Making classifiers less biased one Legendre mode at a time

MoDe as a Method to Enhance Searches for Resonances

Ouail Kitouni Benjamin Nachman Constantin Weisser Michael Williams

Mitigating Classifier Bias

- It is often desired to **control the dependence of a classification algorithm on some protected feature** of the data samples.
- In particle colliders, searches for resonant new physics require classifiers that do not sculpt a peak in an otherwise smooth background spectrum.
- Standard techniques aim to produce classifiers that are **independent** from the protected feature. This is **sufficient** to avoid localized structures, but is **not necessary.**

When a classifier turns cats into dogs (or quarks into W bosons)

- Jets produced from boosted W boson decays are interesting in many extensions of the Standard Model.
- Various features of the substructure of these jets can be used to distinguish the boosted bosons (signal) from generic quark and gluon jets (background.)
- A bump hunt (search for excess signal in the mass spectrum of a sample of events) is performed on the mass (protected attribute) of the W candidate.
- The challenge with substructure classifiers is that they can **introduce artificial bumps** into the mass spectrum because substructure is correlated with mass m (see central figures.)

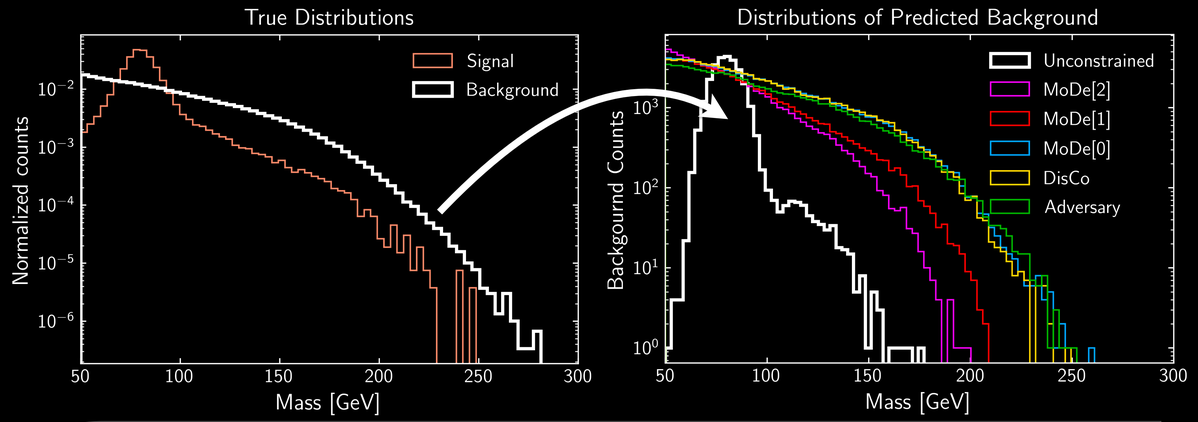
Moment Decomposition (MoDe[ℓ])

MoDe is a new, fast and simple **regularization technique** which produces **unbiased classifiers** with **better classification power** than state-of-the-art methods.

We regularize the standard classification loss by adding the following term

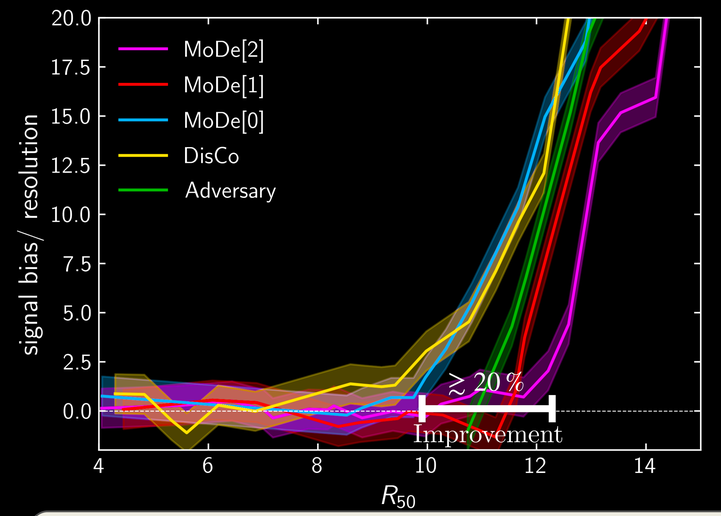
$$L^{\ell}_{\text{MoDe}} \equiv \sum_m \int |F_m(s) - \sum_{i=0}^{\ell} c_i(s) P_i(m)|^2 \mathrm{d}s,$$

where s is the classifier output. We decompose the empirical CDF of s conditioned on the protected attribute m, $F_m(s)$, into its first ℓ Legendre modes. **Higher order moments** are **penalized**, resulting in a classifier with a **simpler dependence** on m.



Left: true distribution of the mass of signal and background jets. **Right:** distribution of samples predicted to be background jets (at 50% true positive rate) by classifiers trained with various regularizations.

- **Unconstrained classifier:** only minimizes the classification loss. We use it as a baseline to show how peak sculpting can arise. This classifier is unusable in a real analysis.
- **MoDe[ℓ]:** penalizes dependence on m of orders higher than ℓ.
- **DisCo:** achieves independence by minimizing distance correlation between m and s.
- **Adversary:** relies on training an adversary to predict m from s. The solution to this minimax optimization problem should be independent from m.



- To asses classification performance we use the background rejection factor R_{50}: the inverse false positive rate at 50% true positive rate.
- To quantify peak sculpting we use the **signal bias** induced by the classifier selection. Values less than unity are consistent with no bias (since the true signal rate is zero), while values significantly larger indicate substantial bias that could result in false claims of observations.

The figure above shows that the **flexibility to go beyond decorrelation** provided by **MoDe[1,2]** results in **unbiased signal estimators** at **larger background-rejection power.** This would directly translate to **improved sensitivity** in a real-world analysis.

https://github.com/okitouni/MoDe

arXiv https://arxiv.org/abs/2010.09745

OK was supported by an MIT fellowship. BPN was supported by the Department of Energy under contract DE-AC02-05CH11231. CW and MW were supported by the National Science Foundation under contracts PHY-1912836, PHY-2019786, and OAC-1739772.