Conditional independence measures for fairer, more reliable models

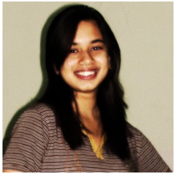
Danica J. Sutherland
UBC + Amii; she/her

based on joint work with:

Roman Pogodin
Gatsby, UCL → McGill + Mila

(both)

Namrata Deka
UBC → CMU

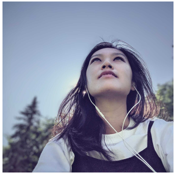


(CIRCE)

Antonin Schrab
Centre for AI + Gatsby, UCL

(SplitKCI)

Yazhe Li
Gatsby, UCL + DeepMind



(both)

Victor Veitch
UChicago + Google

(CIRCE)

Arthur Gretton
Gatsby, UCL + DeepMind

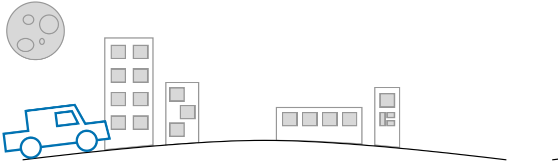
(both)

BIRS workshop: Statistical Aspects of Trustworthy Machine Learning, Feb 2023

CIRCE is arXiv:2212.08645 (ICLR 2023, "notable: top 5%")
SplitKCI is new work in submission; on arXiv soon...

Intro: conditionally invariant representations

- Self-driving car tries to predict its location
- Starts in the morning
- Finishes in the evening
- ...learns to predict **location** from **time of day**



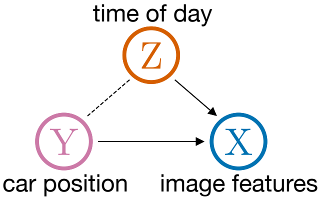
Distribution shift: car starts in the afternoon

- ...and makes lots of errors

Intro: conditionally invariant representations

Idealized solution to this **distribution shift** problem:

- predictions should be **conditionally independent** of time given the car position: $X \perp\!\!\!\perp Z \mid Y$



Same form as a common **domain invariance** objective:

features $\perp\!\!\!\perp$ domain ID $\mid$ true label

Same form as common **fairness** criterion (equalized odds):

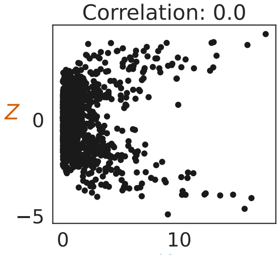
predictions $\perp\!\!\!\perp$ protected attribute $\mid$ true label

Problem: *conditional* dependence is **hard to measure**!

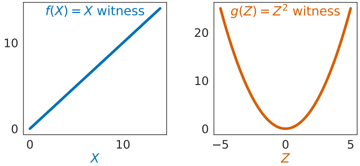
- Discrete Y: check dependence of X and Z for *each* Y value
  - On *each minibatch* during training...
- Continuous Y: classical methods need strong assumptions
  - e.g. joint Gaussianity (then check partial correlation)

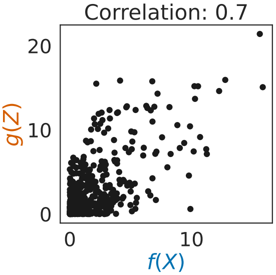
Warmup: detecting **unconditional** dependence

$$
\begin{aligned}
Y &\sim \mathcal{N}(0,1) \\
\xi_1, \xi_2 &\sim \mathcal{N}(0,1) \text{ i.i.d. noise} \\
X &= (Y + \xi_1)^2 \\
Z &= Y + \xi_1 + \xi_2
\end{aligned}
$$



- X and Z are **uncorrelated**



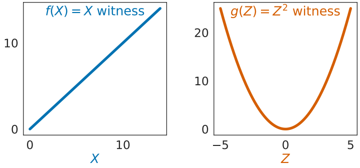


$X \perp\!\!\!\perp Z$ if and only if **all** square-integrable functions $f(X)$ and $g(Z)$ are uncorrelated

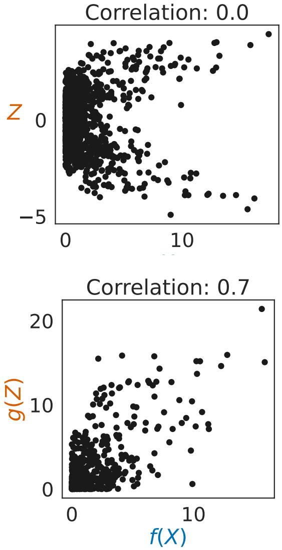
Warmup: detecting **unconditional** dependence

$$\begin{aligned} Y &\sim \mathcal{N}(0,1) \\ \xi_1, \xi_2 &\sim \mathcal{N}(0,1) \text{ i.i.d. noise} \\ X &= (Y + \xi_1)^2 \\ Z &= Y + \xi_1 + \xi_2 \end{aligned}$$

- X and Z are **uncorrelated**
- One way to detect dependence: we can find correlated **nonlinear** functions $f(X)$ and $g(Z)$

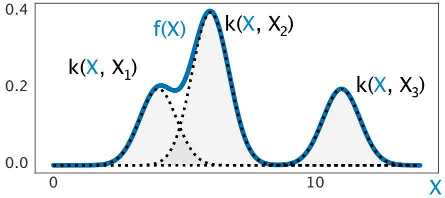


$X \perp\!\!\!\perp Z$ if and only if **all** square-integrable functions $f(X)$ and $g(Z)$ are uncorrelated



Warmup: detecting **unconditional** dependence

- If there aren't *any* correlated $f(X)$ and $g(Z)$, then X and Z are independent
- How to check ~~all~~ *enough* nonlinear functions?
- Check $f(X)$ and $g(Z)$ from **kernel spaces** (RKHSs): $f(X) = \sum_i \alpha_i \, k(X, X_i)$



- From RKHS properties: $\mathrm{Cov}\left(f(X), g(Z)\right) = \langle f, C_{XZ}\, g\rangle$ for the linear operator

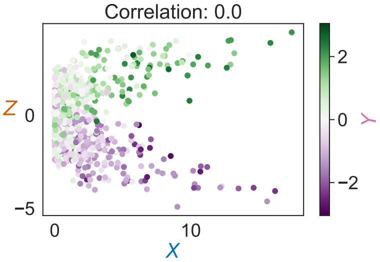
$$C_{XZ} = \mathbb{E}[k(X,\cdot) \otimes k(Z,\cdot)] - \mathbb{E}[k(X,\cdot)] \otimes \mathbb{E}[k(Z,\cdot)]$$

  - With linear kernels, C_{XZ} is just the cross-covariance matrix $\mathbb{E}[XZ^\top] - \mathbb{E}[X]\,\mathbb{E}[Z]^\top$
  - If $C_{XZ} = 0$, all $f(X)$ and $g(Z)$ in the RKHSes are uncorrelated
  - If our kernels are "rich enough" (Gaussian is enough), this implies independence
- Hilbert-Schmidt Independence Criterion: $\mathrm{HSIC}\left(X, Z\right) = \|C_{XZ}\|^2_{\mathrm{HS}} = 0$ iff $C_{XZ} = 0$
  - Can estimate with $\widehat{\mathrm{HSIC}}(X, Z) = \frac{1}{B^2} \mathbf{1}^\top \left(H K_{XX} H \odot K_{ZZ}\right) \mathbf{1}$, where H is "centring matrix"
- Deep nets with features X_θ ~independent of Z: $\min_\phi \mathrm{loss}(\phi(X), Y) + \gamma \, \widehat{\mathrm{HSIC}}(\phi(X), Z)$

Detecting **conditional** dependence

$$
\begin{aligned}
Y &\sim \mathcal{N}(0,1) \\
\xi_1, \xi_2 &\sim \mathcal{N}(0,1) \text{ i.i.d. noise} \\
X &= (Y + \xi_1)^2 \\
Z &= Y + \xi_1 + \xi_2
\end{aligned}
$$

- X and Z are **dependent**
- X and Z are **conditionally dependent** given Y (through ξ_1)



How do we characterize conditional (in)dependence?

- Start by just conditioning everything on Y: $X \perp\!\!\!\perp Z \mid Y$ iff for all $f_Y \in L^2_X$ and $g_Y \in L^2_Z$,

$$\mathbb{E}_{XZ}[f_Y(X)\, g_Y(Z) \mid Y] = \mathbb{E}_X[f_Y(X) \mid Y]\, \mathbb{E}_Z[g_Y(Z) \mid Y] \qquad Y\text{-a.s.}$$

- Equivalent: $X \perp\!\!\!\perp Z \mid Y$ iff for all $f \in L^2_{XY}$ and $g \in L^2_{ZY}$,

$$\mathbb{E}_{XZ}[f(X,Y)\, g(Z,Y) \mid Y] = \mathbb{E}_X[f(X,Y) \mid Y]\, \mathbb{E}_Z[g(Z,Y) \mid Y] \qquad Y\text{-a.s.}$$

- Equivalent (Daudin 1980): $X \perp\!\!\!\perp Z \mid Y$ iff for all $\tilde{f} \in L^2_{XY}$ such that $\mathbb{E}_X[\tilde{f}(X,Y) \mid Y] = 0 \quad Y$-a.s. and all $\tilde{g} \in L^2_{ZY}$ such that $\mathbb{E}_Z[\tilde{g}(Z,Y) \mid Y] = 0 \quad Y$-a.s.,

$$\mathbb{E}\left[\tilde{f}(X,Y)\, \tilde{g}(Z,Y)\right] = 0$$

proof

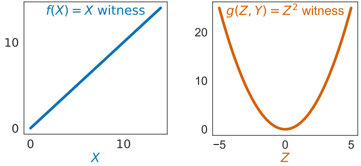
- Equivalent: $X \perp\!\!\!\perp Z \mid Y$ iff for all $f \in L^2_X$, $g \in L^2_{ZY}$,

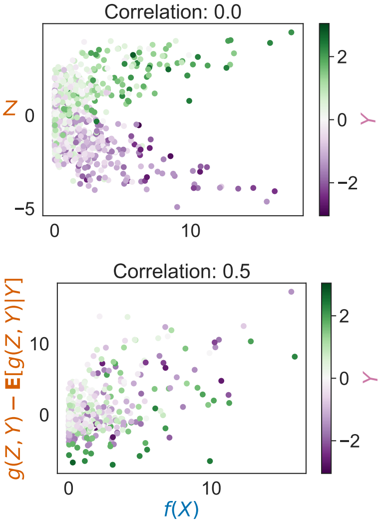
$$\mathbb{E}\Big[f(X)\left(g(Z,Y) - \mathbb{E}_Z[g(Z,Y) \mid Y]\right)\Big] = 0$$

Detecting **conditional** dependence

$$
\begin{aligned}
Y &\sim \mathcal{N}(0,1) \\
\xi_1, \xi_2 &\sim \mathcal{N}(0,1) \text{ i.i.d. noise} \\
X &= (Y + \xi_1)^2 \\
Z &= Y + \xi_1 + \xi_2
\end{aligned}
$$

- X and Z are **dependent**
- X and Z are **conditionally dependent** given Y (through ξ_1)





$X \perp\!\!\!\perp Z \mid Y$ if and only if **all** $f(X)$ are uncorrelated with **all** $g(Z,Y) - \mathbb{E}\left[g(Z,Y) \mid Y\right]$ [Daudin 1980]

CIRCE: Conditional Independence Regression CovariancE

- Want to check covariance of $f(X)$ and $g^c(Z,Y) = g(Z,Y) - \mathbb{E}\left[g(Z,Y) \mid Y\right]$
  - $g^c(Z,Y)$ has mean zero, so they're uncorrelated iff $\mathbb{E}[f(X)\, g^c(Z,Y)] = 0$
- The **CIRCE operator** gives $\langle f, C^c_{XZ|Y}\, g\rangle = \mathbb{E}[f(X)\, g^c(Z,Y)]$, using

$$C^c_{XZ|Y} = \mathbb{E}\Big[k(X,\cdot) \otimes \big(k((Z,Y),\cdot) - \mathbb{E}[k((Z',Y),\cdot) \mid Y]\big)\Big]$$

- $\mathrm{CIRCE}(X,Z \mid Y) = \|C^c_{XZ|Y}\|^2_{\mathrm{HS}} = 0$ iff $X \perp\!\!\!\perp Z \mid Y$, if kernels are "rich enough"
- Important special case: if $k((Z,Y),(Z',Y')) = k(Z,Z')\, k(Y,Y')$, we get

$$C^c_{XZ|Y} = \mathbb{E}[k(X,\cdot) \otimes k(Y,\cdot) \otimes \big(k(Z,\cdot) - \mu_{Z|Y}(Y)\big)]$$

where $\mu_{Z|Y} = \mathbb{E}[k(Z,\cdot) \mid Y]$ is the **conditional mean embedding** of Z given Y

CIRCE estimator

- Want squared norm of $C^c_{XZ|Y} = \mathbb{E}[k(X,\cdot) \otimes k(Y,\cdot) \otimes \left(k(Z,\cdot) - \mu_{Z|Y}(Y)\right)]$
- First, estimate conditional mean embedding $\mu_{Z|Y}$ on a dataset $\{(Z_i, Y_i)\}_{i=1}^M$
  - Use kernel ridge regression: inputs Y, RKHS-valued labels $k(Z,\cdot)$
  - Use this to estimate the conditionally-centred kernel function

$$\begin{aligned}\hat{k}^c((Z,Y),(Z',Y')) &= \langle k(Z,\cdot) - \hat{\mu}_{Z|Y}(Y), k(Z',\cdot) - \hat{\mu}_{Z|Y}(Y')\rangle \\ &\approx k(Z,Z') - \mathbb{E}[k(Z,Z') \mid Y] - \mathbb{E}[k(Z,Z') \mid Y'] + \mathbb{E}[k(Z,Z') \mid Y, Y']\end{aligned}$$

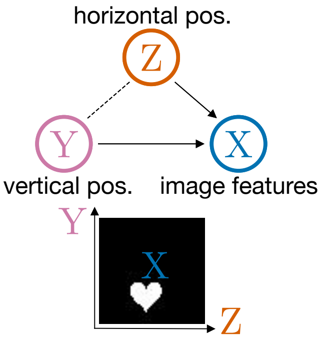
- While training $\phi(X)$, for each batch $\{(\phi(X_i), Z_i, Y_i)\}_{i=1}^B$:
  - Get $(K_{XX})_{ij} = k(\phi(X_i), \phi(X_j))$, $(K_{YY})_{ij} = k(Y_i, Y_j)$, $(\hat{K}^c_{ZZ})_{ij} = \hat{k}^c((Z,Y),(Z',Y'))$
  - Regularize with $\widehat{\mathrm{CIRCE}} = \frac{1}{B(B-1)} \mathbf{1}^\top \left(K_{XX} \odot K_{YY} \odot \hat{K}^c_{ZZ}\right)\mathbf{1}$

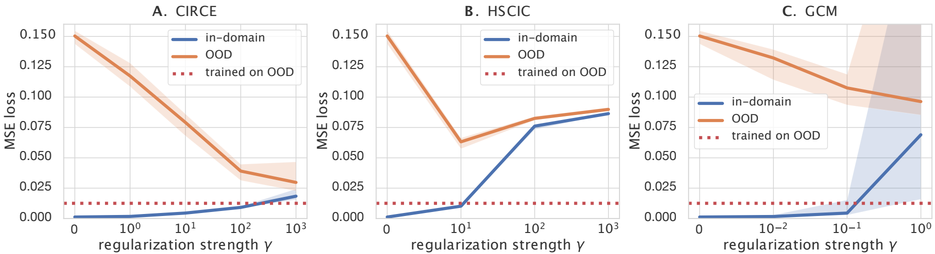
Benefits of CIRCE:

- As $B, M \to \infty$, $\widehat{\mathrm{CIRCE}} \to 0$ iff $\phi(X) \perp\!\!\!\perp Z \mid Y$; rate is known (see paper)
- K_{YY} and $\hat{K}^c_{ZZ}$ don't depend on ϕ:
  - Can precompute them, so only need $k(\phi(X_i), \phi(X_j))$ for each new ϕ
  - Separates (small) batch size B and (big) regression training size M: better convergence

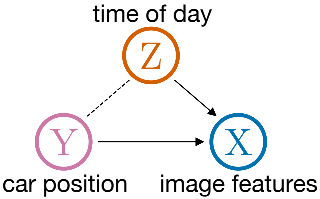
Experiments

- dSprites dataset [Matthey et al., 2017]:
  2D shapes in different locations
- Task: predict vertical position Y
  But be invariant to horizontal position Z
  Z and Y have strong dependence in training
- Compare to HSCIC [Quinzan et al., 2022] (also kernel-based) and GCM [Shah & Peters, 2020] (correlation-based)
- CIRCE wins!





CIRCE discussion



- **CIRCE**: a measure of conditional independence for feature learning
- It works with continuous variables and in deep learning settings
- Applications: domain shift invariance, fairness
- Ongoing: learn kernels on Y (straightforward) and Z (harder)
- Next: testing whether $X \perp\!\!\!\perp Z \mid Y$

arXiv paper

(code link inside)

Testing

- Learning with a CIRCE regularizer tries to learn a model where $X \perp\!\!\!\perp Z \mid Y$
- ...did it work?
- Or: lots of other interesting conditional independence questions to ask!
  - Is car insurance price (X) $\perp\!\!\!\perp$ neighbourhood's racial makeup (Z) | driver risk (Y)?

- We'll take a **null hypothesis significance testing** approach
- $\mathfrak{H}_0 : X \perp\!\!\!\perp Z \mid Y$; alternative hypothesis is just "not that"
- Assuming good-enough kernels, equivalent to ask whether $\mathrm{CIRCE}(X, Z \mid Y) = 0$:

$$\text{is } \left\| \mathbb{E}\left[k(X, \cdot) \otimes k(Y, \cdot) \otimes \big(k(Z, \cdot) - \mu_{Z|Y}(Y) \big) \right] \right\|^2 = 0?$$

- Problem: estimating the conditional mean, $\mu_{Z|Y}(Y)$, is really hard!
  - *Best-case* minimax rate is $\mathcal{O}\left(1/m^{1/4}\right)$; can be *arbitrarily slow* (Li et al. 2022)
  - Rate for "everything else given a $\hat{\mu}_{Z|Y}$" is $\mathcal{O}(1/\sqrt{n})$

Bias

- What happens when $\hat{\mu}_{Z|Y} = \mu_{Z|Y} + \Delta_{Z|Y}$, with $\Delta_{Z|Y} \neq 0$, when $X \perp\!\!\!\perp Z \mid Y$?

$$\left\| \mathbb{E}\Big[k(X,\cdot) \otimes k(Y,\cdot) \otimes \big(k(Z,\cdot) - \mu_{Z|Y}(Y) - \Delta_{Z|Y}(Y) \big) \Big] \right\|^2$$

$$= \Bigg\| \underbrace{\mathbb{E}\left[k(X,\cdot) \otimes k(Y,\cdot) \otimes \big(k(Z,\cdot) - \mu_{Z|Y}(Y) \big) \right]}_{0, \text{ since } X \perp\!\!\!\perp Z \mid Y} - \mathbb{E}\left[k(X,\cdot) \otimes k(Y,\cdot) \otimes \Delta_{Z|Y}(Y) \right] \Bigg\|^2$$

$$= \mathbb{E}\Big[k(X,X')\, k(Y,Y')\, \underbrace{\langle \Delta_{Z|Y}(Y), \Delta_{Z|Y}(Y') \rangle}_{\text{likely big if } k(Y,Y') \text{ is big}} \Big]$$

- If we estimated the regression wrong, it *doesn't matter how many samples we get* for the rest of the estimator: $\widehat{\mathrm{CIRCE}}$ will be big
  - Understanding *how* big is hard

(Split)KCI

- When used during training a deep model, it helped to only use one regression
- For testing, this is less relevant
- Instead of the CIRCE operator, the **KCI operator** (Zhang et al. 2012) is

$$C^{\mathrm{KCI}}_{XZ|Y} = \mathbb{E}\left[\big(k(X,\cdot) - \mu_{X|Y}(Y)\big) \otimes k(Y,\cdot) \otimes \big(k(Z,\cdot) - \mu_{Z|Y}(Y)\big)\right]$$

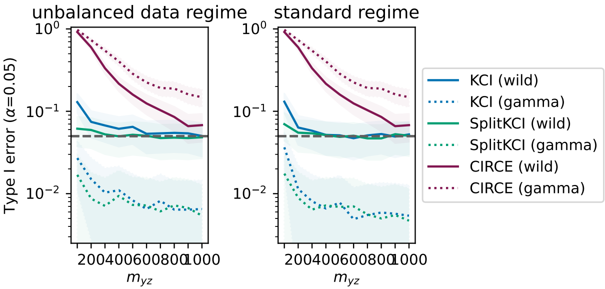
- $C^{\mathrm{KCI}}_{XZ|Y} = 0$ iff $X \perp\!\!\!\perp Z \mid Y$; with incorrect regressions, bias becomes

$$\mathbb{E}\big[\langle \Delta_{X|Y}(X), \Delta_{X|Y}(X') \rangle \, k(Y,Y') \, \langle \Delta_{Z|Y}(Y), \Delta_{Z|Y}(Y') \rangle\big]$$

- Can reduce this by replacing $\langle \Delta_{X|Y}(X), \Delta_{X|Y}(X') \rangle$ with $\langle \Delta^{(1)}_{X|Y}(X), \Delta^{(2)}_{X|Y}(X') \rangle$
  - Compute by using two different regressions: split the data used to train it
  - The regression is really hard, so it's annoying to not use all the data
  - ...but the regression is *so* hard that losing half the data doesn't hurt *that* much
- Everything still works out since other one is centred (like CIRCE)
- Can even use **different kernels** (not necessarily universal!) – any arbitrary functions
  - Simple kernels might help: faster convergence, still debiasing

Testing with SplitKCI

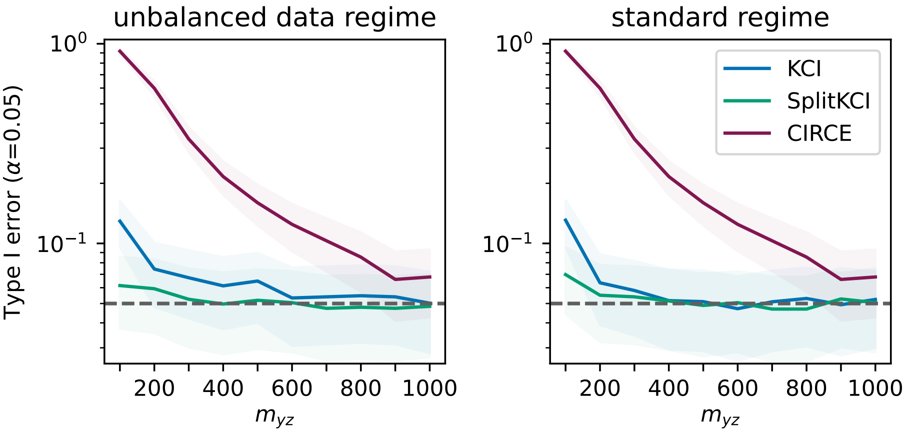
- Zhang et al. (2012) tested based on a gamma approximation to the null distribution
- That approximation can't cope with the bias when mean estimation is poor



- Instead, use **wild bootstrap**
  - Approximate null distribution by element-wise multiplying the centred kernel matrix by $qq^\top$, q a vector of random signs
  - Can prove it works (asymptotically), as long as we have enough regression samples

Better Type I error control

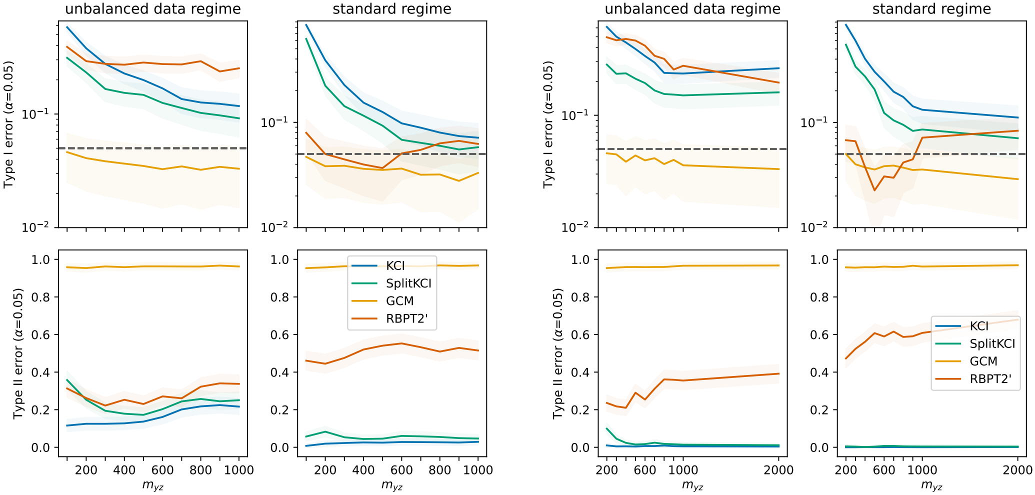
- On synthetic Gaussian data:



- Indications of similar results on real car insurance data

More powerful than competitors

- Different synthetic task; left side is $n = 100$, right is $n = 200$



Discussion

- **CIRCE**: a measure of conditional independence for feature learning
  - Works with continuous variables, in deep learning settings
  - Applications to fairness, domain shift, ...
  - Ongoing extension: learn kernels on Y (straightforward) and/or Z (harder)

- Unfortunately, CIRCE is really bad at testing
- Bias seems to be a big factor for it and its predecessor KCI
- **SplitKCI**: an "in-between" statistic based on data splitting
  - Debiasing with data splitting
  - Want to use a lot more data for regression than rest of test
    - Good setting: limited (X, Z, Y) triples, but lots of (X, Y) and (Z, Y) pairs
  - Wild bootstrap for estimating the test threshold

Characterizing conditional (in)dependence – proof sketch

back

$$\forall f \in L^2_{XY},\ g \in L^2_{ZY}, \qquad \mathbb{E}[fg \mid Y] = \mathbb{E}[f \mid Y]\mathbb{E}[g \mid Y] \tag{A}$$

$\Updownarrow$ [Daudin 1980]

$$\forall \tilde{f} \in L^2_{XY},\ \tilde{g} \in L^2_{ZY} \text{ s.t. } \mathbb{E}[\tilde{f} \mid Y] = 0 = \mathbb{E}[\tilde{g} \mid Y], \qquad \mathbb{E}[\tilde{f}\tilde{g}] = 0 \tag{B}$$

- (A) $\implies$ (B), (C): Just apply (A) to $\tilde{f}$ and $\tilde{g}$, RHS becomes 0
- (B) $\implies$ (A):
  - Choose $\tilde{f}(X,Y) = f(X,Y) - \mathbb{E}[f(X,Y) \mid Y]$ and $\tilde{g}(Z,Y) = g(Z,Y) - \mathbb{E}[g(Z,Y) \mid Y]$.
  - $0 = \mathbb{E}[\tilde{f}\tilde{g}] = \mathbb{E}_Y\left[\mathbb{E}[\tilde{f}\tilde{g} \mid Y]\right] = \mathbb{E}_Y\left[\mathbb{E}[fg \mid Y] - \mathbb{E}[f \mid Y]\mathbb{E}[g \mid Y]\right]$
  - Letting g include an indicator on sets of Y, implies must hold almost surely in Y
- Same basic idea works for uncentred $f \in L^2_{XY}$ and centred $g \in L^2_{ZY}$
- Slightly more argument to drop the Y in f