Robust signal detection with classifiers decorrelated via optimal transport

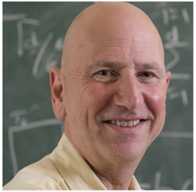
Purvasha Chakravarti

Department of Statistical Science, University College London

Lucas Kania

Olaf Behnke

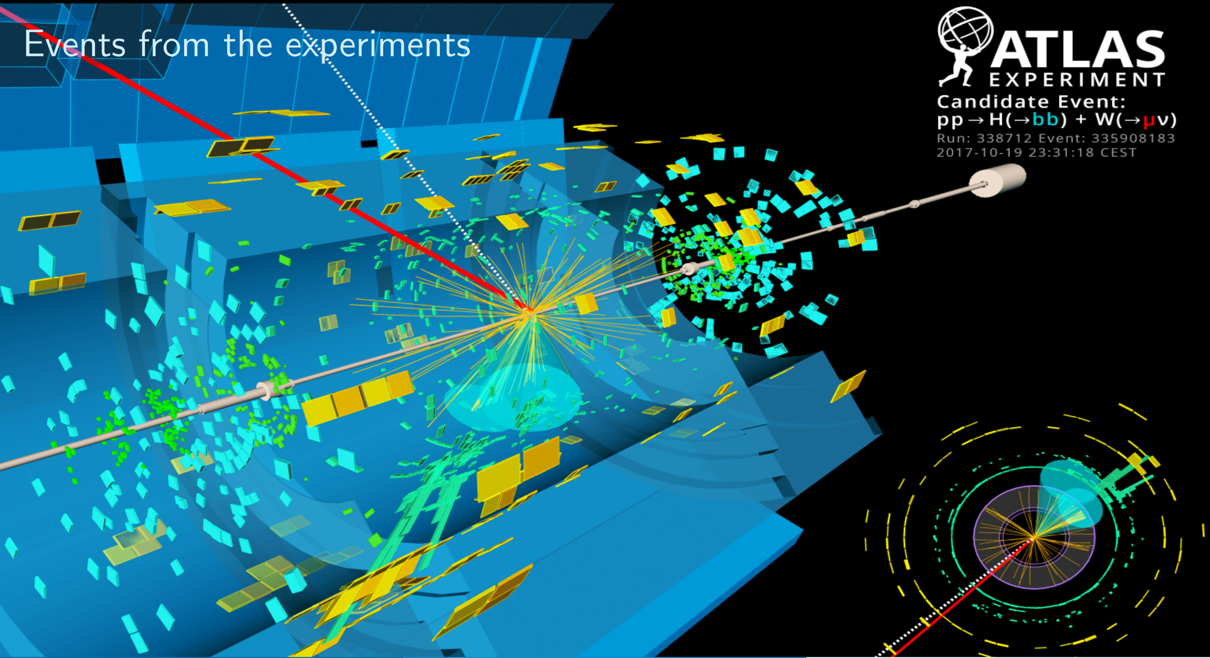
Mikael Kuusela



Larry Wasserman

PHYSTAT - Statistics meets ML, Imperial College London
September 11, 2024

Events from the experiments



Experimental data

Experimental data are generated from one of the two processes:
Background - refers to the known physics (SM).
Signal - represents an interesting event with a known/unknown possible particle.

Experimental data

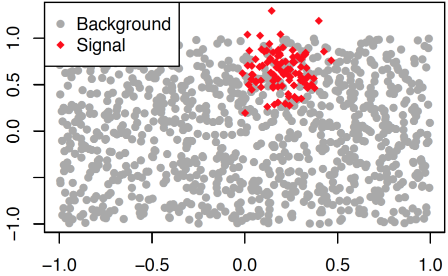
Experimental data are generated from one of the two processes:
Background - refers to the known physics (SM).
Signal - represents an interesting event with a known/unknown possible particle.

$$q = (1-\lambda)p_b + \lambda p_s, \quad \text{No signal: } \lambda = 0$$

Experimental data

Experimental data are generated from one of the two processes:
Background - refers to the known physics (SM).
Signal - represents an interesting event with a known/unknown possible particle.

$$q = (1-\lambda)p_b + \lambda p_s, \quad \text{No signal: } \lambda = 0$$



Two-dimensional toy example.

Model-dependent methods

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_{m_b} \sim p_b$$
$$\text{Signal:} \quad Y_1, \ldots, Y_{m_s} \sim p_s$$

Model-dependent methods

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_{m_b} \sim p_b$$
$$\text{Signal:} \quad Y_1, \ldots, Y_{m_s} \sim p_s$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_n \sim q = (1-\lambda)p_b + \lambda p_s$$

Model-dependent methods

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_{m_b} \sim p_b$$
$$\text{Signal:} \quad Y_1, \ldots, Y_{m_s} \sim p_s$$

- Background + possible signal (experimental) sample - unlabelled observations

$$\text{Experimental:} \quad W_1, \ldots, W_n \sim q = (1-\lambda)p_b + \lambda p_s$$

Testing for signal can be formulated as:

$$H_0 : \lambda = 0 \quad \text{versus} \quad H_1 : \lambda > 0.$$

Train a classifier (h) to separate signal from background.

Problem

Methods assume that the background samples $X_1, \ldots, X_{m_b}$ come from the "true" background distribution p_b.

But X's are MC simulations which are likely to be systematically misspecified.

Problem

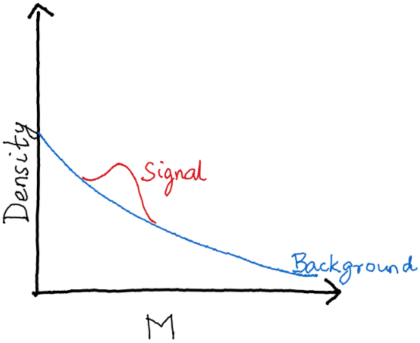
Methods assume that the background samples $X_1, \ldots, X_{m_b}$ come from the "true" background distribution p_b.

But X's are MC simulations which are likely to be systematically misspecified.

Important question: Are the "signals" found true signals or differences between the true background and a misspecified background?

Towards background-agnostic. Signal is localized in some resonant features

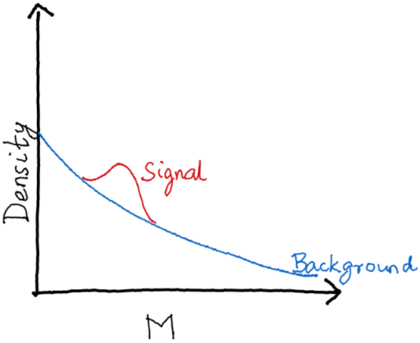
$$q = (1-\lambda)p_b + \lambda p_s, \quad \text{No signal: } \lambda = 0 \text{ or equivalently } q = p_b$$



Localization in resonant feature M.

Towards background-agnostic. Signal is localized in some resonant features

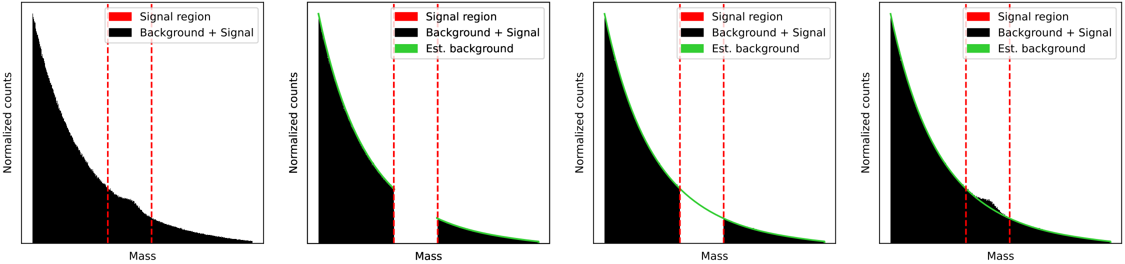
$$q = (1-\lambda)p_b + \lambda p_s, \quad \text{No signal: } \lambda = 0 \text{ or equivalently } q = p_b$$



Localization in resonant feature M.

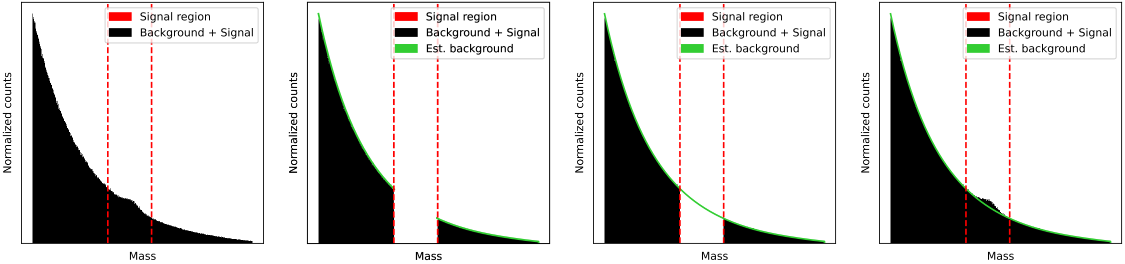
Signal detection is performed on resonant feature of only the experimental data.

Bump hunting



See details: [Chakravarti et al. (2409.06399)]

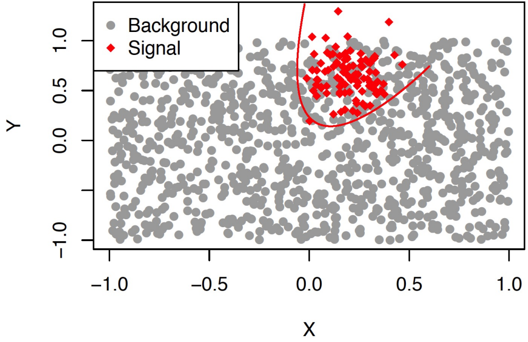
Bump hunting



See details: [Chakravarti et al. (2409.06399)]

Problem: λ is usually very small.

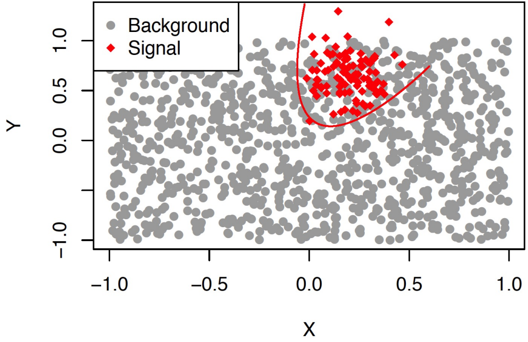
Signal enrichment using auxiliary variables



Two-dimensional toy example.

- Access to MC simulations from assumed Background and Signal models.

Signal enrichment using auxiliary variables



Two-dimensional toy example.

- Access to MC simulations from assumed Background and Signal models.
- Signal enrichment is performed using a classifier trained on auxiliary variables of simulated data before signal detection.

Data

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\begin{aligned} \text{Background:} \quad & X_1, \ldots, X_{m_b} \sim p_b \\ \text{Signal:} \quad & Y_1, \ldots, Y_{m_s} \sim p_s \end{aligned}$$

Used to train the supervised classifier h.

Data

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\begin{aligned} \text{Background:} \quad & X_1, \ldots, X_{m_b} \sim p_b \\ \text{Signal:} \quad & Y_1, \ldots, Y_{m_s} \sim p_s \end{aligned}$$

  Used to train the supervised classifier h.

- Background + possible signal (real, experimental) sample - unlabelled observations

$$\begin{aligned} \text{Auxiliary Variables:} \quad & W_1, \ldots, W_n \sim q = (1-\lambda)p_b + \lambda p_s \\ \text{Resonant/Protected Variable:} \quad & M_1, \ldots, M_n \end{aligned}$$

Data

Two sources of data are at hand:

- Background + signal (Monte Carlo) sample - labelled observations

$$\text{Background:} \quad X_1, \ldots, X_{m_b} \sim p_b$$
$$\text{Signal:} \quad Y_1, \ldots, Y_{m_s} \sim p_s$$

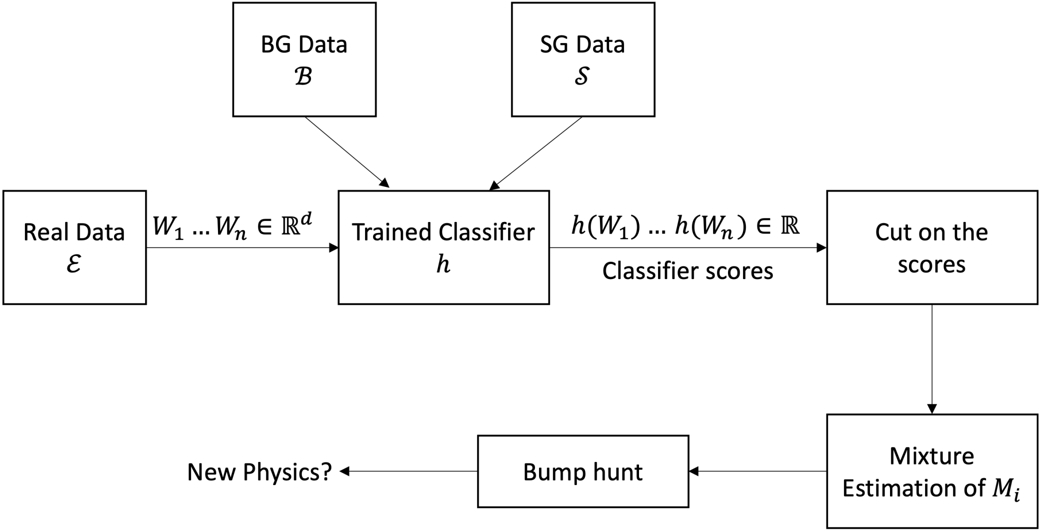
  Used to train the supervised classifier h.

- Background + possible signal (real, experimental) sample - unlabelled observations

$$\text{Auxiliary Variables:} \quad W_1, \ldots, W_n \sim q = (1-\lambda)p_b + \lambda p_s$$
$$\text{Resonant/Protected Variable:} \quad M_1, \ldots, M_n$$

  Use h to perform signal enrichment and $M_i's$ to perform signal detection using bump hunting.

Signal detection process



Problem with BG estimation: sculpting

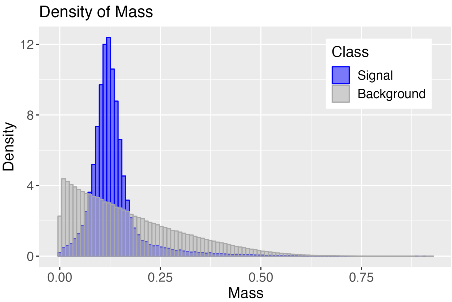
When we cut on the classifier scores the distribution of $M_i's$ changes!

Problem with BG estimation: sculpting

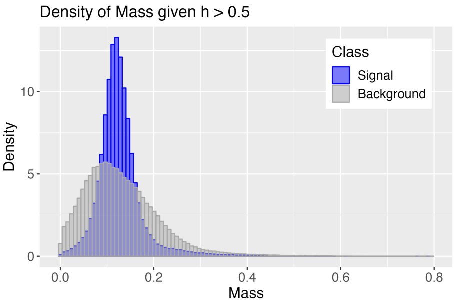
When we cut on the classifier scores the distribution of $M_i's$ changes!

Example: Protected variable: Mass, Cut: Classifier output $h > 0.5$.

Distribution of Mass



Distribution of Mass after Cut

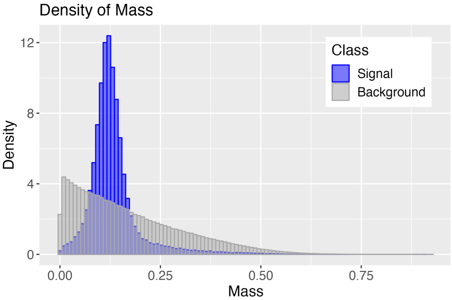


Problem with BG estimation: sculpting

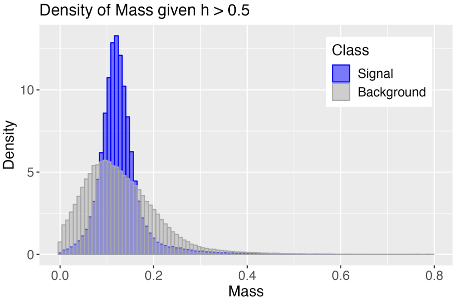
When we cut on the classifier scores the distribution of $M_i's$ changes!

Example: Protected variable: Mass, Cut: Classifier output $h > 0.5$.

Distribution of Mass



Distribution of Mass after Cut



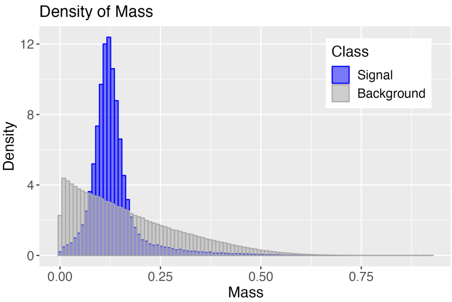
Idea: Can the protected variable have the same background distribution after cuts as before cuts? Yes, if $h(X)$ is independent of M.

What is decorrelation?

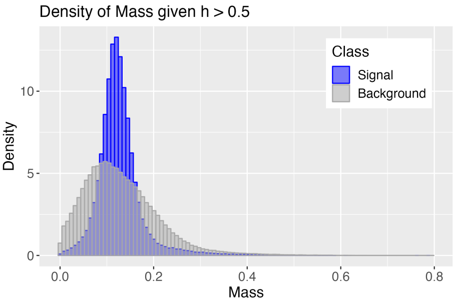
To avoid sculpting need $h(X)$ decorrelated (independent) of M!

Example: Protected variable: Mass, Cut: Classifier output $h > 0.5$.

Distribution of Mass



Distribution of Mass after Cut

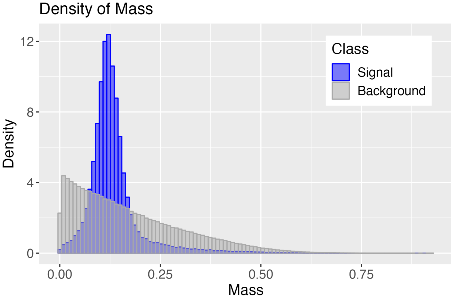


What is decorrelation?

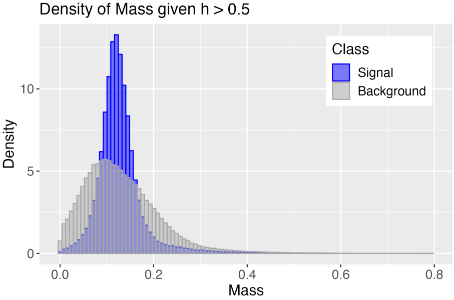
To avoid sculpting need $h(X)$ decorrelated (independent) of M!

Example: Protected variable: Mass, Cut: Classifier output $h > 0.5$.

Distribution of Mass



Distribution of Mass after Cut



Discussion on existing decorrelation methods

- Make classifier inputs decorrelated of the protected variable.
  - Designing Decorrelated Taggers (DDT) [Dolen et al.(1603.00027)]
  - Convolved SubStructure (CSS) [Moult et al. (1710.06859)]

- Enforce decorrelation of classifier during training using regularization.
  - DisCo Fever [Kasieczka, Shih (2001.05310)]
  - MoDe [Kitouni et al. (2010.09745)]
  - Adversarial Neural Networks (ANN) [Louppe et al. (1611.01046)] [Shimmin et al. (1703.03507)]

- Find a transformation of pre-trained classifier to be decorrelated of the protected variable.
  - CDOT (our method) [Chakravarti et al. (2409.06399)]
  - CNOTS [Algren et al. (2307.05187)]
  - Conditional normalizing flows [Klein et al. (2211.02486)]
  - Cuts derived from quantile regression [Moreno et al. (PhysRevD.102.012010)]

Classifier Decorrelated through Optimal Transport (CDOT)

Solution: Make cuts on transformed classifier output $T_M(h(X))$ instead, where $T_M(h(X))$ is independent of the protected variable M for background data.

- Objective: Minimize $(T_M(h(X)) - h(X))^2$ subject to $T_M(h(X))$ independent of M, given $X \sim \mathcal{B}$ and marginal of $h(X)$ and $T_M(h(X))$ are the same.

Classifier Decorrelated through Optimal Transport (CDOT)

Solution: Make cuts on transformed classifier output $T_M(h(X))$ instead, where $T_M(h(X))$ is independent of the protected variable M for background data.

- Objective: Minimize $(T_M(h(X)) - h(X))^2$ subject to $T_M(h(X))$ independent of M, given $X \sim \mathcal{B}$ and marginal of $h(X)$ and $T_M(h(X))$ are the same.

- The optimal transport map T_m from $p(h(x)|M = m, \mathcal{B})$ to the marginal $p(h(x)|\mathcal{B})$ is the solution.

Classifier Decorrelated through Optimal Transport (CDOT)

Solution: Make cuts on transformed classifier output $T_M(h(X))$ instead, where $T_M(h(X))$ is independent of the protected variable M for background data.

- Objective: Minimize $(T_M(h(X)) - h(X))^2$ subject to $T_M(h(X))$ independent of M, given $X \sim \mathcal{B}$ and marginal of $h(X)$ and $T_M(h(X))$ are the same.

- The optimal transport map T_m from $p(h(x)|M = m, \mathcal{B})$ to the marginal $p(h(x)|\mathcal{B})$ is the solution.

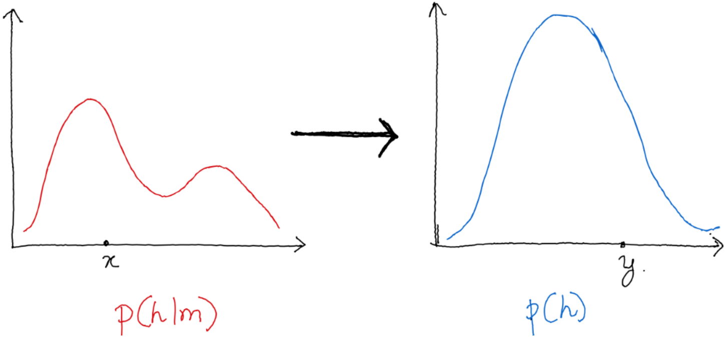
- When $h(X)$ is univariate, closed form solution:

$$T_m(h(X)) = G^{-1}(F_{h|M}(h(X)|M = m))$$

  where G is the marginal cdf of $h(X)$ and $F_{h|M}$ is the conditional distribution of $h(X)$ given $M = m$ and X is from the background distribution.

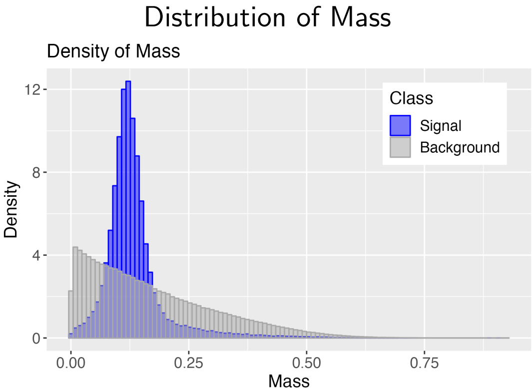
Classifier Decorrelated through Optimal Transport (CDOT)

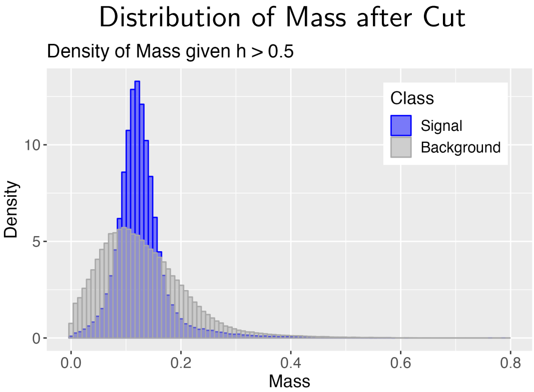
The optimal transport map T_m from $p(h(x)|M = m, \mathcal{B})$ to the marginal $p(h(x)|\mathcal{B})$ is the solution.



Sculpting problem

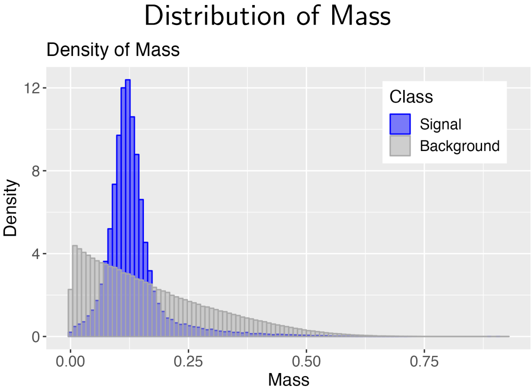
Example: Protected variable: Mass, Cut: Classifier output $h > 0.5$.

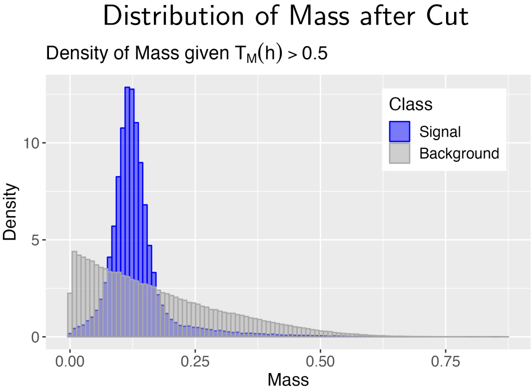




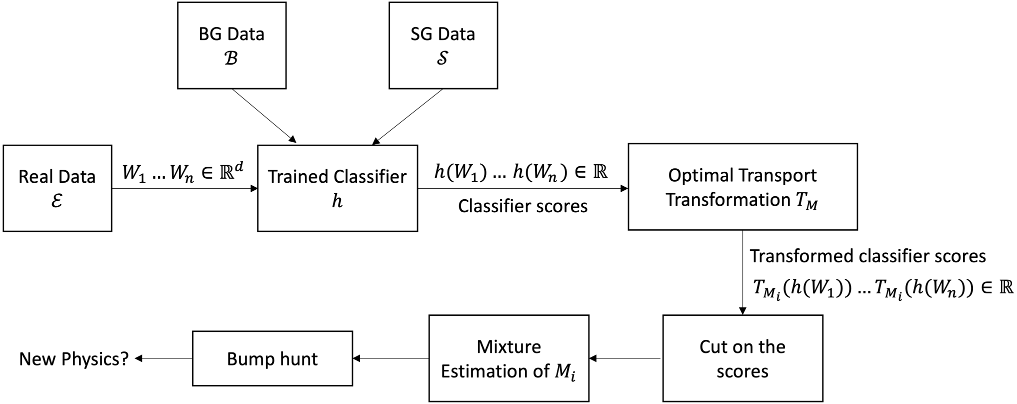
Sculpting problem solved!

Example: Protected variable: Mass, Cut: Classifier output $T_M(h) > 0.5$.



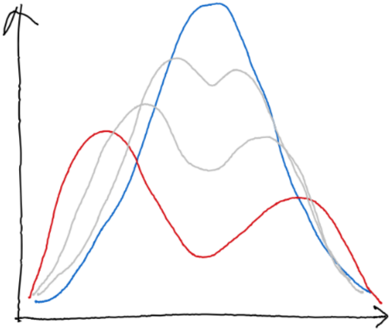


Signal detection process



Geodesic path of Optimal Transport

Solutions can span from $h(X)$ to $T(h(X))$.

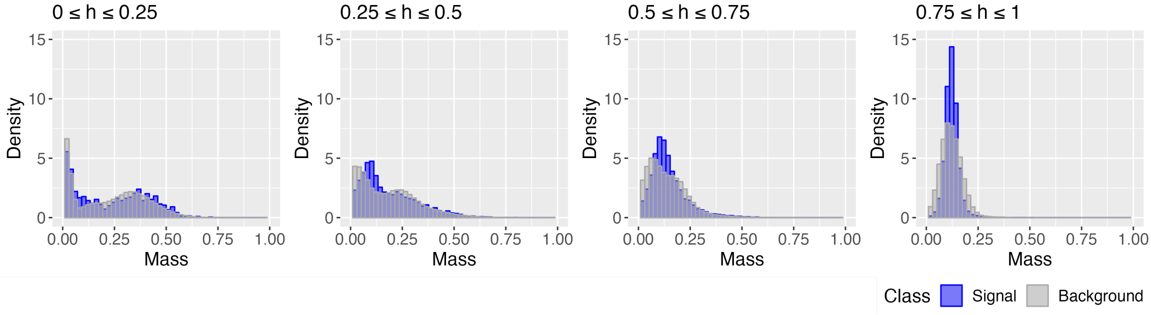


$$\beta h(X) + (1-\beta)T(h(X)), \quad \beta \in [0,1].$$

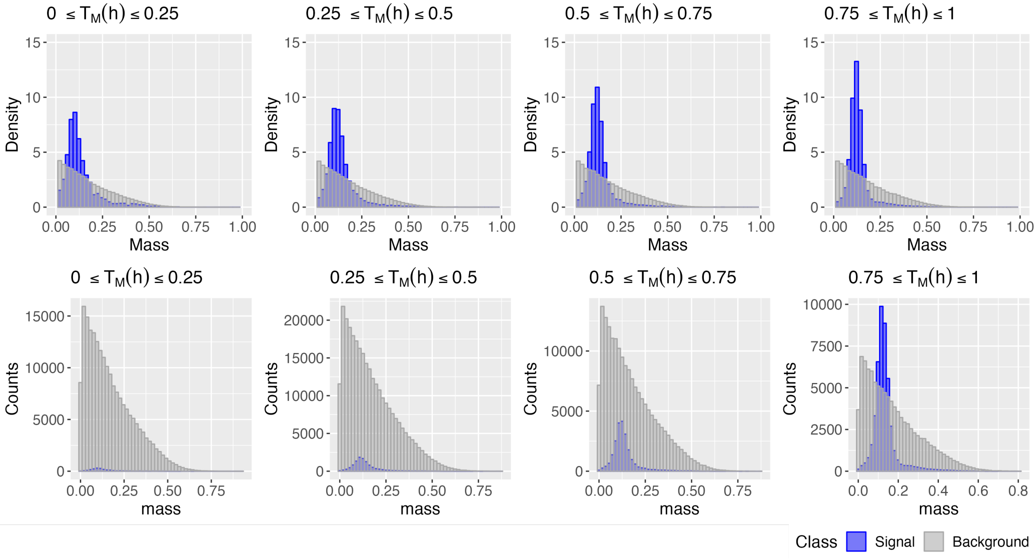
Detection of decaying high-pT W-boson events: WTagging dataset

- Boosted hadronic W tagging dataset: benchmark for studying decorrelation methods.
- Bump hunt is performed on the mass of one W candidate jet and another (possibly W candidate) jet, mJJ.
- Classification is performed on ten representative jet substructure features.
- Details can be found in DDT [Dolen et al. (JHEP 2016)], DisCo Fever [Kasieczka, Shih (2001.05310)], and MoDe [Kitouni et al. (2010.09745)] papers.

WTagging dataset: before OT transformation

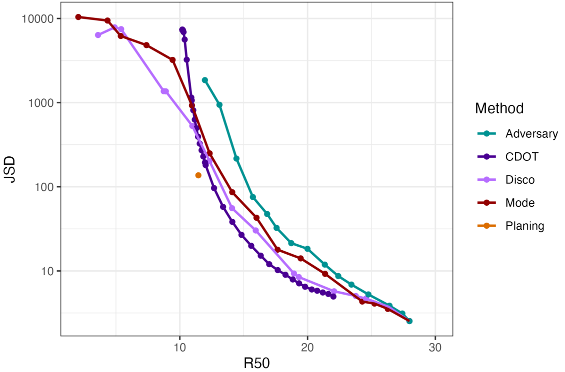


WTagging dataset: after OT transformation



WTagging dataset: comparison

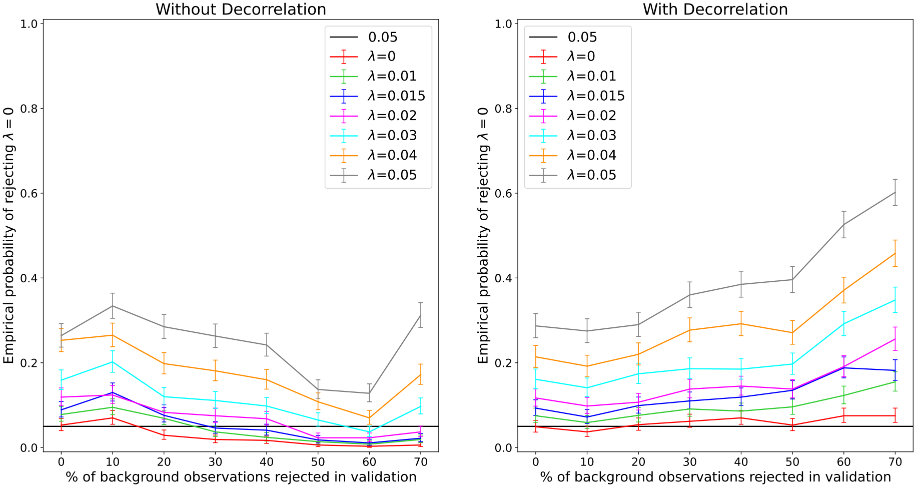
JSD: Jensen–Shannon divergence, $R50$: the background rejection power (inverse false positive rate) at 50% signal efficiency.



CDOT achieves superior signal-to-background ratio for strongly decorrelated classifiers.

Original figure without CDOT taken from the MoDe [Kitouni et al. (2010.09745)] paper.

WTagging dataset: Power



Detection of high-mass resonance events

- Data was generated using the MadGraph particle physics software.
- 4b represents events that were identified as having four b-jets.
- 3b represents events which were identified as having four jets, of which exactly three are b-jets.
- Signal sample (X $\rightarrow$ HH $\rightarrow$ 4b) produced at 400 GeV.
- We train the supervised classifier h on the pT, energy, η and ϕ variables of the four jets.
- More details: [Manole et al. (2208.02807)]

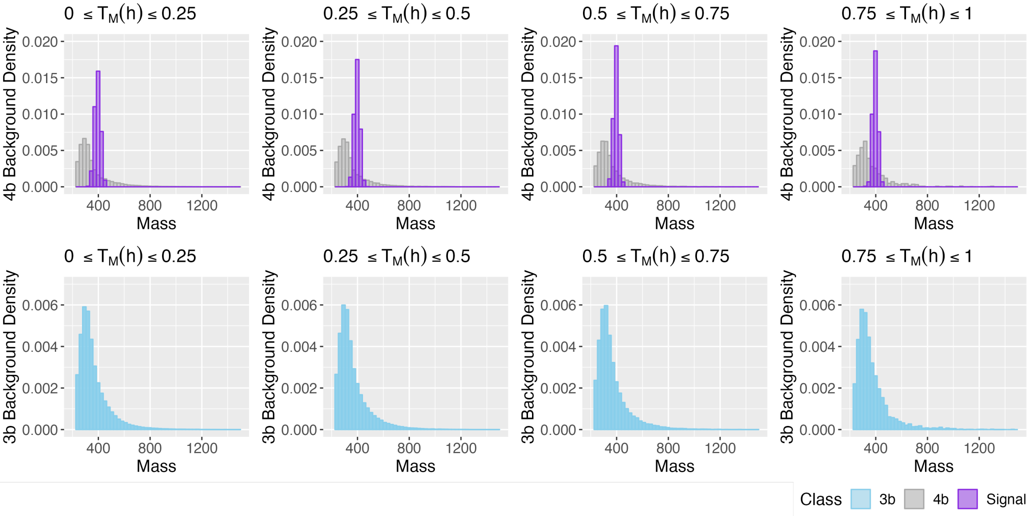
MC Background: $3b$ $(50,000)$

MC Signal: 400 signal $(44,196)$

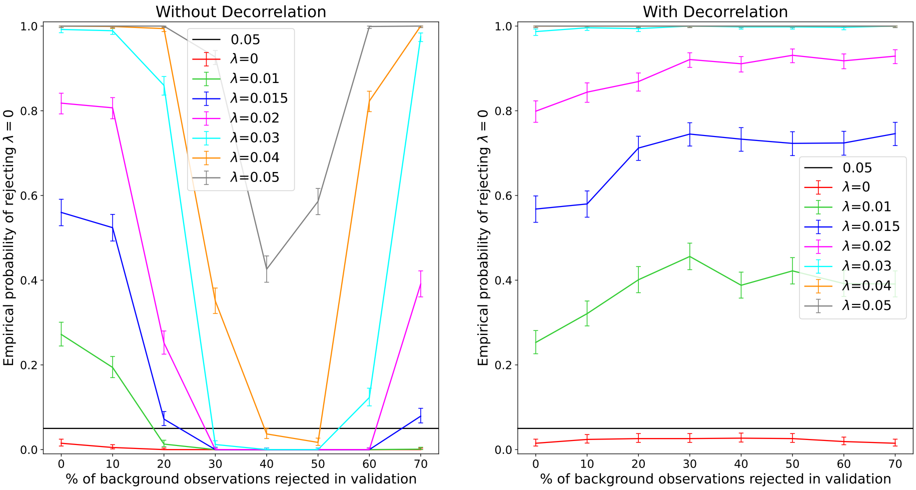
Experimental: $4b +$ 400 signal $(60,000)$

Simulated Data: robust on 4b data with signal

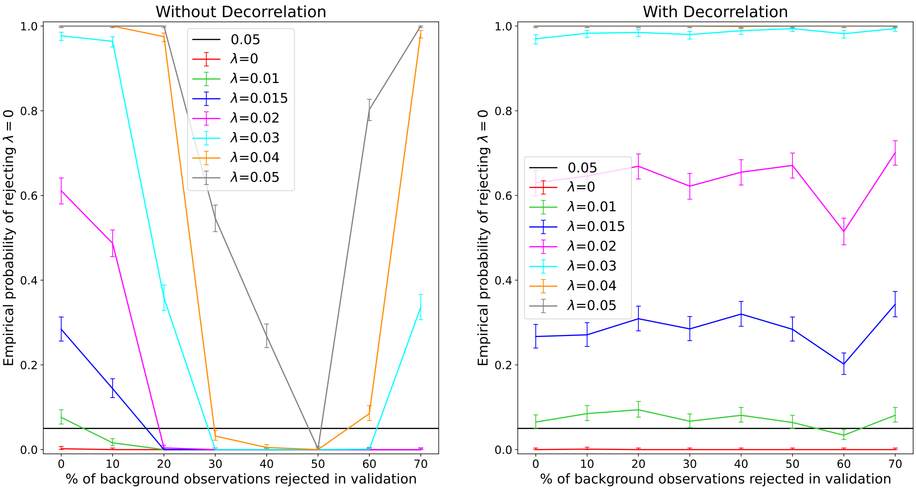
CDOT trained on the 3b data and signal shows robustness on 4b data.



3b: Power


Without Decorrelation
With Decorrelation
Empirical probability of rejecting $\lambda = 0$
% of background observations rejected in validation
0.05
λ=0
λ=0.01
λ=0.015
λ=0.02
λ=0.03
λ=0.04
λ=0.05
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70

4b: Power



Without Decorrelation
With Decorrelation
Empirical probability of rejecting $\lambda = 0$
% of background observations rejected in validation
0.05
λ=0
λ=0.01
λ=0.015
λ=0.02
λ=0.03
λ=0.04
λ=0.05
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70

Comments and discussion

- CDOT can make any pre-trained classifier independent of given protected variables.
- CDOT can handle multiple or multivariate protected variables.
- Can be extended to multiple or multivariate classifiers but computationally expensive.
- Gives a range of transformed classifier using geodesic morphing.
- CDOT is robust to some background model misspecification.
- Overall, showed that both signal enrichment and decorrelation help increase power of detection.

Thank you! Questions?

arXiv: 2409.06399

Email: p.chakravarti@ucl.ac.uk