Limiting false positives in model-independent anomaly detection

Kees Benkendorfer
Reed College

In collaboration with Luc Le Pottier (Michigan) and Benjamin Nachman (Berkeley Lab)

Based in part on arXiv:2009.02205

22 September 2020
PacNW Theory Workshop

1. Motivation
   - Anomaly detection
   - The problem of false positives
2. Limiting false positives
   - Solution: SALAD
   - Solution: SA-CWoLa
   - Results
3. Conclusion

Anomaly detection basics

- The Standard Model (SM) is known to be incomplete
  - No particles Beyond the Standard Model (BSM) have yet been found
  - **There remains a vast space of models with no dedicated search**
- Machine learning opens doors for model-independent anomaly detection
- Many techniques have recently been developed, e.g.
  - CWoLa [1, 2], SALAD [3], ANODE [4]

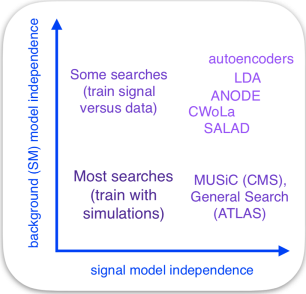


Figure: Snapshot of the current landscape of anomaly detection methods. From [4]

The problem

- Sensitivity of new ML methods rivals or exceeds traditional anomaly searches
- Beginning to see application in data
  - ATLAS analysis using CWoLa was first ML-based anomaly hunt [5]

Problem

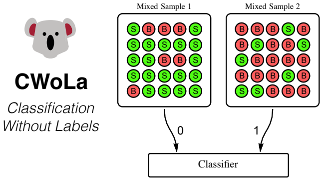
False positives pose a significant challenge for certain ML anomaly detection methods as sensitivity increases

We need new techniques to mitigate this behavior

Example: CWoLa

- Train supervised classifier to distinguish between two groups of mixed (unlabeled) background/signal
- **Classifier effectively learns to distinguish between signal and background**
- For bump hunt, groups determined by course binning in invariant dijet mass
- Use 'signal'-tagging to emphasize significance of signal

,



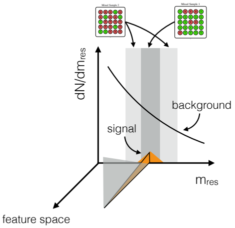
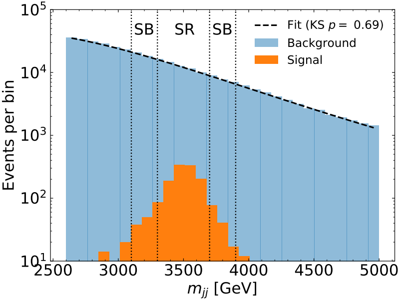


Figure: CWoLa schematic: see e.g. [6]

Example: CWoLa

- Using the 2020 LHC Olympics dataset for prototyping
- Simulation of an LHC-like detector
- Signal is $W' \to XY$ with $m_{W'} = 3.5\,\text{TeV}$, $m_X = 500\,\text{GeV}$, and $m_Y = 100\,\text{GeV}$
  - Event with two jets (sprays of hadronic particles)
- Jet masses artificially correlated to M_{JJ} by taking $m_j \mapsto m_j + 0.1 M_{JJ}$
- Classifier trained between 'signal region' (SR) and 'sideband' (SB)



Example: CWoLa

- CWoLa classifier trained on several jet features
  - Jet mass, N-subjettiness ratio τ_{21}
- Classifier able to infer M_{JJ} from correlations
- SR and SB come from different regions in M_{JJ} $\implies$ classifier tags full SR as 'signal'-like
  - Note: ATLAS result in [5] trained only on jet masses and performed explicit decorrelation to avoid this problem
- Result: severe distribution sculpting

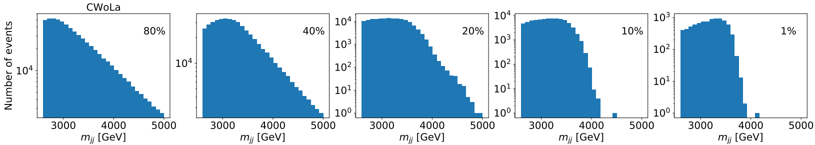


Figure: Sculpting of the M_{JJ} distribution by a CWoLa classifier

Strategy

- Techniques exist to modify data (e.g. explicit decorrelation) to mitigate the problem
    - Can be finicky in practice
    - Might cause sensitivity reduction by removing information

Goal

It would be nice to have false positive mitigation built into the anomaly detection method itself

Two potential solutions:

1. Eliminate M_{JJ} differences by comparing SR to itself
    - SALAD
2. Penalize the classifier for learning M_{JJ}
    - SA-CWoLa

SALAD

- Idea: eliminate SR/SB differences by only looking at SR
- Introduce a simulated dataset (Herwig++)
- Reweight simulation to look like data in sidebands using NN
- Extend reweighting to signal region in simulation
- Train classifier to distinguish simulation signal region and data signal region

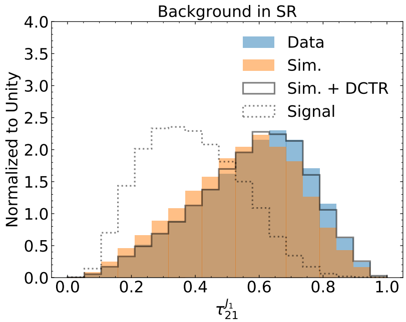


Figure: τ_{21} of the leading jet in data, simulation, signal, and reweighted simulation ('Sim. + DCTR')

SA-CWoLa

- Modification to CWoLa
- Idea: penalize classifier for distinguishing SR and SB in a simulated dataset
  - Use normal 'binary cross-entropy' loss function in data and negative binary cross-entropy in simulation
- Loss function minimized by picking out signal, and only signal

$$\mathcal{L}_{\text{SA-CWola}}[f] = -\sum_{i \in \text{SR,data}} \log(f(x_i)) - \sum_{i \in \text{SB,data}} \log(1 - f(x_i)) \\ + \lambda \left(\sum_{i \in \text{SR,sim.}} \log(f(x_i)) + \sum_{i \in \text{SB,sim.}} \log(1 - f(x_i)) \right)$$

SA-CWoLa limits M_{JJ} sculpting

- Classifier trained on jet masses and N-subjettiness ratio τ_{21} of leading two jets
- SA-CWola exhibits much less bump sculpting than CWoLa

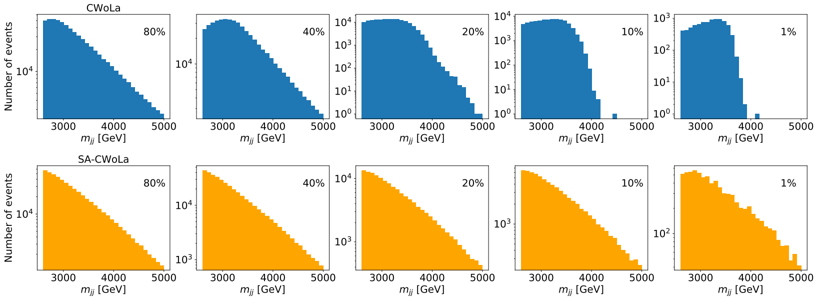
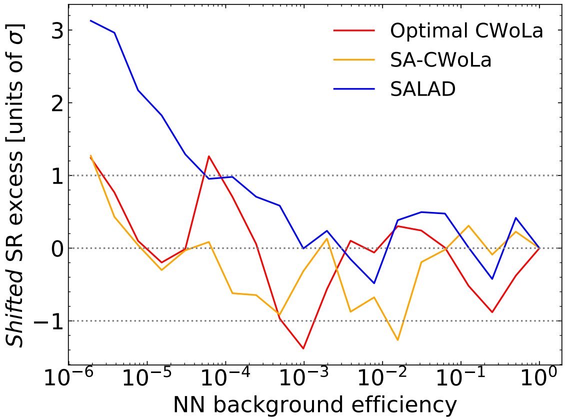


Figure: Sculpting of the M_{JJ} distribution by CWoLa classifier (top) and SA-CWoLa (bottom)

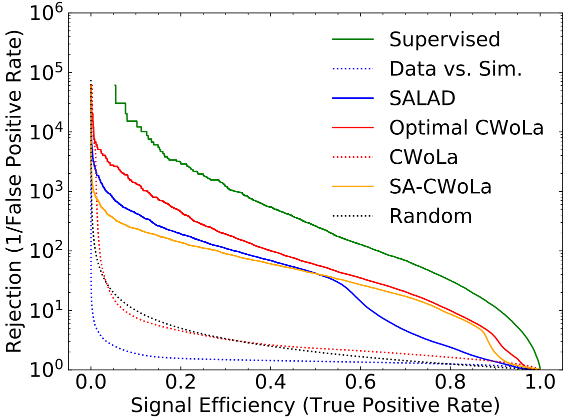
False positive results

- No signal injected, shifted to correct for natural SR deficit
- ‘Optimal CWoLa’ is CWoLa trained on SR vs SR + signal (instead of SB vs SR + signal)



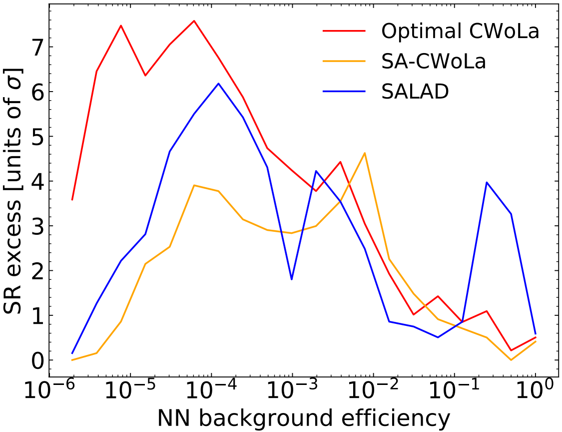
Signal sensitivity results: pure performance

- 2σ signal injected



Signal sensitivity results: with background estimation

- 2σ signal injected



Conclusion

- Limiting false positives is increasingly important as we start applying ML anomaly detection techniques in data
- Currently-used techniques require method-external techniques to correct for this
  - Can be unwieldy
  - Might hurt signal sensitivity
- SALAD and SA-CWoLa are promising techniques that are intrinsically robust
  - We may not need to sacrifice signal sensitivity to prevent false positives

Acknowledgments

- This work was supported in part by NSF PHY REU Grant 194992
- This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231

References

[1] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Anomaly Detection for Resonant New Physics with Machine Learning. *Phys. Rev. Lett.*, 121(24):241803, December 2018.

[2] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Extending the Bump Hunt with Machine Learning. *Phys. Rev. D*, 99(1):014038, January 2019.

[3] Anders Andreassen, Benjamin Nachman, and David Shih. Simulation Assisted Likelihood-free Anomaly Detection. *arXiv:2001.05001 [hep-ex, physics:hep-ph, physics:physics, stat]*, January 2020.

[4] Benjamin Nachman and David Shih. Anomaly Detection with Density Estimation. *Phys. Rev. D*, 101(7):075042, April 2020.

[5] ATLAS Collaboration. Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector. *arXiv:2005.02983 [hep-ex]*, May 2020.

[6] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: Learning from mixed samples in high energy physics. *J. High Energ. Phys.*, 2017(10):174, October 2017.