)
   - TC location, intensity
Two Databases
TC location & intensity + GOES images

1. **HURDAT2**
   - Hurricane best-track data
   - 6-hr resolution (1979-2020)
   - TC location, intensity

2. **MERGIR**
   - Geostationary satellite (GOES) imagery
   - 4-km, 30-min resolution
   - 2000-2020
Evolution of TC Convective Structure as “Structural Trajectories” $S_t$ of Interpretable Functions $X_t$
Evolution of TC Convective Structure as “Structural Trajectories” $S_{<t}$ of Interpretable Functions $X_t$

Structural trajectory is a 24h sequence of cont. functions (at 30 min time res). “Hovmöller diagram”

$$S_{<t} = \{X_{t-24h}, X_{t-23.5h}, X_{t-23h}, \ldots, X_t\}$$
Main Questions as a Two-Sample Testing Problem

\[ Y_t = \begin{cases} 
1 & \text{if RI event at time } t, \\
0 & \text{otherwise} 
\end{cases} \]

\( H_0 : p(s_{<t}|Y_t = 1) = p(s_{<t}|Y_t = 0) \) for all \( s_{<t} \in \mathcal{S} \), versus

\( H_1 : p(s_{<t}|Y_t = 1) \neq p(s_{<t}|Y_t = 0) \) for some \( s_{<t} \in \mathcal{S} \).

Does the distribution of structural trajectories differ between the lead-up to a RI vs. non-RI event? (Statistical significance)

If there is a difference between the distributions, how do they differ? (Scientific interpretability)
Main Questions as a Two-Sample Testing Problem

\[ Y_t = \begin{cases} 
  1 & \text{if RI event at time } t, \\
  0 & \text{otherwise} 
\end{cases} \]

\[ H_0 : p(s_{<t} | Y_t = 1) = p(s_{<t} | Y_t = 0) \text{ for all } s_{<t} \in S, \text{ versus} \]
\[ H_1 : p(s_{<t} | Y_t = 1) \neq p(s_{<t} | Y_t = 0) \text{ for some } s_{<t} \in S. \]

Does the distribution of structural trajectories differ between the lead-up to a RI vs. non-RI event? (Statistical significance)

If there is a difference between the distributions, how do they differ? (Scientific interpretability)
Why the Two-Sample Test is Challenging …

\[ H_0 : p(s_{<t}|Y_t = 1) = p(s_{<t}|Y_t = 0) \text{ for all } s_{<t} \in S, \text{ versus} \]
\[ H_1 : p(s_{<t}|Y_t = 1) \neq p(s_{<t}|Y_t = 0) \text{ for some } s_{<t} \in S. \]

- The complexity of the data itself, with one observation representing an entire sequence \( S_{<t} \) of functions

\[ S_{<t} = \{X_{t-T}, X_{t-T+1}, \ldots, X_t\} \]

- Dependence between labels \( Y_t \) (and sequences \( S_{<t} \)) at nearby time points \( t \)

- IID data ⇒ “Dependent Identically Distributed” (DID) sequence data

\[ \{(S_{<t}, Y_t)\}_{t \geq 0} \]
Two-Sample Test via Regression (HighDim IID data)


Suppose we have two samples:

$$S^0_1, \ldots, S^0_{n_0} \sim P_0 \quad \text{and} \quad S^1_1, \ldots, S^1_{n_1} \sim P_1$$

A two sample-test would ask whether $P_0$ and $P_1$ are the same; i.e., it would test the null hypothesis

$$H_0 : p(s|Y = 0) = p(s|Y = 1) \quad \text{for all } s \in S$$
Two-Sample Test via Regression (HighDim IID data)


Suppose we have two samples:

\[ S_1^0, \ldots, S_{n_0}^0 \sim P_0 \quad \text{and} \quad S_1^1, \ldots, S_{n_1}^1 \sim P_1 \]

A two sample-test would ask whether \( P_0 \) and \( P_1 \) are the same; i.e., it would test the null hypothesis

\[ H_0 : p(s|Y = 0) = p(s|Y = 1) \quad \text{for all} \quad s \in S \]

By Bayes rule, this is equivalent to testing

\[ H_0 : \mathbb{P}(Y = 1|S = s) = \mathbb{P}(Y = 1) \quad \text{for all} \quad s \in S \]
Convert 2-sample testing to a regression problem

Our null and alternative hypotheses are

\[ H_0 : \ P(Y = 1 | S = s) = P(Y = 1) \text{ for all } s \in S \]
\[ H_1 : \ P(Y = 1 | S = s) \neq P(Y = 1) \text{ for some } s \in S \]

Define the regression function \( m_{\text{post}}(s) := P(Y = 1 | S = s). \)
Let \( \hat{m}(s) \) be an estimate of \( m_{\text{post}}(s) \) based on train data \( \mathcal{T} = \{(S_i, Y_i)\}_{i=1}^n \).
Let \( \hat{m}_{\text{prior}}(s) = \frac{1}{n} \sum_{i=1}^n I(Y_i = 1) \) be an estimate of \( m_{\text{prior}} := P(Y = 1). \)
Convert 2-sample testing to a regression problem

Our null and alternative hypotheses are

\[ H_0 : \Pr(Y = 1 | S = s) = \Pr(Y = 1) \text{ for all } s \in S \]
\[ H_1 : \Pr(Y = 1 | S = s) \neq \Pr(Y = 1) \text{ for some } s \in S \]

Define the regression function \( m_{\text{post}}(s) := \Pr(Y = 1 | S = s) \).
Let \( \hat{m}(s) \) be an estimate of \( m_{\text{post}}(s) \) based on train data \( T = \{(S_i, Y_i)\}_{i=1}^n \).
Let \( \hat{m}_{\text{prior}}(s) = \frac{1}{n} \sum_{i=1}^n I(Y_i = 1) \) be an estimate of \( m_{\text{prior}} := \Pr(Y = 1) \).

Define the “local posterior difference” (LPD) at evaluation points \( \mathcal{V} \subset S \):

\[ \lambda(s) := \hat{m}_{\text{post}}(s) - \hat{m}_{\text{prior}} \]

Our global test statistic is

\[ \lambda := \frac{1}{|\mathcal{V}|} \sum_{s \in \mathcal{V}} \lambda(s)^2 \]
Can Detect Distributional Differences in Galaxy Images for HighSF and LowSF Samples [Freeman, Kim & Lee, MNRAS 2017]

Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [40].
Can Detect Distributional Differences in Galaxy Images for HighSF and LowSF Samples [Freeman, Kim & Lee, MNRAS 2017]

Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [40].

But these are I.I.D data and not dependent sequence data...
Dependence Settings for Labeled Sequence Data

\[ S_{<t} = \{X_{t-T}, X_{t-T+1}, \ldots, X_t\} \]

I.I.D pairs

Y's conditionally independent

(a) Setting A: \( \{(S_{<t}, Y_t)\}_{t \geq 0} \) with no temporal dependence between pairs \((S_{<t}, Y_t)\) for different \(t\).

(b) Setting B: \( Y_t \) conditionally independent of \( Y_{t-1} \) given \( S_{<t} \); \( S_{<t} \) is autocorrelated.

(c) Setting C: \( Y_t \) conditionally dependent on \( Y_{t-1} \) given \( S_{<t} \); \( S_{<t} \) and \( Y_t \) are each autocorrelated.

In Settings A and B: Labels Y are conditionally independent given S

⇒ Labels Y are exchangeable under \( H_0 \). A permutation test would be valid [Kim et al 2019]
TC Data are Not Exchangeable.

\[ S_{<t} = \{X_{t-T}, X_{t-T+1}, \ldots, X_t\} \]

(a) Setting A: \( \{(S_{<t}, Y_t)\}_{t \geq 0} \) with no temporal dependence between pairs \((S_{<t}, Y_t)\) for different \(t\).

(b) Setting B: \( Y_t \) conditionally independent of \( Y_{t-1} \) given \( S_{<t} \); \( S_{<t} \) is autocorrelated.

(c) Setting C: \( Y_t \) conditionally dependent on \( Y_{t-1} \) given \( S_{<t} \); \( S_{<t} \) and \( Y_t \) are each autocorrelated.

In TC data, we have auto-correlation in \( Y \) which is inherent or governed by unobserved quantities (Setting C) \( \Rightarrow \) Permutation tests are not valid.
For permutation test:

- Estimate $m_{\text{post}}(s) := P(Y_t = 1|S_{<t} = s)$ using labeled train data $\{(S_{<t}, Y_t)\}_{t \in T_1}$
- Compute test statistic $\lambda = \sum_{s \in V} \lambda^2(s)$, where $\lambda(s) = \hat{m}_{\text{post}}(s) - \hat{m}_{\text{prior}}$
- To estimate the null distribution of $\lambda$:
  - Permute original labels $\{Y_t\}_{t \in T_1}$
  - Recompute test statistic $\lambda$
Permutation Test $\Rightarrow$ Markov Chain (MC) Bootstrap Test

For permutation test:

- Estimate $m_{\text{post}}(s) := \mathbb{P}(Y_t = 1|S_{<t} = s)$ using labeled train data $\{(S_{<t}, Y_t)\}_{t \in \mathcal{T}_1}$
- Compute test statistic $\lambda = \sum_{s \in \mathcal{Y}} \lambda^2(s)$, where $\lambda(s) = \hat{m}_{\text{post}}(s) - \hat{m}_{\text{prior}}$
- To estimate the null distribution of $\lambda$:
  - Permute original labels $\{Y_t\}_{t \in \mathcal{T}_1}$
  - Recompute test statistic $\lambda$

Instead, use train data $\{Y_t\}_{t \in \mathcal{T}_2}$ and regression method to estimate

$$m_{\text{seq}}(Y_{t-1}, \ldots, Y_{t-k}) := \mathbb{P}(Y_t = 1|Y_{t-1}, \ldots, Y_{t-k})$$

Draw new labels

$$\tilde{Y}_t \sim \text{Binom}(\hat{\mathbb{P}}(Y_t = 1|Y_{t-1}, \ldots, Y_{t-k})) \quad \text{for} \quad t \in \mathcal{T}_1$$
TC train data: High-res GOES images back to 2000 (~400 TCs to fit regression of Y on S). However, intensity data goes back to 1979 (>1000 TCs to fit MC on labels)

| Table 1 |
| Sample sizes: Data set summary for each category: (i) labeled sequences $(S_{<t}, Y_t)$ used in training, (ii) unlabeled test sequences $S_{<t}$ and (iii) synoptic labels $Y_t$ used when complete trajectories are not needed |

<table>
<thead>
<tr>
<th></th>
<th>NAL</th>
<th>ENP</th>
<th>Total</th>
<th>Year Range</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(i) Training Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Sequences</td>
<td>47,502</td>
<td>31,549</td>
<td>79,051</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RI Sequences</td>
<td>7015</td>
<td>6742</td>
<td>13,757</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW Sequences</td>
<td>5878</td>
<td>7298</td>
<td>13,176</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unique TCs</strong></td>
<td>209</td>
<td>185</td>
<td>394</td>
<td>2000–2012</td>
<td>13</td>
</tr>
<tr>
<td><strong>(ii) Test Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Sequences</td>
<td>28,368</td>
<td>32,817</td>
<td>61,185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RI Sequences</td>
<td>3965</td>
<td>6386</td>
<td>10,351</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW Sequences</td>
<td>3167</td>
<td>7182</td>
<td>10,349</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unique TCs</strong></td>
<td>125</td>
<td>152</td>
<td>277</td>
<td>2013–2020</td>
<td>8</td>
</tr>
<tr>
<td><strong>(iii) Synoptic Labels</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Labels</td>
<td>14,683</td>
<td>15,274</td>
<td>29,957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RI Labels</td>
<td>1850</td>
<td>2462</td>
<td>4312</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RW Labels</td>
<td>1643</td>
<td>2534</td>
<td>4177</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unique TCs</strong></td>
<td>532</td>
<td>589</td>
<td>1121</td>
<td>1979–2012</td>
<td>34</td>
</tr>
</tbody>
</table>
Theorem: MC Bootstrap Test is Valid Asymptotically

Assume:

1. \( \{(S_{<t}, Y_t)\}_{t \geq 0} \) is a stationary sequence

2. \( \{(S_{<t}, Y_t)\}_{t \geq 0} \) satisfies the DAG of Setting C

3. \( \hat{m}_{\text{post}} \) is a continuous function of the train data \( \mathcal{D} := \{(S_{<t}, Y_t)\}_{t \in \mathcal{T}_1} \)

4. The marginal distribution estimator is consistent; that is, the generator of \( \{Y_t^0\}_{t \in \mathcal{T}_1} \) converges to the true generating process of \( \{Y_t\}_{t \in \mathcal{T}_1} \) under \( H_0 \),

\[
G_{\hat{p}_{t^2}} \xrightarrow{\text{Dist}} \quad G^*
\]
Theorem: MC Bootstrap Test is Valid Asymptotically

**Theorem 1.** Assume 1, 2, 3 and 4. Under the null hypothesis,

$$
\lambda(\mathcal{D}_0^{t_2}) \xrightarrow[\mathcal{D}_0^{t_2} \rightarrow \infty]{\text{Dist}} \lambda(\mathcal{D})
$$

It follows from Theorem 1 that type I error is controlled asymptotically:

**Corollary 1 (Type I error control).** Let

$$
\hat{p}_B^{t_2}(\mathcal{D}) := \frac{1}{B+1} \left( 1 + \sum_{b=1}^{B} \mathbb{I}\left( \lambda(\mathcal{D}^{(b)}) > \lambda(\mathcal{D}) \right) \right)
$$

be the Monte Carlo p-value for $H_0$, where $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(B)} \overset{\text{IID}}{\sim} \mathcal{D}_0^{t_2}$. Assume that Assumptions 1, 2, 3 and 4 hold. Then, under the null hypothesis, for any $0 < \alpha < 1$,

$$
\lim_{t_2 \rightarrow \infty} \lim_{B \rightarrow \infty} \mathbb{P}\left( \hat{p}_B^{t_2}(\mathcal{D}) \leq \alpha \right) = \alpha.
$$
Empirical Results for Synthetic Data Support Our Approach

(Left) Permutation test breaks under Setting C. (Right) MC bootstrap test still valid
TC Analysis by Basin: Reject $H_0: p(s_{<t}|Y_t=1) = p(s_{<t}|Y_t=0)$. Now what?

*How do the distributions of the structural trajectories $s_{<t}$ differ?*
TC Analysis by Basin: Reject $H_0: p(s_{<t}|Y_t=1) = p(s_{<t}|Y_t=0)$. Now what?

How do the distributions of the structural trajectories $s_{<t}$ differ?

- Use contributions to test statistic as a local diagnostic. “Local posterior difference” (LPD):

$$\lambda(s) = \hat{P}(Y_t = 1|S_{<t} = s) - \hat{P}(Y_t = 1)$$
Positive LPD identifies trajectories with "high chance of RI"

Negative LPD identifies trajectories with "low chance of RI"

\[ \lambda(s) = \hat{P}(Y_t = 1|S_{<t} = s) - \hat{P}(Y_t = 1) \]
LPDs can also be used to track development of specific TCs

Analysis by basin ⇒ Case study of Hurricane Jose (2017)

- We interpret high LPD as a TC which is “convectively primed” for RI.
- Hurricane Jose was subject to high vertical wind shear (cause of RW) near Sept 9, which our model does not account for.
We have proposed a two-sample test for D.I.D sequence data \( \{(s_{<t}, Y_t)\}_{t \geq 0} \) with interpretable diagnostics. Two key ideas:

- a test statistic based on the posterior difference \( p(Y=1|s) - p(s) \), estimated via a suitable regression method;

- a bootstrap test where we estimate the marginal distribution of \( \{Y_t\}_{t \geq 0} \); consistency guarantees asymptotic validity.
Extend inputs $S$ to include other functional features and data sources.

Can extend to a *conditional* test $H_0: p(s|Y=1,z) = P(z|Y=0,z)$ by considering the posterior differences $P(Y=1|s,z)-P(Y=1|z)$. 
Acknowledgments

- **Trey McNeely** (CMU, now Microsoft Research)
- Galen Vincent (CMU, now Maxar)
- **Dr Kimberly M Wood** (MSU, Geosciences)
- Dr Rafael Izbicki (UFSCar)

*This work is funded in part by NSF DMS-2053804 and NSF PHY-2020295.*