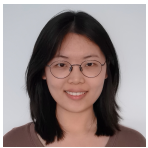
Detecting conditional dependence in practice: why it's so hard, and what we can do

Danica J. Sutherland
UBC + Amii; she/her

based on joint work with:

Zheng He
UBC



(hardness)

Roman Pogodin
UCL → McGill + Mila → Google

(all)

Namrata Deka
UBC → CMU

(CIRCE, hardness)

Antonin Schrab
UCL

(SplitKCI)

Yazhe Li
UCL + DeepMind → Microsoft

(all)

Victor Veitch
UChicago + Google

(CIRCE)

Arthur Gretton
UCL + DeepMind

(all)

UW Biostatistics Seminar, November 2025

CIRCE is arXiv:2212.08645 (ICLR 2023, "notable: top 5%")
SplitKCI is arXiv:2402.13196 (in submission)
hardness is not on arXiv yet (soon!) (NeurIPS 2025, spotlight)

Conditional independence

- Are X and Y associated, even if I "control for" Z?
- Are the shapes of cancer cells associated with drug dosage, controlling for disease stage?

Conditional independence

- Are X and Y associated, even if I "control for" Z?
- Are the shapes of cancer cells associated with drug dosage, controlling for disease stage?
- Should I have an edge between X and Y in my causal DAG? If $X \perp\!\!\!\perp Y \mid Z$, no!

Conditional independence

- Are X and Y associated, even if I "control for" Z?
- Are the shapes of cancer cells associated with drug dosage, controlling for disease stage?
- Should I have an edge between X and Y in my causal DAG? If $X \perp\!\!\!\perp Y \mid Z$, no!
- Are loan predictions associated with applicant's race, conditioned on whether they'll pay back the loan?

Conditional independence

- Are X and Y associated, even if I "control for" Z?
- Are the shapes of cancer cells associated with drug dosage, controlling for disease stage?
- Should I have an edge between X and Y in my causal DAG? If $X \perp\!\!\!\perp Y \mid Z$, no!
- Are loan predictions associated with applicant's race, conditioned on whether they'll pay back the loan?
- Do my predictions of where pedestrians are depend on if I'm driving in Manhattan or Manitoba, conditioned on where the pedestrians actually are?

Conditional independence

- Are X and Y associated, even if I "control for" Z?
- Are the shapes of cancer cells associated with drug dosage, controlling for disease stage?
- Should I have an edge between X and Y in my causal DAG? If $X \perp\!\!\!\perp Y \mid Z$, no!
- Are loan predictions associated with applicant's race, conditioned on whether they'll pay back the loan?
- Do my predictions of where pedestrians are depend on if I'm driving in Manhattan or Manitoba, conditioned on where the pedestrians actually are?

- $X \perp\!\!\!\perp Y \mid Z$ iff the joint distribution factorizes

Conditional independence

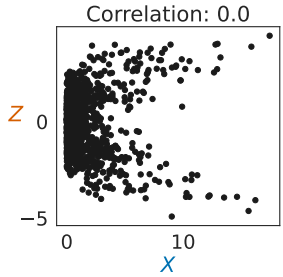
- Are X and Y associated, even if I "control for" Z?
- Are the shapes of cancer cells associated with drug dosage, controlling for disease stage?
- Should I have an edge between X and Y in my causal DAG? If $X \perp\!\!\!\perp Y \mid Z$, no!
- Are loan predictions associated with applicant's race, conditioned on whether they'll pay back the loan?
- Do my predictions of where pedestrians are depend on if I'm driving in Manhattan or Manitoba, conditioned on where the pedestrians actually are?

- $X \perp\!\!\!\perp Y \mid Z$ iff the joint distribution factorizes
- If X, Y, Z are jointly Gaussian, occurs iff partial correlation is zero

Warmup: detecting **unconditional** dependence

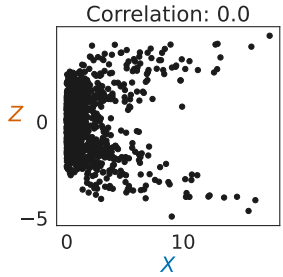
$$
\begin{aligned}
Z &\sim \mathcal{N}(0,1) \\
\xi_1, \xi_2 &\sim \mathcal{N}(0,1) \text{ i.i.d. noise} \\
X &= (Z + \xi_1)^2 \\
Y &= Z + \xi_1 + \xi_2
\end{aligned}
$$

- X and Y are **uncorrelated**



Warmup: detecting **unconditional** dependence

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

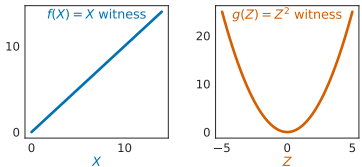


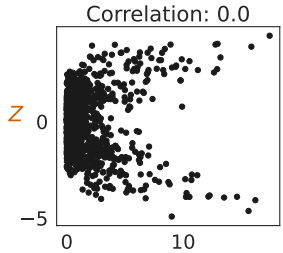
- X and Y are **uncorrelated**
- One way to detect dependence: we can find correlated **nonlinear** functions $f(X)$ and $g(Y)$

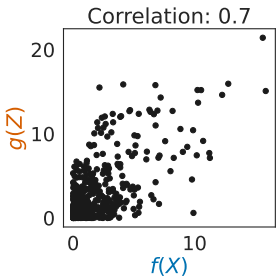
Warmup: detecting **unconditional** dependence

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

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



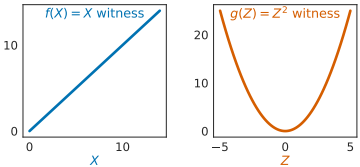




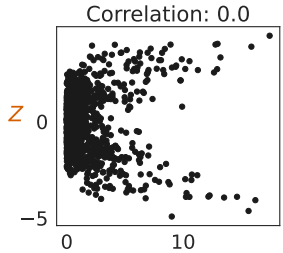
Warmup: detecting **unconditional** dependence

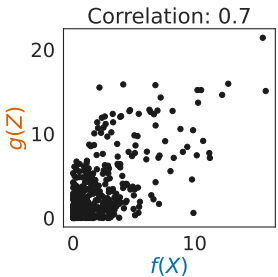
$$
\begin{aligned}
Z &\sim \mathcal{N}(0,1) \\
\xi_1, \xi_2 &\sim \mathcal{N}(0,1) \text{ i.i.d. noise} \\
X &= (Z+\xi_1)^2 \\
Y &= Z+\xi_1+\xi_2
\end{aligned}
$$

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



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





Warmup: detecting **unconditional** dependence

- If there aren't *any* correlated $f(X)$ and $g(Y)$, then X and Y are independent
- How to check all nonlinear functions?

Warmup: detecting **unconditional** dependence

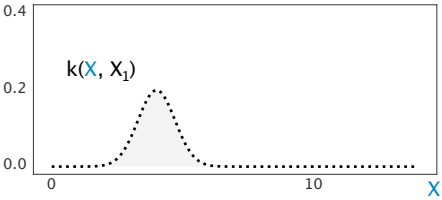
- If there aren't *any* correlated $f(X)$ and $g(Y)$, then X and Y are independent
- How to check ~~all~~ *enough* nonlinear functions?

Warmup: detecting **unconditional** dependence

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

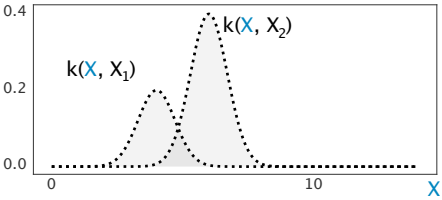
Warmup: detecting **unconditional** dependence

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



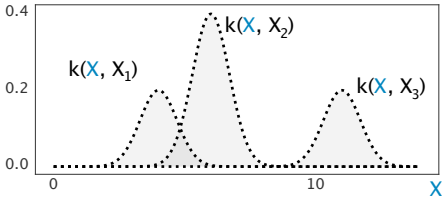
Warmup: detecting **unconditional** dependence

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



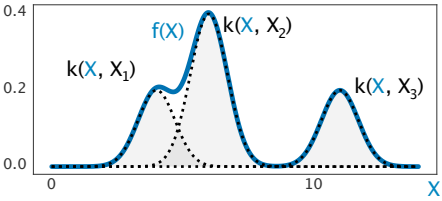
Warmup: detecting **unconditional** dependence

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



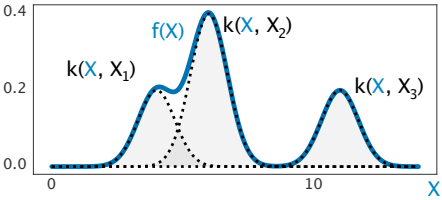
Warmup: detecting **unconditional** dependence

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



Warmup: detecting **unconditional** dependence

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



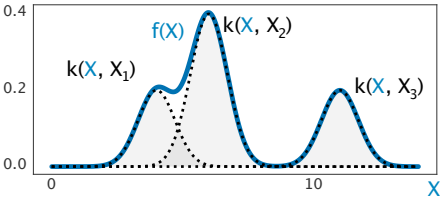
- From RKHS properties: $\mathrm{Cov}\,(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle$ for the linear operator

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

  - With linear kernels, C_{XY} is just the cross-covariance matrix $\mathbb{E}[XY^\top] - \mathbb{E}[X]\,\mathbb{E}[Y]^\top$

Warmup: detecting **unconditional** dependence

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



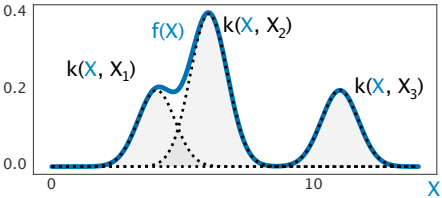
- From RKHS properties: $\mathrm{Cov}\,(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle$ for the linear operator

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

  - With linear kernels, C_{XY} is just the cross-covariance matrix $\mathbb{E}[XY^\top] - \mathbb{E}[X]\,\mathbb{E}[Y]^\top$
  - If $C_{XY} = 0$, all $f(X)$ and $g(Y)$ in the RKHSes are uncorrelated

Warmup: detecting **unconditional** dependence

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



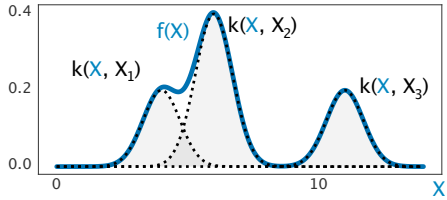
- From RKHS properties: $\mathrm{Cov}\,(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle$ for the linear operator

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

  - With linear kernels, C_{XY} is just the cross-covariance matrix $\mathbb{E}[XY^\top] - \mathbb{E}[X]\,\mathbb{E}[Y]^\top$
  - If $C_{XY} = 0$, all $f(X)$ and $g(Y)$ in the RKHSes are uncorrelated
  - If our kernels are "rich enough" (Gaussian is enough), this implies independence

Warmup: detecting **unconditional** dependence

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



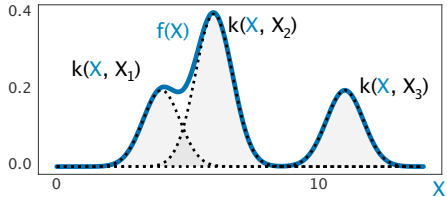
- From RKHS properties: $\mathrm{Cov}\,(f(X), g(Y)) = \langle f, C_{XY}\, g \rangle$ for the linear operator

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

  - With linear kernels, C_{XY} is just the cross-covariance matrix $\mathbb{E}[XY^\top] - \mathbb{E}[X]\,\mathbb{E}[Y]^\top$
  - If $C_{XY} = 0$, all $f(X)$ and $g(Y)$ in the RKHSes are uncorrelated
  - If our kernels are "rich enough" (Gaussian is enough), this implies independence
- Hilbert-Schmidt Independence Criterion: $\mathrm{HSIC}\,(X, Y) = \|C_{XY}\|^2_{\mathrm{HS}} = 0$ iff $C_{XY} = 0$

Warmup: detecting **unconditional** dependence

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



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

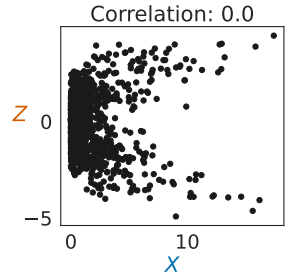
$$C_{XY} = \mathbb{E}[k(X,\cdot) \otimes k(Y,\cdot)] - \mathbb{E}[k(X,\cdot)] \otimes \mathbb{E}[k(Y,\cdot)]$$

  - With linear kernels, C_{XY} is just the cross-covariance matrix $\mathbb{E}[XY^\top] - \mathbb{E}[X]\,\mathbb{E}[Y]^\top$
  - If $C_{XY} = 0$, all $f(X)$ and $g(Y)$ in the RKHSes are uncorrelated
  - If our kernels are "rich enough" (Gaussian is enough), this implies independence
- Hilbert-Schmidt Independence Criterion: $\mathrm{HSIC}\,(X, Y) = \|C_{XY}\|^2_{\mathrm{HS}} = 0$ iff $C_{XY} = 0$
  - Can estimate with $\widehat{\mathrm{HSIC}}(X,Y) = \frac{1}{B^2}\mathbf{1}^\top \big(HK_{XX}H \odot K_{ZZ}\big)\mathbf{1}$, where H is "centring matrix"

Detecting **conditional** dependence

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

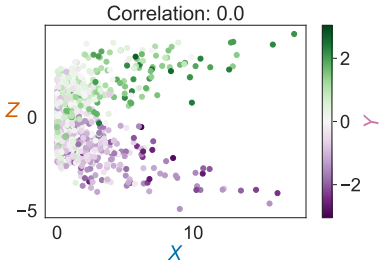
- X and Y are **dependent**



Detecting **conditional** dependence

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

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



How do we characterize conditional (in)dependence?

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

$$\mathop{\mathbb{E}}_{XY}[f_Z(X)\, g_Z(Y) \mid Z] = \mathop{\mathbb{E}}_{X}[f_Z(X) \mid Z]\ \mathop{\mathbb{E}}_{Y}[g_Z(Y) \mid Z] \qquad Z\text{-a.s.}$$

How do we characterize conditional (in)dependence?

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

$$\mathop{\mathbb{E}}_{XY}[f_Z(X)\, g_Z(Y) \mid Z] = \mathop{\mathbb{E}}_{X}[f_Z(X) \mid Z]\ \mathop{\mathbb{E}}_{Y}[g_Z(Y) \mid Z] \qquad Z\text{-a.s.}$$

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

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

How do we characterize conditional (in)dependence?

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

$$\mathop{\mathbb{E}}_{XY}[f_Z(X)\, g_Z(Y) \mid Z] = \mathop{\mathbb{E}}_{X}[f_Z(X) \mid Z]\ \mathop{\mathbb{E}}_{Y}[g_Z(Y) \mid Z] \qquad Z\text{-a.s.}$$

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

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

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

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

How do we characterize conditional (in)dependence?

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

$$\mathop{\mathbb{E}}_{XY}[f_Z(X)\, g_Z(Y) \mid Z] = \mathop{\mathbb{E}}_{X}[f_Z(X) \mid Z]\ \mathop{\mathbb{E}}_{Y}[g_Z(Y) \mid Z] \qquad Z\text{-a.s.}$$

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

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

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

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

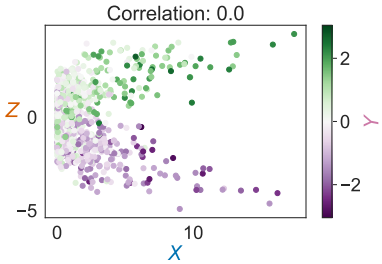
- Equivalent (us): $X \perp\!\!\!\perp Y \mid Z$ iff for all $f \in L^2_X$, $g \in L^2_Y$, $w \in L^2_Z$

$$\mathop{\mathbb{E}}_{Z}\Big[w(Z)\ \mathop{\mathbb{E}}_{XY}\big[\big(f(X) - \mathbb{E}\left[f(X) \mid Z\right]\big)\ \big(g(Y) - \mathbb{E}\left[g(Y) \mid Z\right]\big) \mid Z\big]\Big] = 0$$

Detecting **conditional** dependence

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

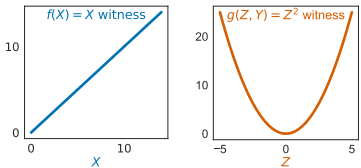
- X and Y are **dependent**
- X and Y are **conditionally dependent** given Z (through ξ_1)

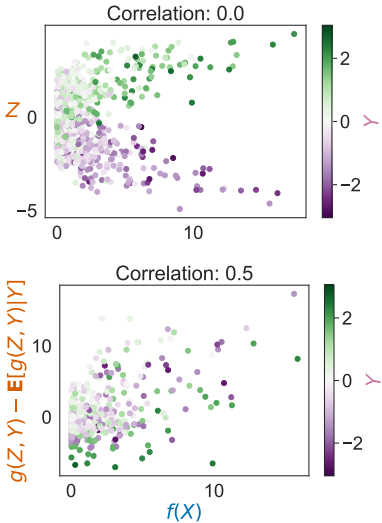


Detecting **conditional** dependence

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

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

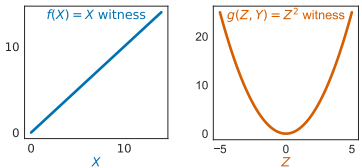




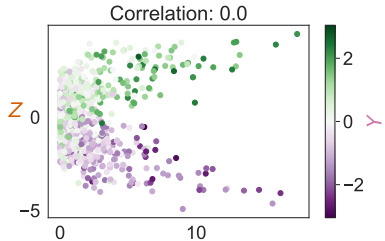
Detecting **conditional** dependence

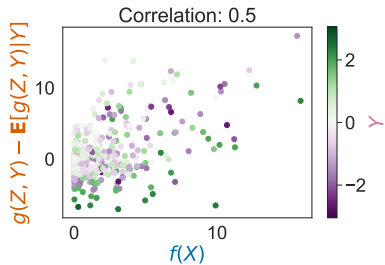
$$
\begin{aligned}
Z &\sim \mathcal{N}(0,1) \\
\xi_1, \xi_2 &\sim \mathcal{N}(0,1) \text{ i.i.d. noise} \\
X &= (Z+\xi_1)^2 \\
Y &= Z+\xi_1+\xi_2
\end{aligned}
$$

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



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





KCI: Kernel-based Conditional Independence (Zhang et al., UAI 2012; us)

- Want to check $w(Z)$-weighted covariance of $f(X) - \mathbb{E}\left[f(X) \,|\, Z\right]$ and $g(Y) - \mathbb{E}\left[g(Y) \,|\, Z\right]$
- The **conditional mean embedding** $\mu_{X|Z}(z) = \mathbb{E}\left[k(X, \cdot) \mid Z = z\right]$ gives us

$$\langle \mu_{X|Z}(z), f \rangle = \mathbb{E}[f(X) \mid Z = z] \quad \text{for any } f \text{ in the RKHS for } X$$

KCI: Kernel-based Conditional Independence (Zhang et al., UAI 2012; us)

- Want to check $w(Z)$-weighted covariance of $f(X) - \mathbb{E}\left[f(X) \,|\, Z\right]$ and $g(Y) - \mathbb{E}\left[g(Y) \,|\, Z\right]$
- The **conditional mean embedding** $\mu_{X|Z}(z) = \mathbb{E}\left[k(X, \cdot) \mid Z = z\right]$ gives us

$$\langle \mu_{X|Z}(z), f \rangle = \mathbb{E}[f(X) \mid Z = z] \quad \text{for any } f \text{ in the RKHS for } X$$

  - For $k(X, X') = X \cdot X'$, this is $\mu_{X|Z}(z) = \mathbb{E}[X \mid Z = z]$

KCI: Kernel-based Conditional Independence (Zhang et al., UAI 2012; us)

- Want to check $w(Z)$-weighted covariance of $f(X) - \mathbb{E}\left[f(X) \mid Z\right]$ and $g(Y) - \mathbb{E}\left[g(Y) \mid Z\right]$
- The **conditional mean embedding** $\mu_{X|Z}(z) = \mathbb{E}\left[k(X, \cdot) \mid Z = z\right]$ gives us

$$\langle \mu_{X|Z}(z), f \rangle = \mathbb{E}[f(X) \mid Z = z] \quad \text{for any } f \text{ in the RKHS for } X$$

  - For $k(X, X') = X \cdot X'$, this is $\mu_{X|Z}(z) = \mathbb{E}[X \mid Z = z]$
- The **conditional cross-covariance operator**,

$$\mathfrak{C}_{XY|Z}(z) = \mathbb{E}\left[\left(k(X, \cdot) - \mu_{X|Z}(z)\right) \otimes \left(k(Y, \cdot) - \mu_{Y|Z}(z)\right) \mid Z = z\right]$$

  lets us evaluate $\langle f, \mathfrak{C}_{XY|Z=z} g \rangle = \mathrm{Cov}(f(X), g(Y) \mid Z = z)$

KCI: Kernel-based Conditional Independence (Zhang et al., UAI 2012; us)

- Want to check $w(Z)$-weighted covariance of $f(X) - \mathbb{E}\left[f(X) \mid Z\right]$ and $g(Y) - \mathbb{E}\left[g(Y) \mid Z\right]$
- The **conditional mean embedding** $\mu_{X|Z}(z) = \mathbb{E}\left[k(X, \cdot) \mid Z = z\right]$ gives us

$$\langle \mu_{X|Z}(z), f \rangle = \mathbb{E}[f(X) \mid Z = z] \quad \text{for any } f \text{ in the RKHS for } X$$

  - For $k(X, X') = X \cdot X'$, this is $\mu_{X|Z}(z) = \mathbb{E}[X \mid Z = z]$
- The **conditional cross-covariance operator**,

$$\mathfrak{C}_{XY|Z}(z) = \mathbb{E}\left[\left(k(X, \cdot) - \mu_{X|Z}(z)\right) \otimes \left(k(Y, \cdot) - \mu_{Y|Z}(z)\right) \mid Z = z\right]$$

  lets us evaluate $\langle f, \mathfrak{C}_{XY|Z=z} g \rangle = \mathrm{Cov}(f(X), g(Y) \mid Z = z)$
- The **KCI operator** aggregates this over Z: $\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z}(Z) \otimes k(Z, \cdot)\right]$ so

$$\langle f \otimes g, \mathfrak{C}_{KCI} w \rangle = \mathbb{E}\left[w(Z)\, \mathbb{E}\left[(f(X) - \mathbb{E}\left[f(X) \mid Z\right])(g(Y) - \mathbb{E}\left[g(Y) \mid Z\right]) \mid Z\right]\right]$$

KCI: Kernel-based Conditional Independence (Zhang et al., UAI 2012; us)

- Want to check $w(Z)$-weighted covariance of $f(X) - \mathbb{E}\left[f(X) \mid Z\right]$ and $g(Y) - \mathbb{E}\left[g(Y) \mid Z\right]$
- The **conditional mean embedding** $\mu_{X|Z}(z) = \mathbb{E}\left[k(X, \cdot) \mid Z = z\right]$ gives us

$$\langle \mu_{X|Z}(z), f \rangle = \mathbb{E}[f(X) \mid Z = z] \quad \text{for any } f \text{ in the RKHS for } X$$

  - For $k(X, X') = X \cdot X'$, this is $\mu_{X|Z}(z) = \mathbb{E}[X \mid Z = z]$
- The **conditional cross-covariance operator**,

$$\mathfrak{C}_{XY|Z}(z) = \mathbb{E}\left[\left(k(X, \cdot) - \mu_{X|Z}(z)\right) \otimes \left(k(Y, \cdot) - \mu_{Y|Z}(z)\right) \mid Z = z\right]$$

  lets us evaluate $\langle f, \mathfrak{C}_{XY|Z=z} g \rangle = \mathrm{Cov}(f(X), g(Y) \mid Z = z)$
- The **KCI operator** aggregates this over Z: $\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z}(Z) \otimes k(Z, \cdot)\right]$ so

$$\langle f \otimes g, \mathfrak{C}_{KCI} w \rangle = \mathbb{E}\left[w(Z)\, \mathbb{E}\left[(f(X) - \mathbb{E}\left[f(X) \mid Z\right])(g(Y) - \mathbb{E}\left[g(Y) \mid Z\right]) \mid Z\right]\right]$$

- Means that, if the kernels are "rich enough," $\mathrm{KCI} = \|\mathfrak{C}_{KCI}\|^2_{\mathrm{HS}}$ is zero iff $X \perp\!\!\!\perp Y \mid Z$

Estimating KCI

- Define centred kernel $k^{\mu}\big((X,Z),(X',Z')\big) = \langle k(X,\cdot) - \mu_{X|Z}(Z), k(X',\cdot) - \mu_{X|Z}(Z)\rangle$
- Then we can compute that

$$\|\mathfrak{C}_{KCI}\|^2_{\mathrm{HS}} = \mathbb{E}\left[k(Z,Z')\; k^{\mu}\big((X,Z),(X',Z')\big)\; k^{\mu}\big((Y,Z),(Y',Z')\big)\right]$$

Estimating KCI

- Define centred kernel $k^{\mu}\big((X,Z),(X',Z')\big) = \langle k(X,\cdot) - \mu_{X|Z}(Z), k(X',\cdot) - \mu_{X|Z}(Z)\rangle$
- Then we can compute that

$$\|\mathfrak{C}_{KCI}\|^2_{\mathrm{HS}} = \mathbb{E}\left[k(Z,Z')\; k^{\mu}\big((X,Z),(X',Z')\big)\; k^{\mu}\big((Y,Z),(Y',Z')\big)\right]$$

- Gives a natural U-statistic estimator

$$\mathrm{KCI}_n = \frac{1}{n(n-1)} \sum_{i \neq j} k(Z,Z')\; k^{\mu}\big((x_i,z_i),(x_j,z_j)\big)\; k^{\mu}\big((y_i,z_i),(x_j,z_j)\big)$$

Estimating KCI

- Define centred kernel $k^\mu\big((X,Z),(X',Z')\big) = \langle k(X,\cdot) - \mu_{X|Z}(Z), k(X',\cdot) - \mu_{X|Z}(Z)\rangle$
- Then we can compute that

$$\|\mathfrak{C}_{KCI}\|^2_{\mathrm{HS}} = \mathbb{E}\left[k(Z,Z')\; k^\mu\big((X,Z),(X',Z')\big)\; k^\mu\big((Y,Z),(Y',Z')\big)\right]$$

- Gives a natural U-statistic estimator

$$\mathrm{KCI}_n = \frac{1}{n(n-1)} \sum_{i \neq j} k(Z,Z')\; k^\mu\big((x_i,z_i),(x_j,z_j)\big)\; k^\mu\big((y_i,z_i),(x_j,z_j)\big)$$

- Only problem is . . . we don't know $\mu_{X|Z}(z) = \mathbb{E}[k(X,\cdot) \mid Z = z]$ or $\mu_{Y|Z}$!

Estimating KCI

- Define centred kernel $k^\mu\big((X,Z),(X',Z')\big) = \langle k(X,\cdot) - \mu_{X|Z}(Z), k(X',\cdot) - \mu_{X|Z}(Z)\rangle$
- Then we can compute that

$$\|\mathfrak{C}_{KCI}\|^2_{\mathrm{HS}} = \mathbb{E}\left[k(Z,Z')\; k^\mu\big((X,Z),(X',Z')\big)\; k^\mu\big((Y,Z),(Y',Z')\big)\right]$$

- Gives a natural U-statistic estimator

$$\mathrm{KCI}_n = \frac{1}{n(n-1)} \sum_{i \neq j} k(Z,Z')\; k^\mu\big((x_i,z_i),(x_j,z_j)\big)\; k^\mu\big((y_i,z_i),(x_j,z_j)\big)$$

- Only problem is . . . we don't know $\mu_{X|Z}(z) = \mathbb{E}[k(X,\cdot) \mid Z = z]$ or $\mu_{Y|Z}$!
- We can **estimate** them from datasets $\{(x_i,z_i)\}$ and $\{(y_i,z_i)\}$. Two major options:
  - Kernel ridge regression: inputs Z, RKHS-valued outputs $k(X,\cdot)$ or $k(Y,\cdot)$

Estimating KCI

- Define centred kernel $k^{\mu}\big((X,Z),(X',Z')\big) = \langle k(X,\cdot) - \mu_{X|Z}(Z), k(X',\cdot) - \mu_{X|Z}(Z)\rangle$
- Then we can compute that

$$\|\mathfrak{C}_{KCI}\|^2_{\mathrm{HS}} = \mathbb{E}\left[k(Z,Z')\; k^{\mu}\big((X,Z),(X',Z')\big)\; k^{\mu}\big((Y,Z),(Y',Z')\big)\right]$$

- Gives a natural U-statistic estimator

$$\mathrm{KCI}_n = \frac{1}{n(n-1)} \sum_{i \neq j} k(Z,Z')\; k^{\mu}\big((x_i,z_i),(x_j,z_j)\big)\; k^{\mu}\big((y_i,z_i),(x_j,z_j)\big)$$

- Only problem is . . . we don't know $\mu_{X|Z}(z) = \mathbb{E}[k(X,\cdot) \mid Z = z]$ or $\mu_{Y|Z}$!
- We can **estimate** them from datasets $\{(x_i,z_i)\}$ and $\{(y_i,z_i)\}$. Two major options:
  - Kernel ridge regression: inputs Z, RKHS-valued outputs $k(X,\cdot)$ or $k(Y,\cdot)$
  - Fit a **conditional generative model** for X given Z, then $\hat{\mu}_{X|Z}(z) = \widehat{\mathbb{E}}[k(X,\cdot) \mid Z = z]$

Estimating KCI

- Define centred kernel $k^{\mu}\big((X,Z),(X',Z')\big) = \langle k(X,\cdot) - \mu_{X|Z}(Z), k(X',\cdot) - \mu_{X|Z}(Z)\rangle$
- Then we can compute that

$$\|\mathfrak{C}_{KCI}\|^2_{\mathrm{HS}} = \mathbb{E}\left[k(Z,Z')\; k^{\mu}\big((X,Z),(X',Z')\big)\; k^{\mu}\big((Y,Z),(Y',Z')\big)\right]$$

- Gives a natural U-statistic estimator

$$\mathrm{KCI}_n = \frac{1}{n(n-1)} \sum_{i \neq j} k(Z,Z')\; k^{\mu}\big((x_i,z_i),(x_j,z_j)\big)\; k^{\mu}\big((y_i,z_i),(x_j,z_j)\big)$$

- Only problem is . . . we don't know $\mu_{X|Z}(z) = \mathbb{E}[k(X,\cdot) \mid Z = z]$ or $\mu_{Y|Z}$!
- We can **estimate** them from datasets $\{(x_i,z_i)\}$ and $\{(y_i,z_i)\}$. Two major options:
  - Kernel ridge regression: inputs Z, RKHS-valued outputs $k(X,\cdot)$ or $k(Y,\cdot)$
  - Fit a **conditional generative model** for X given Z, then $\hat{\mu}_{X|Z}(z) = \widehat{\mathbb{E}}[k(X,\cdot) \mid Z = z]$
- Get $k^{\hat{\mu}}\big((X,Z),(X',Z')\big) = \langle k(X,\cdot) - \hat{\mu}_{X|Z}(Z), k(X',\cdot) - \hat{\mu}_{X|Z}(Z)\rangle$ and

$$\widehat{\mathrm{KCI}}_n = \frac{1}{n(n-1)} \sum_{i \neq j} k(Z,Z')\; k^{\hat{\mu}}\big((x_i,z_i),(x_j,z_j)\big)\; k^{\hat{\mu}}\big((y_i,z_i),(x_j,z_j)\big)$$

Testing

- Our initial goal was to tell whether $X \perp\!\!\!\perp Y \mid Z$ based on samples
- We'll take a traditional **null hypothesis significance testing** approach
- $\mathfrak{H}_0 : X \perp\!\!\!\perp Y \mid Z$; alternative hypothesis is "not that"
- Assuming good-enough kernels, equivalent to ask whether $\mathrm{KCI} = 0$

Testing

- Our initial goal was to tell whether $X \perp\!\!\!\perp Y \mid Z$ based on samples
- We'll take a traditional **null hypothesis significance testing** approach
- $\mathfrak{H}_0 : X \perp\!\!\!\perp Y \mid Z$; alternative hypothesis is "not that"
- Assuming good-enough kernels, equivalent to ask whether $\mathrm{KCI} = 0$
- If we know the mean embeddings ("model-X" (and Y) regime), we can estimate KCI_n and reject $\mathfrak{H}_0$ if it's big enough

Testing

- Our initial goal was to tell whether $X \perp\!\!\!\perp Y \mid Z$ based on samples
- We'll take a traditional **null hypothesis significance testing** approach
- $\mathfrak{H}_0 : X \perp\!\!\!\perp Y \mid Z$; alternative hypothesis is "not that"
- Assuming good-enough kernels, equivalent to ask whether $\mathrm{KCI} = 0$
- If we know the mean embeddings ("model-X" (and Y) regime), we can estimate KCI_n and reject $\mathfrak{H}_0$ if it's big enough
- Can estimate how big "big enough" is based on estimating asymptotic distribution or (better) wild bootstrap; threshold will be $\Theta(1/n)$
  - Guaranteed conservative test: set $\Theta(1/\sqrt{n})$ threshold with Hoeffding for U-statistics

Testing

- Our initial goal was to tell whether $X \perp\!\!\!\perp Y \mid Z$ based on samples
- We'll take a traditional **null hypothesis significance testing** approach
- $\mathfrak{H}_0 : X \perp\!\!\!\perp Y \mid Z$; alternative hypothesis is "not that"
- Assuming good-enough kernels, equivalent to ask whether $\mathrm{KCI} = 0$
- If we know the mean embeddings ("model-X" (and Y) regime), we can estimate KCI_n and reject $\mathfrak{H}_0$ if it's big enough
- Can estimate how big "big enough" is based on estimating asymptotic distribution or (better) wild bootstrap; threshold will be $\Theta(1/n)$
  - Guaranteed conservative test: set $\Theta(1/\sqrt{n})$ threshold with Hoeffding for U-statistics
- In practice: get our best guess at conditional mean embeddings, do above with $\widehat{\mathrm{KCI}}_n$

Testing

- Our initial goal was to tell whether $X \perp\!\!\!\perp Y \mid Z$ based on samples
- We'll take a traditional **null hypothesis significance testing** approach
- $\mathfrak{H}_0 : X \perp\!\!\!\perp Y \mid Z$; alternative hypothesis is "not that"
- Assuming good-enough kernels, equivalent to ask whether $\mathrm{KCI} = 0$
- If we know the mean embeddings ("model-X" (and Y) regime), we can estimate KCI_n and reject $\mathfrak{H}_0$ if it's big enough
- Can estimate how big "big enough" is based on estimating asymptotic distribution or (better) wild bootstrap; threshold will be $\Theta(1/n)$
  - Guaranteed conservative test: set $\Theta(1/\sqrt{n})$ threshold with Hoeffding for U-statistics
- In practice: get our best guess at conditional mean embeddings, do above with $\widehat{\mathrm{KCI}}_n$
- . . . will this work?

Conditional independence testing is impossible (Shah and Peters, Annals 2020)

- Let $X, Y, Z \sim \mathrm{Unif}([0,1])$ be independent
- Let $Z_{100} \in \{0, 1, \ldots, 9\}$ be the 100th decimal digit of Z
- Let X' be X but with its first decimal digit replaced by Z_{100}
- Let Y' be Y but with its first decimal digit replaced by Z_{100}

Conditional independence testing is impossible (Shah and Peters, Annals 2020)

- Let $X, Y, Z \sim \mathrm{Unif}([0,1])$ be independent
- Let $Z_{100} \in \{0, 1, \ldots, 9\}$ be the 100th decimal digit of Z
- Let X' be X but with its first decimal digit replaced by Z_{100}
- Let Y' be Y but with its first decimal digit replaced by Z_{100}

- X' and Y' are strongly dependent
- Z looks totally irrelevant, unless you know to check its 100th decimal place
- But actually $X' \perp\!\!\!\perp Y' \mid Z$

Conditional independence testing is impossible (Shah and Peters, Annals 2020)

- Let $X, Y, Z \sim \mathrm{Unif}([0,1])$ be independent
- Let $Z_{100} \in \{0, 1, \ldots, 9\}$ be the 100th decimal digit of Z
- Let X' be X but with its first decimal digit replaced by Z_{100}
- Let Y' be Y but with its first decimal digit replaced by Z_{100}

- X' and Y' are strongly dependent
- Z looks totally irrelevant, unless you know to check its 100th decimal place
- But actually $X' \perp\!\!\!\perp Y' \mid Z$

- Take independent $Z' \sim \mathrm{Unif}([0,1])$; X' and Y' are strongly dependent given Z'

Conditional independence testing is impossible (Shah and Peters, Annals 2020)

- Let $X, Y, Z \sim \mathrm{Unif}([0,1])$ be independent
- Let $Z_{100} \in \{0, 1, \ldots, 9\}$ be the 100th decimal digit of Z
- Let X' be X but with its first decimal digit replaced by Z_{100}
- Let Y' be Y but with its first decimal digit replaced by Z_{100}

- X' and Y' are strongly dependent
- Z looks totally irrelevant, unless you know to check its 100th decimal place
- But actually $X' \perp\!\!\!\perp Y' \mid Z$

- Take independent $Z' \sim \mathrm{Unif}([0,1])$; X' and Y' are strongly dependent given Z'

- "Reasonable" tests can only distinguish (X', Y', Z) from (X', Y', Z') with ridiculous numbers of samples; if we want guaranteed Type I error, we can't meaningfully reject

Conditional independence testing is impossible (Shah and Peters, Annals 2020)

- Let $X, Y, Z \sim \mathrm{Unif}([0,1])$ be independent
- Let $Z_{100} \in \{0, 1, \ldots, 9\}$ be the 100th decimal digit of Z
- Let X' be X but with its first decimal digit replaced by Z_{100}
- Let Y' be Y but with its first decimal digit replaced by Z_{100}

- X' and Y' are strongly dependent
- Z looks totally irrelevant, unless you know to check its 100th decimal place
- But actually $X' \perp\!\!\!\perp Y' \mid Z$

- Take independent $Z' \sim \mathrm{Unif}([0,1])$; X' and Y' are strongly dependent given Z'

- "Reasonable" tests can only distinguish (X', Y', Z) from (X', Y', Z') with ridiculous numbers of samples; if we want guaranteed Type I error, we can't meaningfully reject

- Theorem: for any test on any conditionally-dependent continuous distribution, there's some possible independent distribution indistinguishable to the test

What really happens, though?

- The Shah and Peters construction "feels like a trick":
  real distributions don't hide information in lower-order bits
- But the way that $\widehat{\mathrm{KCI}}_n$ fails is in estimating $\mu_{X|Z}$, $\mu_{Y|Z}$
- That *absolutely* is the way that things really fail

What really happens, though?

- The Shah and Peters construction "feels like a trick":
  real distributions don't hide information in lower-order bits
- But the way that $\widehat{\mathrm{KCI}}_n$ fails is in estimating $\mu_{X|Z}$, $\mu_{Y|Z}$
- That *absolutely* is the way that things really fail
  - Proof: we could have an exactly conservative test based on Hoeffding if we knew μs

What really happens, though?

- The Shah and Peters construction "feels like a trick":
  real distributions don't hide information in lower-order bits
- But the way that $\widehat{\mathrm{KCI}}_n$ fails is in estimating $\mu_{X|Z}$, $\mu_{Y|Z}$
- That *absolutely* is the way that things really fail
  - Proof: we could have an exactly conservative test based on Hoeffding if we knew μs
  - Increasingly large literature on model-X is the same
    - If I can model X, I can estimate $\mu_{X|Z}$ arbitrarily accurately

What really happens, though?

- The Shah and Peters construction "feels like a trick":
  real distributions don't hide information in lower-order bits
- But the way that $\widehat{\mathrm{KCI}}_n$ fails is in estimating $\mu_{X|Z}$, $\mu_{Y|Z}$
- That *absolutely* is the way that things really fail
  - Proof: we could have an exactly conservative test based on Hoeffding if we knew μs
  - Increasingly large literature on model-X is the same
    - If I can model X, I can estimate $\mu_{X|Z}$ arbitrarily accurately

- With "rich" kernels, the *best-case* minimax rate for μ estimation is $\mathcal{O}\left(1/m^{1/4}\right)$;
  on harder problems the minimax rate can be *arbitrarily slow* (Li et al., JMLR 2022)

What really happens, though?

- The Shah and Peters construction "feels like a trick":
  real distributions don't hide information in lower-order bits
- But the way that $\widehat{\mathrm{KCI}}_n$ fails is in estimating $\mu_{X|Z}$, $\mu_{Y|Z}$
- That *absolutely* is the way that things really fail
  - Proof: we could have an exactly conservative test based on Hoeffding if we knew μs
  - Increasingly large literature on model-X is the same
    - If I can model X, I can estimate $\mu_{X|Z}$ arbitrarily accurately

- With "rich" kernels, the *best-case* minimax rate for μ estimation is $\mathcal{O}\left(1/m^{1/4}\right)$;
  on harder problems the minimax rate can be *arbitrarily slow* (Li et al., JMLR 2022)
- Conditional density estimation can also be arbitrarily hard

What really happens, though?

- The Shah and Peters construction "feels like a trick":
  real distributions don't hide information in lower-order bits
- But the way that $\widehat{\mathrm{KCI}}_n$ fails is in estimating $\mu_{X|Z}$, $\mu_{Y|Z}$
- That *absolutely* is the way that things really fail
  - Proof: we could have an exactly conservative test based on Hoeffding if we knew μs
  - Increasingly large literature on model-X is the same
    - If I can model X, I can estimate $\mu_{X|Z}$ arbitrarily accurately

- With "rich" kernels, the *best-case* minimax rate for μ estimation is $\mathcal{O}\left(1/m^{1/4}\right)$;
  on harder problems the minimax rate can be *arbitrarily slow* (Li et al., JMLR 2022)
- Conditional density estimation can also be arbitrarily hard
  - Actually, true even for unconditional density estimation, and there's no simple characterization of distributions where it can be done (Lechner and Ben-David, COLT 2024)

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y
- Requires regression estimates of $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y
- Requires regression estimates of $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$
- Studentized estimate for easier null distribution

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y
- Requires regression estimates of $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$
- Studentized estimate for easier null distribution
- For multivariate X, Y, takes biggest of (X_i, Y_j) dimension pairs

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y
- Requires regression estimates of $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$
- Studentized estimate for easier null distribution
- For multivariate X, Y, takes biggest of (X_i, Y_j) dimension pairs
- A "nonlinear partial covariance" – cannot handle every dependence type

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y
- Requires regression estimates of $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$
- Studentized estimate for easier null distribution
- For multivariate X, Y, takes biggest of (X_i, Y_j) dimension pairs
- A "nonlinear partial covariance" – cannot handle every dependence type
- Weighted GCM (Scheidegger et al., JMLR 2022): $\mathbb{E}_Z\left[w(Z)\,\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y
- Requires regression estimates of $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$
- Studentized estimate for easier null distribution
- For multivariate X, Y, takes biggest of (X_i, Y_j) dimension pairs
- A "nonlinear partial covariance" – cannot handle every dependence type

- Weighted GCM (Scheidegger et al., JMLR 2022): $\mathbb{E}_Z\left[w(Z)\,\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$
- One w based on estimating $w(z) = \mathrm{sign}(\mathrm{Cov}(X, Y \mid Z = z))$

Generalized Covariance Measure (GCM)

- Shah and Peters say: since we can't hope to do it in general, let's relax our aims
- GCM estimates $\mathbb{E}_Z\left[\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$ for scalar X, Y
- Requires regression estimates of $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$
- Studentized estimate for easier null distribution
- For multivariate X, Y, takes biggest of (X_i, Y_j) dimension pairs
- A "nonlinear partial covariance" – cannot handle every dependence type

- Weighted GCM (Scheidegger et al., JMLR 2022): $\mathbb{E}_Z\left[w(Z)\,\mathrm{Cov}(X, Y \mid Z) \mid Z\right]$
- One w based on estimating $w(z) = \mathrm{sign}(\mathrm{Cov}(X, Y \mid Z = z))$

- These are (basically) just KCI with linear X, Y kernels!
- For GCM, $k(z, z') = 1$; for wGCM, it's $w(z)w(z')$

Choosing w / $k(Z, Z')$

- KCI-type measures: five kernel/representation choices
- $k(X, X')$, $k(Y, Y')$: GCM sets them as linear, KCI usually Gaussian by default
  - Controls "how" you look at the data
  - Clearly important, generally not obvious how to pick

Choosing w / $k(Z, Z')$

- KCI-type measures: five kernel/representation choices
- $k(X, X')$, $k(Y, Y')$: GCM sets them as linear, KCI usually Gaussian by default
  - Controls "how" you look at the data
  - Clearly important, generally not obvious how to pick
- Conditional mean embedding input kernels, $k_{Z \to X}$ and $k_{Z \to Y}$
  - These are just regressions, can choose parameters to minimize leave-one-out error

Choosing w / $k(Z, Z')$

- KCI-type measures: five kernel/representation choices
- $k(X, X')$, $k(Y, Y')$: GCM sets them as linear, KCI usually Gaussian by default
  - Controls "how" you look at the data
  - Clearly important, generally not obvious how to pick
- Conditional mean embedding input kernels, $k_{Z \to X}$ and $k_{Z \to Y}$
  - These are just regressions, can choose parameters to minimize leave-one-out error
- $k(Z, Z')$: GCM uses 1, wGCM has schemes to estimate w in $w(z)w(z')$

Choosing w / $k(Z, Z')$

- KCI-type measures: five kernel/representation choices
- $k(X, X')$, $k(Y, Y')$: GCM sets them as linear, KCI usually Gaussian by default
  - Controls "how" you look at the data
  - Clearly important, generally not obvious how to pick
- Conditional mean embedding input kernels, $k_{Z \to X}$ and $k_{Z \to Y}$
  - These are just regressions, can choose parameters to minimize leave-one-out error
- $k(Z, Z')$: GCM uses 1, wGCM has schemes to estimate w in $w(z)w(z')$
  - We previously just reused $k_{Z \to X}$

Choosing w / $k(Z, Z')$

- KCI-type measures: five kernel/representation choices
- $k(X, X')$, $k(Y, Y')$: GCM sets them as linear, KCI usually Gaussian by default
  - Controls "how" you look at the data
  - Clearly important, generally not obvious how to pick
- Conditional mean embedding input kernels, $k_{Z \to X}$ and $k_{Z \to Y}$
  - These are just regressions, can choose parameters to minimize leave-one-out error
- $k(Z, Z')$: GCM uses 1, wGCM has schemes to estimate w in $w(z)w(z')$
  - We previously just reused $k_{Z \to X}$
  - Actually, a good choice of $k(Z, Z')$ is very important!

Importance of $k(Z, Z')$

- A toy model:

$$Z \sim \mathcal{N}(0,1), \quad X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma(Z) \\ \gamma(Z) & 1 \end{bmatrix} \right), \quad \gamma(Z) = \sin(\beta Z)$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$

Importance of $k(Z, Z')$

- A toy model:

$$Z \sim \mathcal{N}(0,1), \quad X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma(Z) \\ \gamma(Z) & 1 \end{bmatrix} \right), \quad \gamma(Z) = \sin(\beta Z)$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$
- Intuitively, to estimate $X \mid Y$ we need a kernel with lengthscale around that of f_X

Importance of $k(Z, Z')$

- A toy model:

$$Z \sim \mathcal{N}(0,1), \quad X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

  where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma(Z) \\ \gamma(Z) & 1 \end{bmatrix} \right), \quad \gamma(Z) = \sin(\beta Z)$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$
- Intuitively, to estimate $X \mid Y$ we need a kernel with lengthscale around that of f_X
- ... and to "see" dependence in the residuals, we need a lengthscale around β

Importance of $k(Z, Z')$

- A toy model:

$$Z \sim \mathcal{N}(0,1), \quad X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma(Z) \\ \gamma(Z) & 1 \end{bmatrix} \right), \quad \gamma(Z) = \sin(\beta Z)$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$
- Intuitively, to estimate $X \mid Y$ we need a kernel with lengthscale around that of f_X
- ... and to "see" dependence in the residuals, we need a lengthscale around β
- If we take linear kernels on X and Y, Gaussian on Z with lengthscale ℓ, perfect μs:

$$\mathfrak{C}_{\mathrm{KCI}} = \tau^2 \mathop{\mathbb{E}}_{Z}[k(Y, \cdot)\, \mathbb{E}\left[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z])\right]]$$

Importance of $k(Z, Z')$

- A toy model:

$$Z \sim \mathcal{N}(0,1), \quad X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma(Z) \\ \gamma(Z) & 1 \end{bmatrix} \right), \quad \gamma(Z) = \sin(\beta Z)$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$
- Intuitively, to estimate $X \mid Y$ we need a kernel with lengthscale around that of f_X
- ... and to "see" dependence in the residuals, we need a lengthscale around β
- If we take linear kernels on X and Y, Gaussian on Z with lengthscale ℓ, perfect μs:

$$\mathfrak{C}_{\mathrm{KCI}} = \tau^2 \mathop{\mathbb{E}}_{Z}[k(Y,\cdot)\, \mathbb{E}\left[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z])\right]] = \tau^2 \mathop{\mathbb{E}}_{Z}[k(Y,\cdot)\gamma(Z)]$$

Importance of $k(Z, Z')$

- A toy model:

$$Z \sim \mathcal{N}(0,1), \quad X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma(Z) \\ \gamma(Z) & 1 \end{bmatrix} \right), \quad \gamma(Z) = \sin(\beta Z)$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$
- Intuitively, to estimate $X \mid Y$ we need a kernel with lengthscale around that of f_X
- ... and to "see" dependence in the residuals, we need a lengthscale around β
- If we take linear kernels on X and Y, Gaussian on Z with lengthscale ℓ, perfect μs:

$$\mathfrak{C}_{\mathrm{KCI}} = \tau^2 \mathop{\mathbb{E}}_{Z}[k(Y, \cdot)\, \mathbb{E}\left[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z])\right]] = \tau^2 \mathop{\mathbb{E}}_{Z}[k(Y, \cdot)\gamma(Z)]$$

$$\mathrm{KCI} = \tau^4 \mathop{\mathbb{E}}_{Z, Z'}[k(Z, Z')\gamma(Z)\gamma(Z')]$$

Importance of $k(Z, Z')$

- A toy model:

$$Z \sim \mathcal{N}(0,1), \quad X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma(Z) \\ \gamma(Z) & 1 \end{bmatrix} \right), \quad \gamma(Z) = \sin(\beta Z)$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$
- Intuitively, to estimate $X \mid Y$ we need a kernel with lengthscale around that of f_X
- ... and to "see" dependence in the residuals, we need a lengthscale around β
- If we take linear kernels on X and Y, Gaussian on Z with lengthscale ℓ, perfect μs:

$$\mathfrak{C}_{\mathrm{KCI}} = \tau^2 \mathop{\mathbb{E}}_{Z}[k(Y,\cdot)\, \mathbb{E}\left[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z])\right]] = \tau^2 \mathop{\mathbb{E}}_{Z}[k(Y,\cdot)\gamma(Z)]$$

$$\mathrm{KCI} = \tau^4 \mathop{\mathbb{E}}_{Z,Z'}[k(Z,Z')\gamma(Z)\gamma(Z')] = \frac{1}{2}\tau^4 \sqrt{\frac{\ell^2}{\ell^2+2}} e^{-\beta^2} \left(e^{2\beta^2/(\ell^2+2)} - 1 \right)$$

Bias

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

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

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

Bias

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

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

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

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

Bias

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

$$\begin{aligned}
&\left\| \mathbb{E}\Big[k(X,\cdot) \otimes k(Z,\cdot) \otimes \big(k(Y,\cdot) - \mu_{Y|Z}(Z) - \Delta_{Y|Z}(Z) \big) \Big] \right\|^2 \\
&\qquad = \Big\| \underbrace{\mathbb{E}\left[k(X,\cdot) \otimes k(Z,\cdot) \otimes \big(k(Y,\cdot) - \mu_{Y|Z}(Z) \big) \right]}_{0, \text{ since } X \perp\!\!\!\perp Y \mid Z} - \mathbb{E}\left[k(X,\cdot) \otimes k(Z,\cdot) \otimes \Delta_{Y|Z}(Z) \right] \Big\|^2 \\
&\qquad = \mathbb{E}\Big[k(X,X')\, k(Y,Y') \underbrace{\langle \Delta_{Y|Z}(Z), \Delta_{Y|Z}(Z') \rangle}_{\text{likely big if } k(Y,Y') \text{ is big}} \Big]
\end{aligned}$$

- 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

Bias

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

$$\begin{aligned}
&\left\| \mathbb{E}\Big[k(X,\cdot) \otimes k(Z,\cdot) \otimes \big(k(Y,\cdot) - \mu_{Y|Z}(Z) - \Delta_{Y|Z}(Z) \big) \Big] \right\|^2 \\
&\qquad = \Big\| \underbrace{\mathbb{E}\left[k(X,\cdot) \otimes k(Z,\cdot) \otimes \big(k(Y,\cdot) - \mu_{Y|Z}(Z) \big) \right]}_{0, \text{ since } X \perp\!\!\!\perp Y \mid Z} - \mathbb{E}\left[k(X,\cdot) \otimes k(Z,\cdot) \otimes \Delta_{Y|Z}(Z) \right] \Big\|^2 \\
&\qquad = \mathbb{E}\Big[k(X,X')\, k(Y,Y') \underbrace{\langle \Delta_{Y|Z}(Z), \Delta_{Y|Z}(Z') \rangle}_{\text{likely big if } k(Y,Y') \text{ is big}} \Big]
\end{aligned}$$

- 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

(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., UAI 2012) is

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

(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., UAI 2012) is

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

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

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

(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., UAI 2012) is

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

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

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

- Can reduce this by replacing $\langle \Delta_{X|Z}(X), \Delta_{X|Z}(X')\rangle$ with $\langle \Delta^{(1)}_{X|Z}(X), \Delta^{(2)}_{X|Z}(X')\rangle$
  - Compute by using two different regressions: split the data used to train it

(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., UAI 2012) is

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

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

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

- Can reduce this by replacing $\langle \Delta_{X|Z}(X), \Delta_{X|Z}(X')\rangle$ with $\langle \Delta^{(1)}_{X|Z}(X), \Delta^{(2)}_{X|Z}(X')\rangle$
  - 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

(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., UAI 2012) is

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

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

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

- Can reduce this by replacing $\langle \Delta_{X|Z}(X), \Delta_{X|Z}(X')\rangle$ with $\langle \Delta^{(1)}_{X|Z}(X), \Delta^{(2)}_{X|Z}(X')\rangle$
  - 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

(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., UAI 2012) is

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

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

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

- Can reduce this by replacing $\langle \Delta_{X|Z}(X), \Delta_{X|Z}(X')\rangle$ with $\langle \Delta^{(1)}_{X|Z}(X), \Delta^{(2)}_{X|Z}(X')\rangle$
  - 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)

(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., UAI 2012) is

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

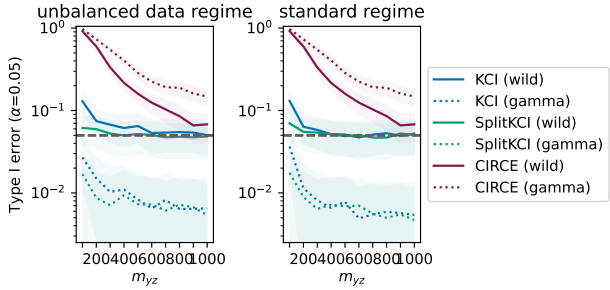
- $C^{\mathrm{KCI}}_{XY|Z} = 0$ iff $X \perp\!\!\!\perp Y \mid Z$; with incorrect regressions, bias becomes

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

- Can reduce this by replacing $\langle \Delta_{X|Z}(X), \Delta_{X|Z}(X')\rangle$ with $\langle \Delta^{(1)}_{X|Z}(X), \Delta^{(2)}_{X|Z}(X')\rangle$
  - 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

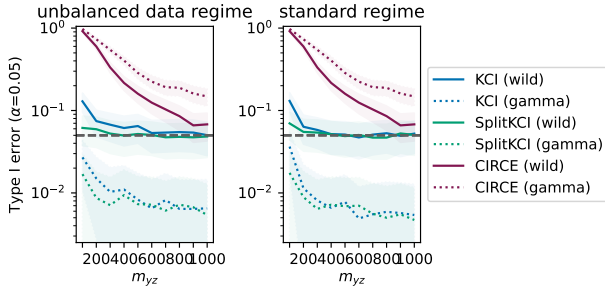
Testing with SplitKCI

- 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



Testing with SplitKCI

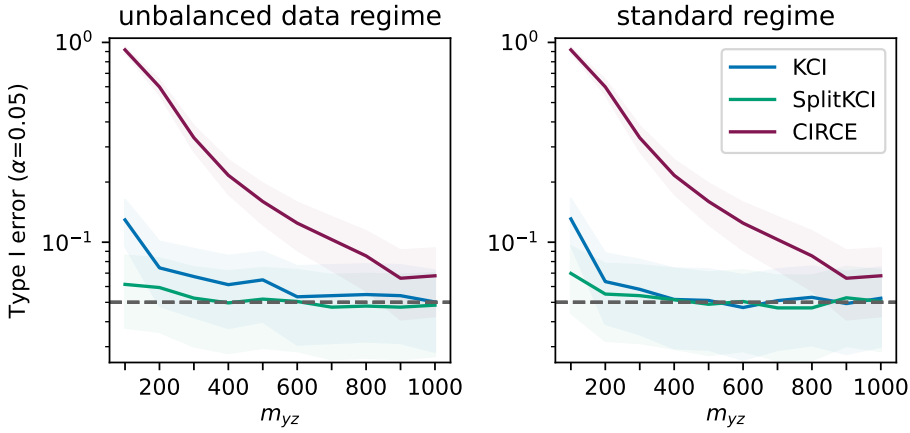
- Zhang et al. (2012) tested based on a gamma approximation to the null distribution
- That approximation can't cope with the bias when mean estimation is poor



- 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

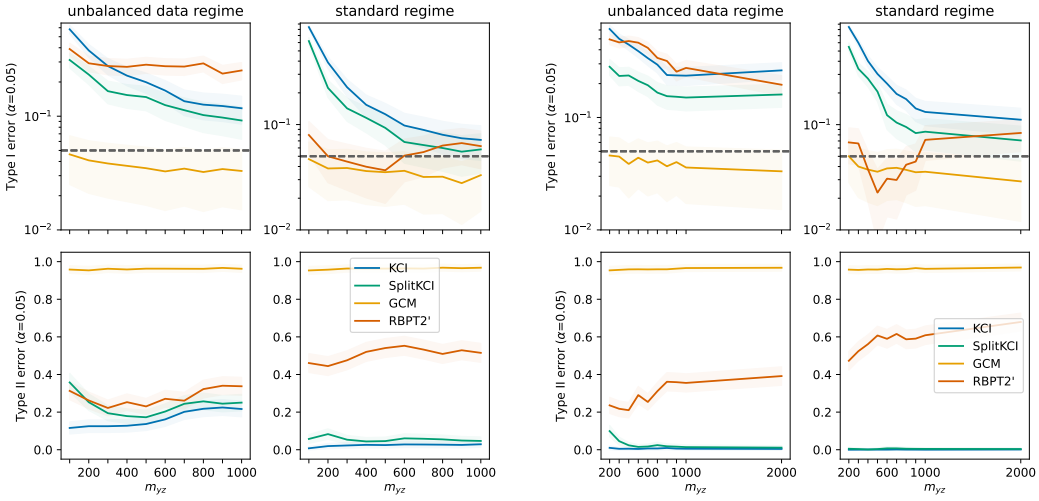
- On synthetic Gaussian data:



- 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 Z (straightforward) and/or Y (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, Y, Z) triples, but lots of (X, Z) and (Y, Z) pairs
  - Wild bootstrap for estimating the test threshold