Challenges and Opportunities for Anomaly Detection

David Shih
Rutgers University

Snowmass Community Planning Meeting
Collider Data Analysis Strategies
October 6, 2020
By now, hundreds (thousands?) of searches for new physics at the LHC.

No sign of new physics yet...
By now, hundreds (thousands?) of searches for new physics at the LHC.

No sign of new physics yet...

99.9999% of the searches at the LHC are model-specific.

Are we covering every possibility?
History of Model Independent Searches at Colliders

- **D0**
  - “Sleuth”
  - PRD 64:012004 (2001)
  - PRL 86:3712 (2001)

- **H1 (Hera)**
  - “General Search”
  - 0705.3721

- **CDF**
  - “Sleuth/Vista”
  - 0712.1311 PRD 78:012002 (2008)
  - 0712.2534 (submitted to PRL, NEVER PUBLISHED)
  - 0809.3781 PRD 79:011101 (2009)

- **CMS**
  - “MUSIC”
  - CMS-PAS-EXO-14-016

- **ATLAS**
  - “Model independent general search”
  - 1807.07447 EPJC 79:120 (2019)
History of Model Independent Searches at Colliders

- **D0**
  - "Sleuth"
  - PRD 64:012004 (2001)
  - PRL 86:3712 (2001)

- **CDF**

- **CMS**
  - "MUSIC"
  - CMS-PAS-EXO-14-016

- **ATLAS**
  - "Model independent general search"
  - 1807.07447 EPJC 79:120 (2019)

**Why are there so few model independent searches at collider experiments?**

**Why are they all following the same approach — comparing data to simulation?**
The ANODE approach and provides a brief introduction to normalizing flows. The reminder of the paper illustrates ANODE through an example based on a dijet search using jet substructure. Details of the simulated samples are provided in Sec. ?? and the results for the signal sensitivity and background specificity are presented in Sec. ?? and ??, respectively.

A study of correlations between the discriminating features and the resonant feature is in Sec. ??. The paper ends with conclusions and outlook in Sec. ??.

An Overview of Model (In)dependent Searches

A viable search for new physics generally must have two essential components: it must be sensitive to new phenomena and it must also be able to estimate the background under the null hypothesis (Standard Model only). The categorization of a search's degree of model (in)dependence requires consideration of both of these components. Figure ?? illustrates how to characterize model independence for both BSM sensitivity and SM background specificity.

We will now consider each in turn.

![Graphical representation of searches for new particles in terms of the background and signal model dependence for achieving signal sensitivity (a) and background specificity (b). The Model Unspecific Search for New Physics (MUSiC) and General Search strategies are from CMS and ATLAS, respectively. LDA stands for Latent Dirichlet Allocation, ANODE stands for ANomaly detection with Density estimation. CWoLa stands for Classification Without Labels, and SALAD stands for Simulation Assisted Likelihood-free Anomaly Detection. Direct density estimation is a form of side-banding where the multidimensional feature space density is learned conditional on the resonant feature (see Sec. ??).](from Nachman & DS 2001.04990)
A study of correlations between the discriminating features and the resonant feature is in Sec. ??

There have been many breakthroughs in unsupervised ML since 2000.
details of the ANODE approach and provides a brief introduction to normalizing flows. The remainder of the paper illustrates ANODE through an example based on a dijet search using jet substructure. Details of the simulated samples are provided in Sec. ?? and the results for the signal sensitivity and background specificity are presented in Sec. ?? and ??, respectively.

A study of correlations between the discriminating features and the resonant feature is in Sec. ??.

The paper ends with conclusions and outlook in Sec. ??.

2 An Overview of Model (In)dependent Searches

A viable search for new physics generally must have two essential components: it must be sensitive to new phenomena and it must also be able to estimate the background under the null hypothesis (Standard Model only). The categorization of a search's degree of model (in)dependence requires consideration of both of these components. Figure ?? illustrates how to characterize model independence for both BSM sensitivity and SM background specificity.

We will now consider each in turn.

<table>
<thead>
<tr>
<th>Signal model independence</th>
<th>Background (SM) model independence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUSiC (CMS), General Search (ATLAS)</td>
<td>autoencoders, LDA, ANODE, CWoLa, SALAD</td>
</tr>
<tr>
<td>Some searches (train signal versus data)</td>
<td>Most searches (train with simulations)</td>
</tr>
<tr>
<td>Pure MC prediction</td>
<td>MUSiC (CMS), General Search (ATLAS)</td>
</tr>
</tbody>
</table>

Figure 1. A graphical representation of searches for new particles in terms of the background and signal model dependence for achieving signal sensitivity (a) and background specificity (b). The Model Unspecific Search for New Physics (MUSiC) \[ ?? \] and General Search \[ ?? \] strategies are from CMS and ATLAS, respectively. LDA stands for Latent Dirichlet Allocation \[ ?? \], ANOmaly detection with Density Estimation (ANODE) is the method presented in this paper, CWoLa stands for Classification Without Labels \[ ?? \] and SALAD stands for Simulation Assisted Likelihood-free Anomaly Detection \[ ?? \].

Direct density estimation is a form of side-banding where the multidimensional feature space density is learned conditional on the resonant feature (see Sec. ??).

There have been many breakthroughs in unsupervised ML since 2000.

These have inspired many new ideas for model-independent searches at the LHC recently!

from Nachman & DS 2001.04990
details of the ANODE approach and provides a brief introduction to normalizing flows. The reminder of the paper illustrates ANODE through an example based on a dijet search using jet substructure. Details of the simulated samples are provided in Sec. ?? and the results for the signal sensitivity and background specificity are presented in Sec. ??, respectively.

A study of correlations between the discriminating features and the resonant feature is in Sec. ??.

The paper ends with conclusions and outlook in Sec. ??.

2 An Overview of Model (In)dependent Searches

A viable search for new physics generally must have two essential components: it must be sensitive to new phenomena and it must also be able to estimate the background under the null hypothesis (Standard Model only). The categorization of a search's degree of model (in)dependence requires consideration of both of these components. Figure ?? illustrates how to characterize model independence for both BSM sensitivity and SM background specificity.

We will now consider each in turn.

![Diagram showing the categorization of model independence for signal sensitivity and background specificity.]

From Nachman & DS 2001.04990

There have been many breakthroughs in unsupervised ML since 2000.

These have inspired many new ideas for model-independent searches at the LHC recently!

Some challenges:

- Robust background estimation
- Quantifying performance
- Recasting?
This is just the beginning!
Backup
A Benchmark Example

LHC Olympics 2020 R&D Dataset
https://doi.org/10.5281/zenodo.2629072

No explicit search at the LHC for this scenario.

Could be hiding in the dijet resonance search at >5sigma significance!!