Robust Signal Detection using a Classifier Decorrelated through Optimal Transport (CDOT)

Purvasha Chakravarti

Department of Statistical Science
University College London
p.chakravarti@ucl.ac.uk

Joint work with Mikael Kuusela and Larry Wasserman, Carnegie Mellon University

ML4Jets 2022
Nov 3, 2022

GOAL: supervised signal detection when signal is known

- Model-dependent search: Search for NP signals when the signal model is known.
- Supervised classifier: Use a supervised classifier trained on MC simulations to perform cuts on the data.
- Decorrelation via Optimal Transport: Use Optimal Transport to make the classifier cuts independent of the protected variables (resonant features), e.g. the invariant mass.
- Test combining multiple cuts: Fit the BG distribution of the protected variable jointly for the different cuts.
- Robust to background misspecification: Check whether the procedure is robust to background misspecification.

Data

Two sources of data are at hand:

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

$$\text{Background:} \quad \mathcal{B}$$
$$\text{Signal:} \quad \mathcal{S}$$

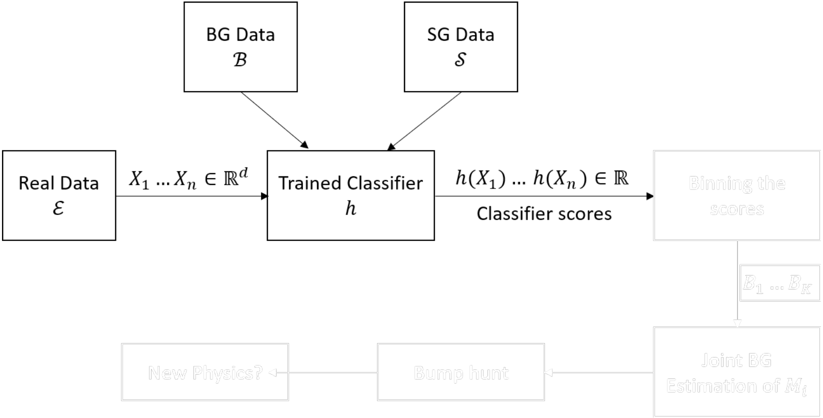
Used to train the classifier

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

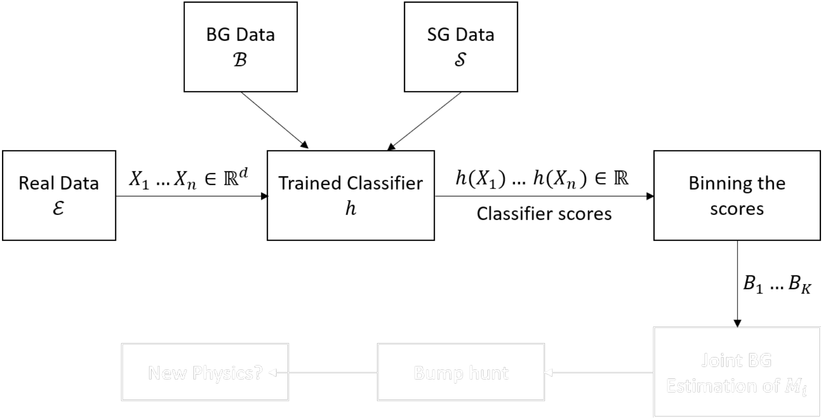
$$\text{Experimental:} \quad \mathcal{E} = \{X_1, \ldots, X_n\}$$
$$\text{Protected Variable:} \quad M_1, \ldots, M_n$$

Use $\mathcal{E}$ to perform cuts and $M_i's$ to perform signal detection using bump hunting.

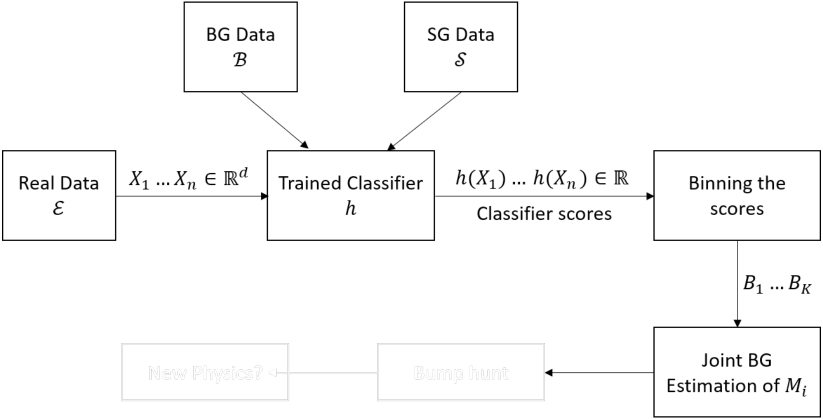
Signal detection process



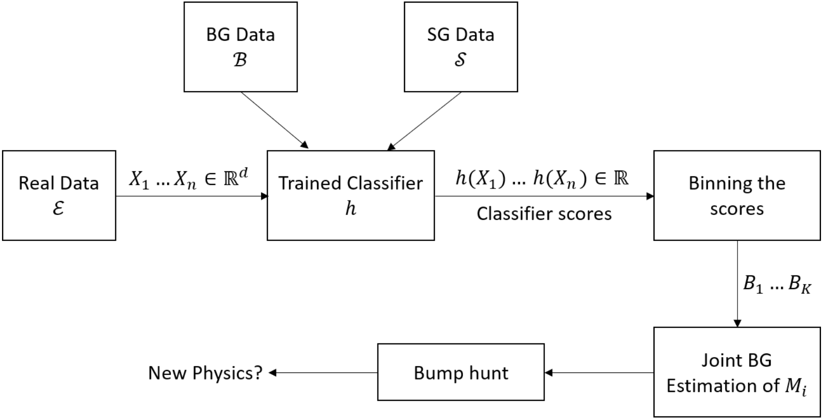
Signal detection process



Signal detection process



Signal detection process



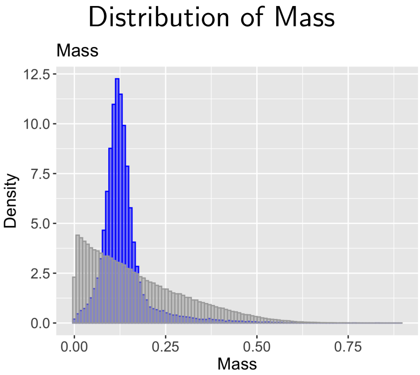
Problem with BG estimation: sculpting

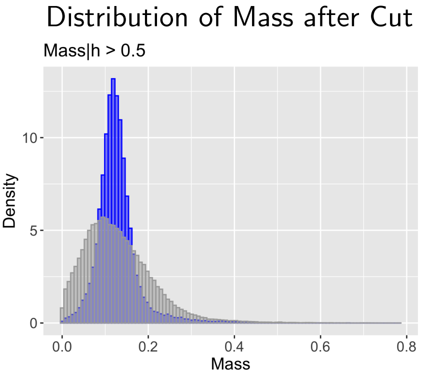
When we cut on the classifier scores the distribution of $M_i's$ changes!

Problem with BG estimation: sculpting

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

Example: Protected variable: Mass, Cut: Classfier output $h > 0.5$.
Grey: BG, Blue: SG





Idea behind decorrelation

Idea: Can the protected variable have the same background distribution after cuts as before cuts?

Idea behind decorrelation

Idea: Can the protected variable have the same background distribution after cuts as before cuts?

Need to make classifier output independent (not just *decorrelated*) of the protected variable for background data. (DisCo Fever [Kasieczka, Shih (2001.05310)], MoDe [Kitouni et al. (2010.09745)], etc)

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

Decorrelation via Optimal Transport

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

- Objective: Minimize $(T(h(X)) - h(X))^2$ subject to $T(h(X))$ independent of $M = m(X)$, given $X \sim \mathcal{B}$

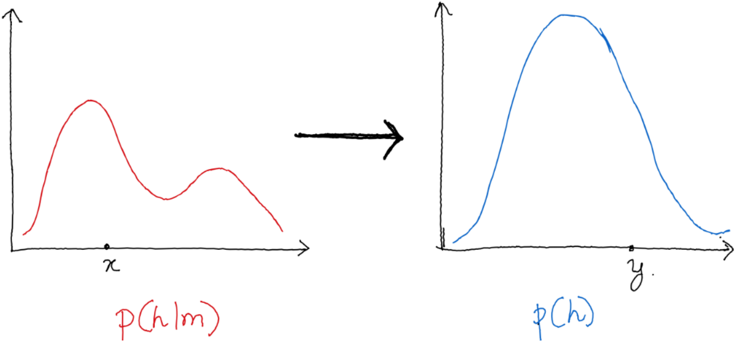
Decorrelation via Optimal Transport

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

- Objective: Minimize $(T(h(X)) - h(X))^2$ subject to $T(h(X))$ independent of $M = m(X)$, given $X \sim \mathcal{B}$
- When $T(h(X))|M$ has the same distribution as $T(h(X))$, then $T(h(X))$ is independent of M.
- The optimal transport map T_a from $p(h(x)|M = a, \mathcal{B})$ to the marginal $p(h(x)|\mathcal{B})$ is the solution.

Decorrelation via Optimal Transport

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



Decorrelation via Optimal Transport

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

- $h(X)$ is univariate.
- Closed form solution to Optimal Transport problem.

$$T_a(h(X)) = G^{-1}(F_{h|M}(h(X)|M = a))$$

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

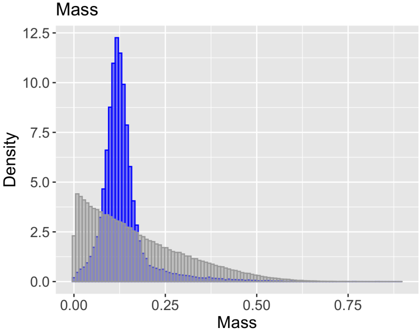
Solution is found by estimating G and $F_{h|M}$.

We call this Classifier Decorrelated through Optimal Transport (CDOT).

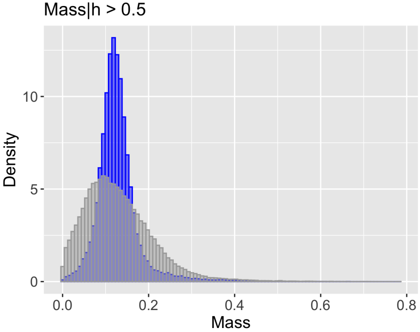
Sculpting problem solved!

Example: Protected variable: Mass, Cut: Classfier output $h > 0.5$.
Grey: BG, Blue: SG

Distribution of Mass



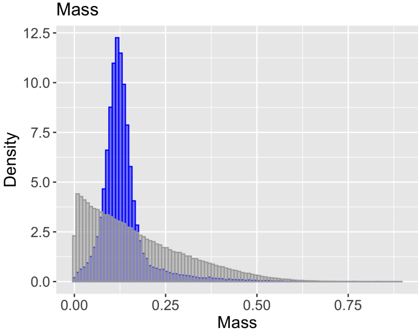
Distribution of Mass after Cut



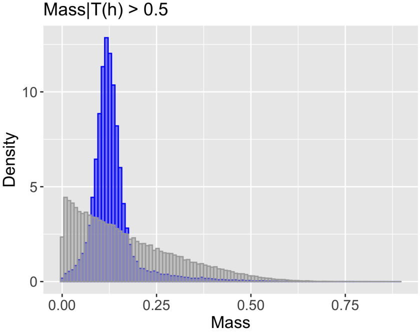
Sculpting problem solved!

Example: Protected variable: Mass, Cut: Classfier output $h > 0.5$.
Grey: BG, Blue: SG

Distribution of Mass

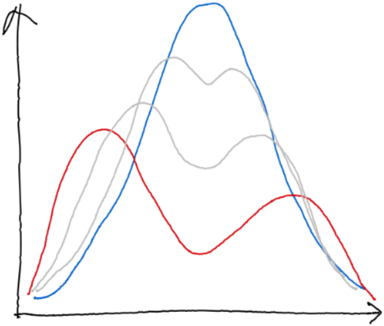


Distribution of Mass after Cut



Geodesic path of Optimal Transport

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



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

Discussion on existing decorrelation methods

- DisCo Fever [Kasieczka, Shih (2001.05310)]:
  - Based on "distance correlation", which is 0 iff variables are independent.
  - Added as a regularization term to the classifier loss function.

Discussion on existing decorrelation methods

- DisCo Fever [Kasieczka, Shih (2001.05310)]:
  - Based on "distance correlation", which is 0 iff variables are independent.
  - Added as a regularization term to the classifier loss function.

- MoDe [Kitouni et al. (2010.09745)]:
  - Regularization term is based on Legendre moments of conditional CDF of $h|M$.
  - MoDe loss with l^{th} moment is optimal when the mass dependence of the classifier is at most an l^{th} order polynomial.
  - $l = 0$ case is minimized iff variables are independent.

Discussion on existing decorrelation methods

- DisCo Fever [Kasieczka, Shih (2001.05310)]:
  - Based on "distance correlation", which is 0 iff variables are independent.
  - Added as a regularization term to the classifier loss function.

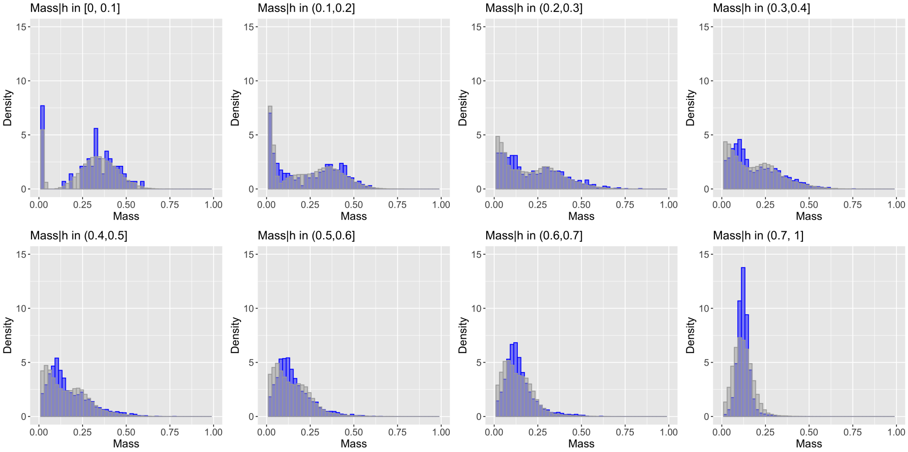
- MoDe [Kitouni et al. (2010.09745)]:
  - Regularization term is based on Legendre moments of conditional CDF of $h|M$.
  - MoDe loss with l^{th} moment is optimal when the mass dependence of the classifier is at most an l^{th} order polynomial.
  - $l = 0$ case is minimized iff variables are independent.

- Cuts derived from quantile regression [Moreno et al. (PhysRevD.102.012010)]:
  - Performs quantile regression to find cut $= \hat{Q}_{h|M}(\alpha)$.
  - $P(h > \text{cut}|M) = 1 - \alpha \; \forall \; m$.
  - Binning is a function of m and hence random.

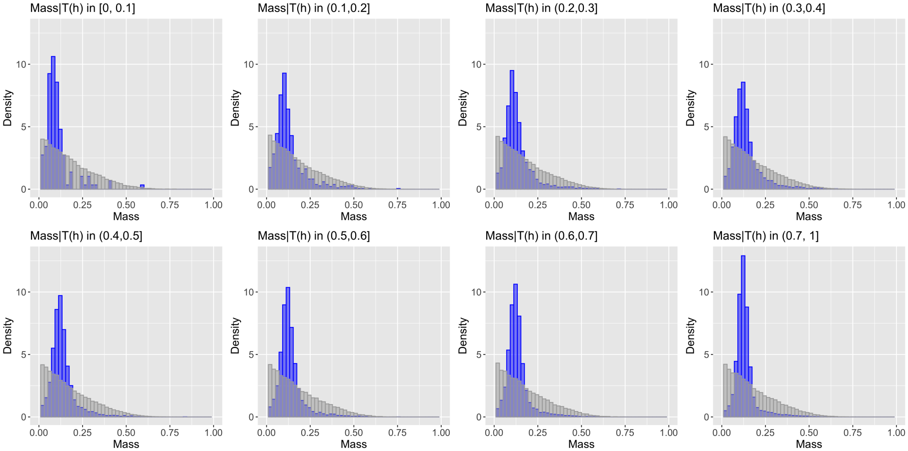
WTagging dataset

- 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

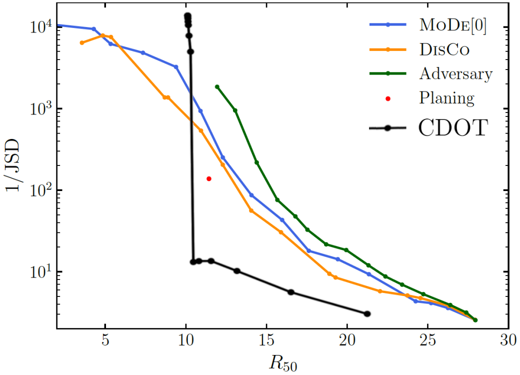


WTagging dataset: after OT transformation



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.

Simulated Data

- 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 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)]

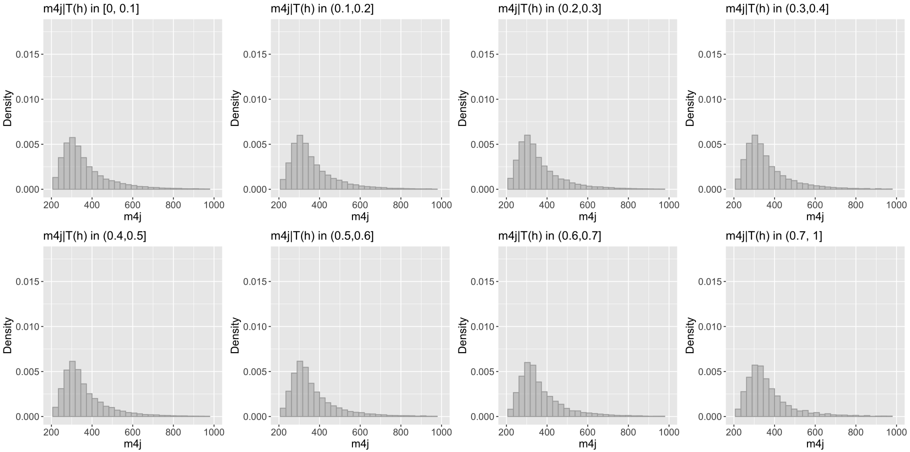
MC Background: $3b$ $(50,000)$

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

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

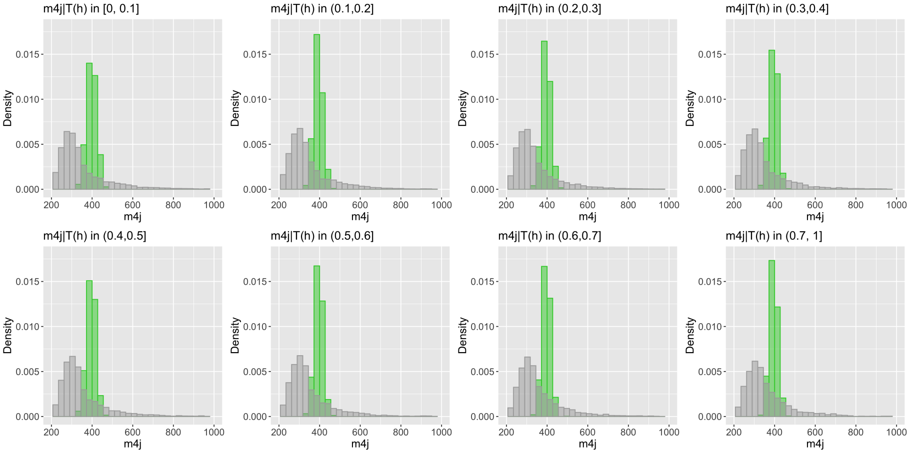
Simulated Data: OT and classifier trained on 3b data

CDOT trained on the 3b data and signal.



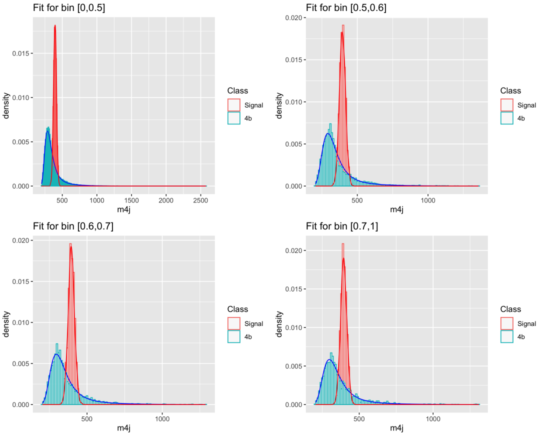
Simulated Data: robust on 4b data with signal

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



BG joint estimation and bump hunt

- Fit a joint model for all the bins to estimate the BG distribution.
- Assume signal model is known.
- Perform bump hunt.



Summary

- Used a supervised classifier trained on MC simulations to perform cuts on the data.
- Used Optimal Transport to make the classifier cuts independent of the protected variables (resonant features).
- Fit the BG distribution of the protected variable jointly for the different cuts.
- Compared CDOT to other decorrelation methods.
- Checked that the procedure is robust to background misspecification.

Future work

- Find the ideal test for bump hunting jointly in all the bins.
- Compare the decorrelation method to when used with quantile regression.
- Analyze to find what perturbations in background the method is robust towards.

Thank you!

Contact email: p.chakravarti@ucl.ac.uk