Guiding New Physics Searches with Unsupervised Learning

[DS, Jacques - 1807.06038]

Andrea De Simone

andrea.desimone@sissa.it
New Physics ?

Searches for New Physics Beyond the Standard Model have been negative so far…

**MAYBE:**

1. **New Physics (NP) is not accessible by LHC**
   - new particles are too light/heavy
   - or interacting too weakly

2. **We have not explored all the possibilities**
   - new physics may be buried under large bkg
   - or hiding behind unusual signatures
“Don’t want to miss a thing” (in data)

closer look at current data
get ready for upcoming data from next run

Model-independent search

searches for specific models may be:
- time-consuming
- insensitive to unexpected/unknown processes
Want a statistical test for NP which is:

1. **model-independent:**
   - no assumption about underlying physical model to interpret data
   - more general

2. **non-parametric:**
   - compare two samples as a whole (not just their means, etc.)
   - fewer assumptions, no max likelihood estim.

3. **un-binned:**
   - high-dim feature space partitioned without rectangular bins
   - retain full multi-dim info of data
1. Statistical test of dataset compatibility
   - Nearest-Neighbors Two-Sample Test
   - Identify Discrepancies
   - Include Uncertainties

2. Applications to High-Energy Physics
1. Statistical test of dataset compatibility
   • Nearest-Neighbors Two-Sample Test
     • Identify Discrepancies
     • Include Uncertainties

2. Applications to High-Energy Physics
Two-sample Test

[a.k.a. “homogeneity test”]

Two sets:

Trial: \( \mathcal{T} = \{ x_1, \ldots, x_{N_T} \} \) iid \( \sim p_T \)

Benchmark: \( \mathcal{B} = \{ x'_1, \ldots, x'_{N_B} \} \) iid \( \sim p_B \)

\( x_i, x'_i \in \mathbb{R}^D \)

probability distributions \( p_B, p_T \) unknown

e.g.: simulated SM bkg real measured data

A. De Simone
Two-sample Test

Two sets:

Trial: \( \mathcal{T} = \{x_1, \ldots, x_{N_T}\} \overset{\text{iid}}{\sim} p_T \)

Benchmark: \( \mathcal{B} = \{x'_1, \ldots, x'_{N_B}\} \overset{\text{iid}}{\sim} p_B \)

probability distributions \( p_B, p_T \) unknown

Are \( \mathcal{B}, \mathcal{T} \) drawn from the same prob. distribution?

easy...
> Two-sample Test

Two sets:

Trial: \( \mathcal{T} = \{x_1, \ldots, x_{N_T}\} \overset{iid}{\sim} p_T \)

Benchmark: \( \mathcal{B} = \{x'_1, \ldots, x'_{N_B}\} \overset{iid}{\sim} p_B \)

probability distributions \( p_B, p_T \) unknown

Are \( B, T \) drawn from the same prob. distribution?

... hard!
Two-sample Test

RECIPE:

1. Density Estimator
   - reconstruct PDFs from samples

2. Test Statistic (TS)
   - measure “distance” between PDFs

3. TS distribution
   - associate probabilities to TS under null hypothesis $H_0: p_B = p_T$

4. $p$ -value
   - accept/reject $H_0$
> 1. Density Estimator

Divide the space in squared bins?

 ✓ easy
 ✓ can use simple statistics (e.g. $\chi^2$)
 ✘ hard/slow/impossible in high-$D$

Need un-binned multivariate approach

Find PDFs estimators: $\hat{p}_B(x), \hat{p}_T(x)$
e.g. based on densities of points:

$$\hat{p}_{B,T}(x) = \frac{\rho_{B,T}(x)}{N_{B,T}}$$

Nearest Neighbors!

[Schilling - 1986][Henze - 1988]
[Wang et al. - 2005, 2006]
[Dasu et al. - 2006][Perez-Cruz - 2008]
[Sugiyama et al. - 2011][Kremer et al., 2015]
Fix integer $K$.

Choose query point $x_j$ in $T$ and draw it in $B$. 

> 1. Density Estimator
> 1. Density Estimator

- Fix integer $K$.
- Choose query point $x_j$ in $T$ and draw it in $B$.
- Find the distance $r_{j,B}$ of the $K^{th}$-NN of $x_j$ in $B$. 

![Diagram](image.png)
• Fix integer $K$.

• Choose query point $x_j$ in $T$ and draw it in $B$.

• Find the distance $r_{j,B}$ of the $K^{th}$-NN of $x_j$ in $B$.

• Find the distance $r_{j,T}$ of the $K^{th}$-NN of $x_j$ in $T$. 

\[ > 1. \text{Density Estimator} \]
• Fix integer $K$.

• Choose query point $x_j$ in $T$ and draw it in $B$.

• Find the distance $r_{j,B}$ of the $K^{th}$-NN of $x_j$ in $B$.

• Find the distance $r_{j,T}$ of the $K^{th}$-NN of $x_j$ in $T$.

• Estimate PDFs:

$$\hat{p}_B(x_j) = \frac{K}{N_B} \frac{1}{\omega_D r_{j,B}^D}$$

$$\hat{p}_T(x_j) = \frac{K}{N_T - 1} \frac{1}{\omega_D r_{j,T}^D}$$
> 2. Test Statistic

• Measure of the “distance” between 2 PDFs

• Define **Test Statistic**: (detect under-/over-densities)

\[
TS(B, T) = \frac{1}{N_T} \sum_{j=1}^{N_T} \log \frac{\hat{p}_T(x_j)}{\hat{p}_B(x_j)}
\]

• Related to Kullback-Leibler divergence as:  
  \[TS(B, T) = \hat{D}_{KL}(\hat{p}_T \| \hat{p}_B)\]
  
  \[D_{KL}(p \| q) = \int_{\mathbb{R}^D} p(x) \log \frac{p(x)}{q(x)} dx\]

• From NN-estimated PDFs:  
  \[TS_{\text{obs}} = \frac{D}{N_T} \sum_{j=1}^{N_T} \log \frac{r_{j,B}}{r_{j,T}} + \log \frac{N_B}{N_T - 1}\]

• **Theorem:** this estimator converges to \(D_{KL}(p_B \| p_T)\), in large sample limit
  
  [Wang et al. - 2005, 2006]
3. Test Statistic Distribution

How is TS distributed? **Permutation test!**

Assume $\rho_B = \rho_T$. Union set: $\mathcal{U} = \mathcal{T} \cup \mathcal{B}$

Random reshuffle $\mathcal{U}$

Repeat many times.

Distribution of TS under $H_0$: $f(TS|H_0) \leftarrow \{TS_n\}$
[asymptotically normal with zero mean]
> 4. \( p \)-value

- \( \hat{\mu}, \hat{\sigma} \): mean, variance of TS distribution \( f(TS|H_0) \)

- Standardize the TS: \( TS \rightarrow TS' \equiv \frac{TS - \hat{\mu}}{\hat{\sigma}} \)

- TS’ distributed according to \( f'(TS'|H_0) = \hat{\sigma} f(\hat{\mu} + \hat{\sigma} TS'|H_0) \)

- Two-sided \( p \)-value:

\[
p = 2 \int_{|TS'_\text{obs}|}^{+\infty} f'(TS'|H_0) dTS'
\]

- Equivalent significance: \( Z \equiv \Phi^{-1}(1 - p/2) \)
2D Gaussian Example

\[ p_B = \mathcal{N}(\mu_B, \Sigma_B) \quad p_T = \mathcal{N}(\mu_T, \Sigma_T) \]

\[ \mu_B = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \quad \mu_T = \begin{pmatrix} 1.2 \\ 1.2 \end{pmatrix} \]

\[ \Sigma_B = \Sigma_T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]

exact KL divergence

more data, more power

\[ K = 5, N_{\text{perm}} = 1000 \]

\[ \mu_B = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \quad \mu_T = \begin{pmatrix} 1.15 \\ 1.15 \end{pmatrix} \]

\[ Z = 5\sigma \]
> NN2ST: Summary

INPUT:

Trial sample: \( \mathcal{T} = \{x_1, \ldots, x_{N_T}\} \overset{iid}{\sim} p_T \quad x_i, x'_i \in \mathbb{R}^D \)

Benchmark sample: \( \mathcal{B} = \{x'_1, \ldots, x'_{N_B}\} \overset{iid}{\sim} p_B \)

K: number of nearest neighbors

N\text{perm}: number of permutations

OUTPUT:

\( \rho \)-value of the null hypothesis \( H_0: p_B = p_T \)

[check compatibility between 2 samples]
> **NN2ST: Summary**

Python code:
[github.com/de-simone/NN2ST](https://github.com/de-simone/NN2ST)
1. Statistical test of dataset compatibility

• Nearest-Neighbors Two-Sample Test
• Identify Discrepancies
• Include Uncertainties

2. Applications to High-Energy Physics
> Where are the discrepancies?

**Bonus: Characterize regions with significant discrepancies**

1. “Score” field over $T$: 
   \[ Z(x_j) = \frac{u(x_j) - \bar{u}}{\sigma_u} \]

2. Identify points where $Z(x) > c$
   They contribute the most to large $T_{\text{obs}}$
   → high-discrepancy (anomalous) regions

3. Apply a clustering algorithm to group them

\[ T_{\text{obs}} = D \bar{u} + \text{const} \]
1. Statistical test of dataset compatibility

- Nearest-Neighbors Two-Sample Test
- Identify Discrepancies
- Include Uncertainties

2. Applications to High-Energy Physics
Sample Uncertainties

How to include sample uncertainties?

1. Model feature uncertainties
   \[ F_B(x), F_T(x) \]
   [e.g. zero-mean gaussians]

2. New samples by adding random noise sampled from \( F_{B,T} \):
   \[
   \mathcal{T}_u = \{ x_i + \Delta x_i \}_{i=1}^{N_T} \\
   \mathcal{B}_u = \{ x'_i + \Delta x'_i \}_{i=1}^{N_B}
   \]

3. Compute TS on new samples
   \[ \text{TS}_u \equiv \text{TS}(\mathcal{B}_u, \mathcal{T}_u) = \text{TS}_{\text{obs}} + U \]

4. Repeat many times to reconstruct \( f(U) \)
> Sample Uncertainties

How to include sample uncertainties?

- $f(TS_u)$ is a convolution: $f(TS_u|H_0) = f(TS|H_0) \ast f(U)$
  - $f(TS_u)$ more spread than $f(TS)$

- $p$-value computed from $f(TS_u)$

- weaker significance, power degradation
2D Gaussian with Uncertainties

\( B, T \) gaussian samples:

\[
p_B = \mathcal{N}(\mu_B, \Sigma_B) \quad p_T = \mathcal{N}(\mu_T, \Sigma_T)
\]

\[
\mu_B = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \quad \mu_T = \begin{pmatrix} 1.15 \\ 1.15 \end{pmatrix}
\]

\[
\Sigma_B = \Sigma_T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\]

uncorrelated errors (diagonal covariance) with fixed relative uncertainty

\[
\sigma_i = \epsilon x_i
\]

for each feature component \( i \)}
> **NN2ST: Summary**

- ✔ general, model-independent
- ✔ fast, no optimization
  - \[ N_{B,T}=20k, \; K=5, \; N_{perm}=1k, \; D=2: \; t \sim 2 \; \text{mins} \]
  - \[ N_{B,T}=20k, \; K=5, \; N_{perm}=1k, \; D=8: \; t \sim 50 \; \text{mins} \]
- ✔ sensitive to **unspecified** signals
- ✔ useful when no variable can separate sig/bkg
- ✔ helps finding signal regions, optimal cuts, …
- ✔ flexible to incorporate uncertainties

- ✗ need to run for each sample pair
- ✗ permutation test is bottleneck
1. Statistical test of dataset compatibility
   - Nearest-Neighbors Two-Sample Test
   - Identify Discrepancies
   - Include Uncertainties

2. Applications to High-Energy Physics
Our Method

- **Bkg Simulation** (Benchmark)
- **Data** (Trial)

**NN2ST**

- **Reject null hypothesis?**
  - no → **No signal in data**
  - yes → **hint of new physics!**
    - select regions to explore
DM search @ LHC

- “proof-of-principle” study
- bkg: $Z \rightarrow \nu \bar{\nu} + (1, 2) j$ ($\sigma_{\text{bkg}}=202.6 \text{ pb}$) sub-leading bkgs not included
- no full detector effects (generic Delphes profile)

Benchmark: BKG$_1$
Trials: BKG$_2$ + SIG
$K = 5$
$N_{\text{perm}} = 3000$

$8 \text{ features:}$
- n. of jets
- $p_T, \eta$ of 2 leading jets
- $E_T^{\text{miss}}, H_T$
- $\Delta \phi_{E_T^{\text{miss}}, j_1}$
> DM search @ LHC

**B**: $\text{BKG}_1$ (20k events)

**T1**: $\text{BKG}_2$ (20k events) + $\text{SIG}_1$ (2010 events)

**T2**: $\text{BKG}_2$ (20k events) + $\text{SIG}_2$ (375 events)

**T3**: $\text{BKG}_2$ (20k events) + $\text{SIG}_3$ (59 events)

$$N_{\text{sig}} = N_B \times \frac{\sigma_{\text{signal}}}{\sigma_{\text{bkg}}}$$

<table>
<thead>
<tr>
<th>Sample</th>
<th>$M_{Z'}$</th>
<th>$\sigma_{\text{signal}}$</th>
<th>$Z_{\text{no uncert.}}$</th>
<th>$Z_{10% \text{ rel uncert.}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.2 TeV</td>
<td>20.4 pb</td>
<td>40 $\sigma$</td>
<td>26 $\sigma$</td>
</tr>
<tr>
<td>T2</td>
<td>2 TeV</td>
<td>3.8 pb</td>
<td>13 $\sigma$</td>
<td>12 $\sigma$</td>
</tr>
<tr>
<td>T3</td>
<td>3 TeV</td>
<td>0.6 pb</td>
<td>2.7 $\sigma$</td>
<td>2.5 $\sigma$</td>
</tr>
</tbody>
</table>

- systematics: expect further degradation of results
- the method has value, it is worth exploring

still not real-world
$N_B = 20000$

$N_T = N_B + N_{\text{sig}}$

more data, more power

stronger signal
easier to discover

$N_{\text{sig}} = N_B \times \frac{\sigma_{\text{signal}}}{\sigma_{\text{bkg}}}$

$Z = 5\sigma$

$Z = 5\sigma$
Directions for future work:

- adaptive choice of $K$
- identifying discrepant regions in realistic situations (with $Z$-score method)
- validation tool for bkg: compatibility between MC sims. and data in control regions
- scalability
- … your suggestions?
1. New Statistical Test for BSM Physics
   - assess degree of compatibility between 2 samples
   - rooted on nearest neighbors, solid math foundations

2. NN2ST as a discovery tool
   - powerful and model-independent
   - lots of applications

3. NN2ST to guide searches
   - identify regions of discrepancies
BACK UP
how to choose $K$? **Model Selection**!

True: $r(x) = \frac{p_T(x)}{p_B(x)}$

Estimated: $\hat{r}(x) = \frac{\hat{p}_T(x)}{\hat{p}_B(x)}$

Define the mean-square error:

$$L(r, \hat{r}) = \frac{1}{2} \int [\hat{r}(x') - r(x')]^2 p_B(x') dx'$$

$$= \frac{1}{2} \int \hat{r}(x')^2 p_B(x') dx' - \int \hat{r}(x)p_T(x) dx + \frac{1}{2} \int r(x')^2 p_B(x') dx'$$

Estimate loss:

$$\hat{L}(r, \hat{r}) = \frac{1}{2N_B} \sum_{x' \in B} \hat{r}(x')^2 - \frac{1}{N_T} \sum_{x \in T} \hat{r}(x)$$

Select optimal $K$ minimizing the loss.

Alternatively: Point-Adaptive k-NN (PAk) [1802.10549]