Causal Information Splitting: Engineering Proxy Features for Robustness to Distribution Shifts

Bijan Mazaheri [1,2] Atalanti Mastakouri [2] Dominik Janzing [2] Michaela Hardt [2]

[1]California Institute of Technology [2]Amazon Causality Lab, Tübingen Germany

POSTER PAPER

Transportability

Distribution Shift

We want to train models to minimize an error function within a testing distribution $(\mathbf{X}_{\text{TEST}}, Y_{\text{TEST}})$.

If $(\mathbf{X}_{\text{TRAIN}}, Y_{\text{TRAIN}}) \sim (\mathbf{X}_{\text{TEST}}, Y_{\text{TEST}})$, we can do *empirical risk minimization*:

$$f = \arg\min_f \mathbb{E}_{\mathbf{x}\sim\mathbf{X}_{\text{TRAIN}}, y\sim Y_{\text{TRAIN}}|\mathbf{X}_{\text{TRAIN}}}[\text{Err}(f(\mathbf{x}), y)] \quad (1)$$

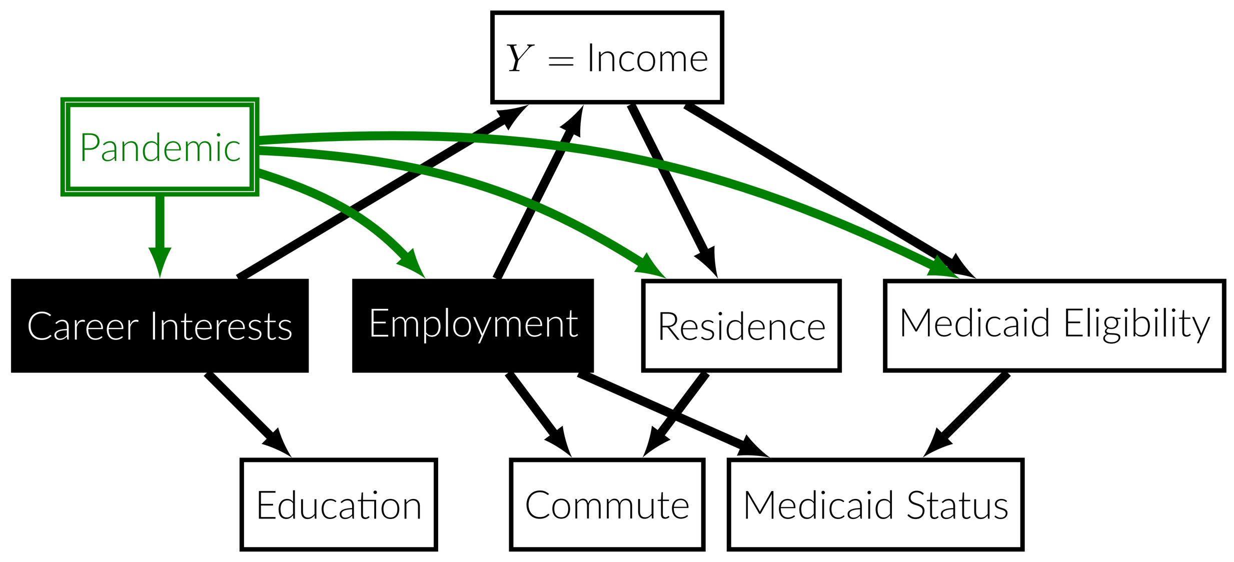
If $(\mathbf{X}_{\text{TRAIN}}, Y_{\text{TRAIN}}) \not\sim (\mathbf{X}_{\text{TEST}}, Y_{\text{TEST}})$, we have *distribution/dataset shift*.

Covariate shift re-weighting techniques require invariance in the label function (Shimodaira, 2000; Sugiyama et al., 2008)

$$\Pr(Y_{\text{TEST}} \mid \mathbf{X}_{\text{TEST}}) = \Pr(Y_{\text{TRAIN}} \mid \mathbf{X}_{\text{TRAIN}}) \quad (2)$$

This is not always true! We can instead search for an *invariant set* with respect to the label function. (Magliacane et al., 2018; Muandet, Balduzzi, and Schölkopf, 2013; Rojas-Carulla et al., 2018)

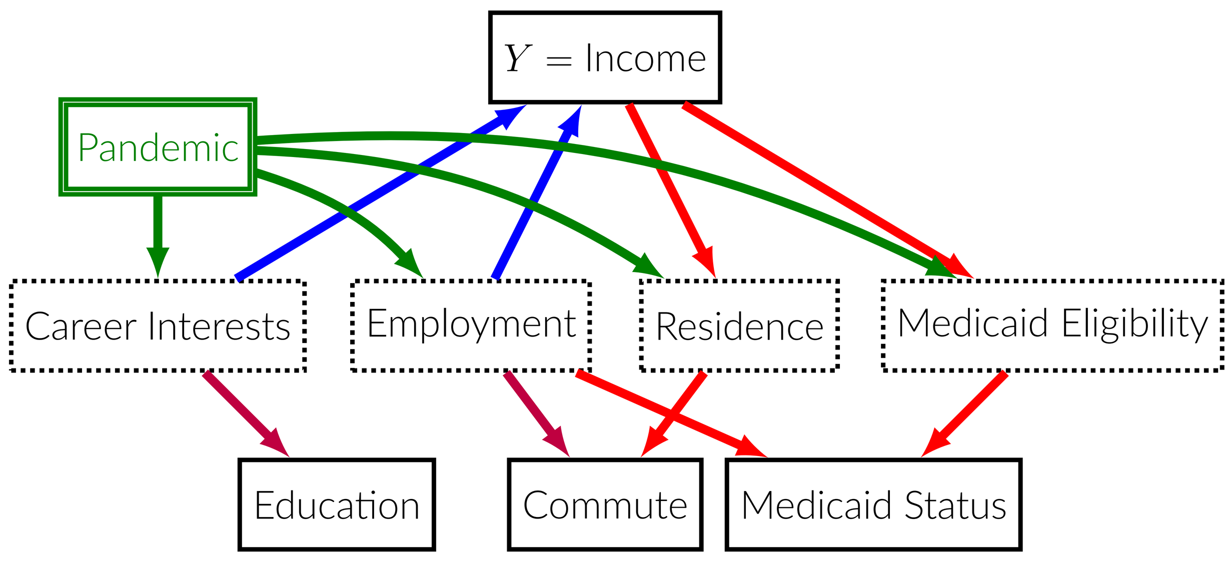
Selection Diagrams (Pearl and Bareinboim, 2011)



$\mathbf{X}$ = {Career Interests, Employment} is an invariant set because it d-separates Pandemic and Income.

No invariant set if Career Interests and Employment are unobserved.

Stable Paths (Subbaswamy and Saria, 2018)

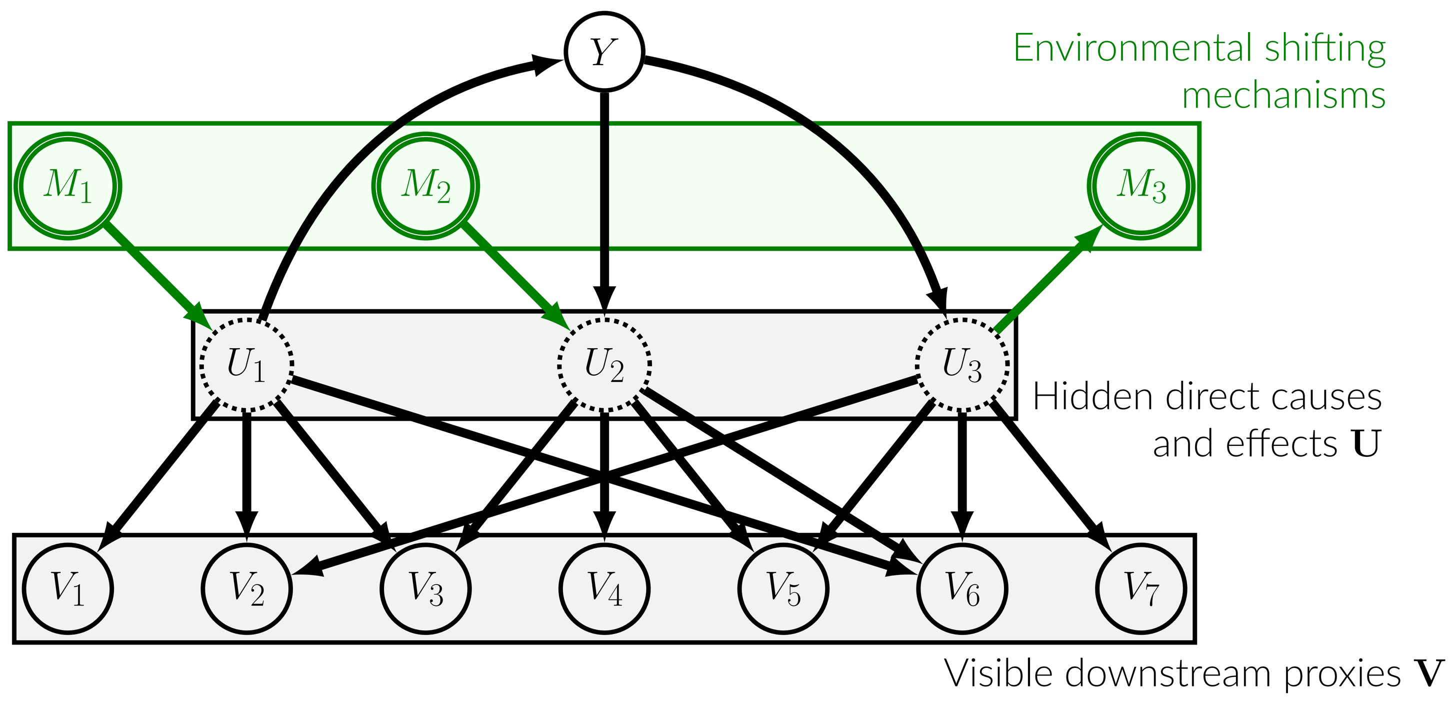


Subbaswamy and Saria, 2018 suggest restricting to stable paths.

Career Interests → Education is not stable (unless Career Interests are included in $\mathbf{X}$). But it still helps!

Proxy-Based Transportability

Problem Setup

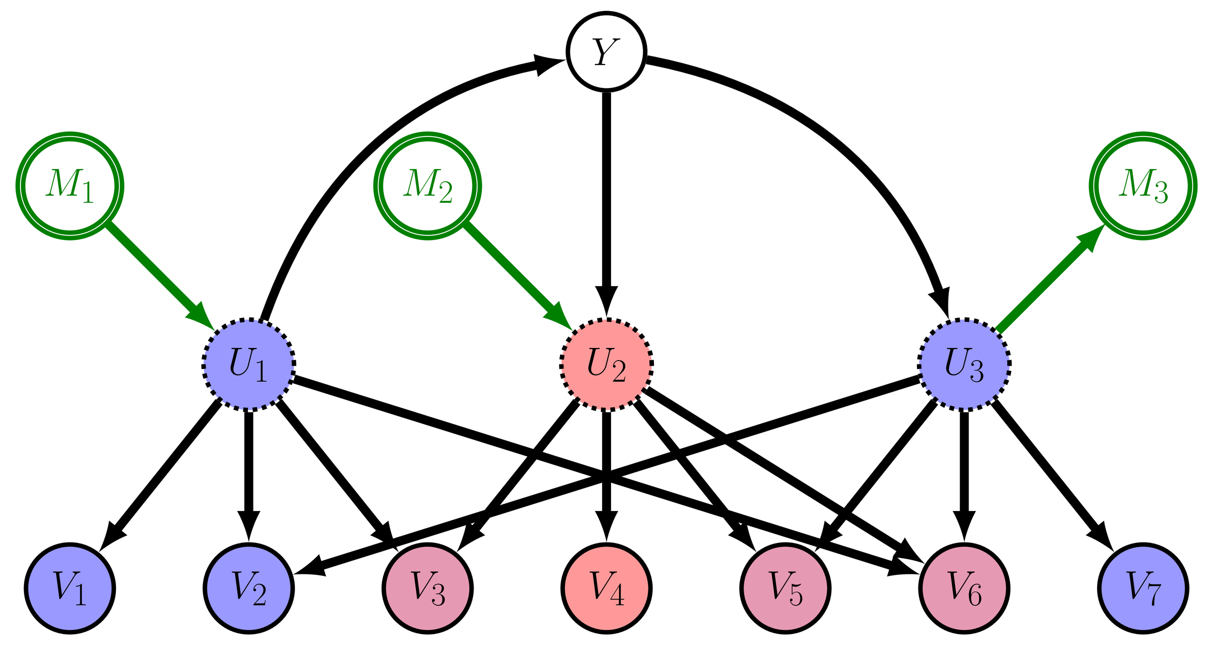


Goals:

We want to find $\mathbf{X}$ that generally:

1) Minimizes a quantitative notion of robustness, called *context sensitivity*: $\mathcal{I}(\mathbf{M} : Y \mid \mathbf{X})$
2) Maximizes predictive potential, called *relevance*: $\mathcal{I}(Y : \mathbf{X})$.

Proxy Classification



Good proxies decrease context sensitivity: $\mathcal{I}(\mathbf{M} : Y \mid \mathbf{X} \cup \{V\}) \leq \mathcal{I}(\mathbf{M} : Y \mid \mathbf{X})$

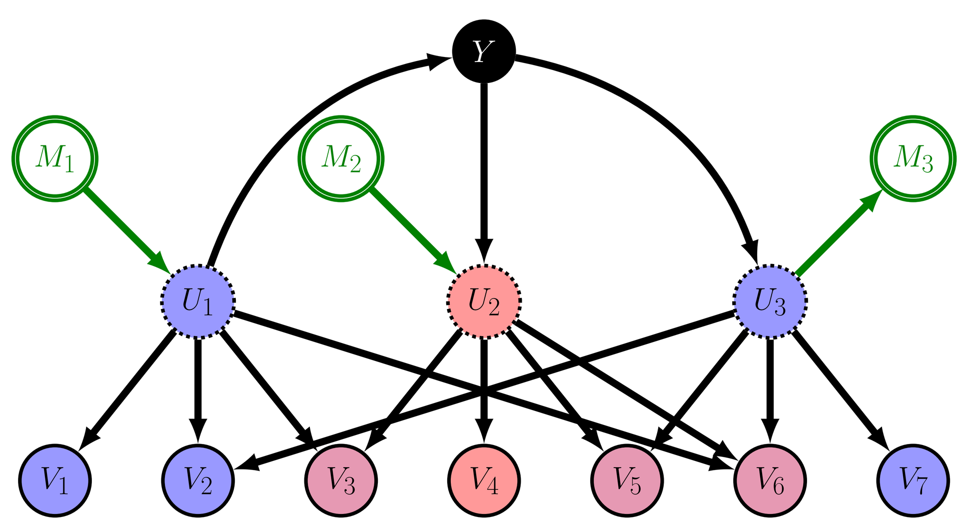
Bad proxies increase context sensitivity: $\mathcal{I}(\mathbf{M} : Y \mid \mathbf{X} \cup \{V\}) > \mathcal{I}(\mathbf{M} : Y \mid \mathbf{X})$

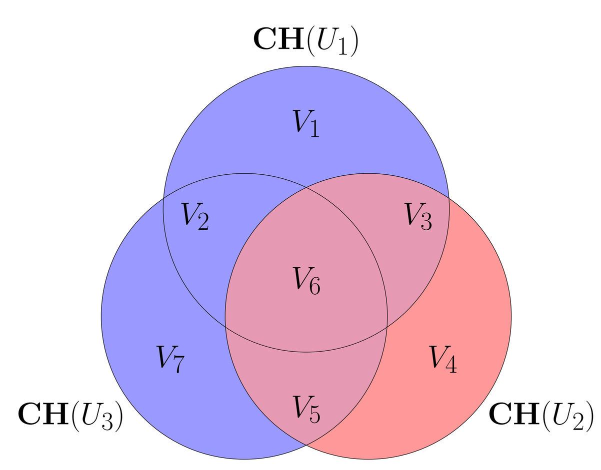
Ambiguous proxies could do either.

We develop *one setting* with both graphical and functional constraints where we have clean definitions for these concepts.

Techniques

Post-selecting on Y

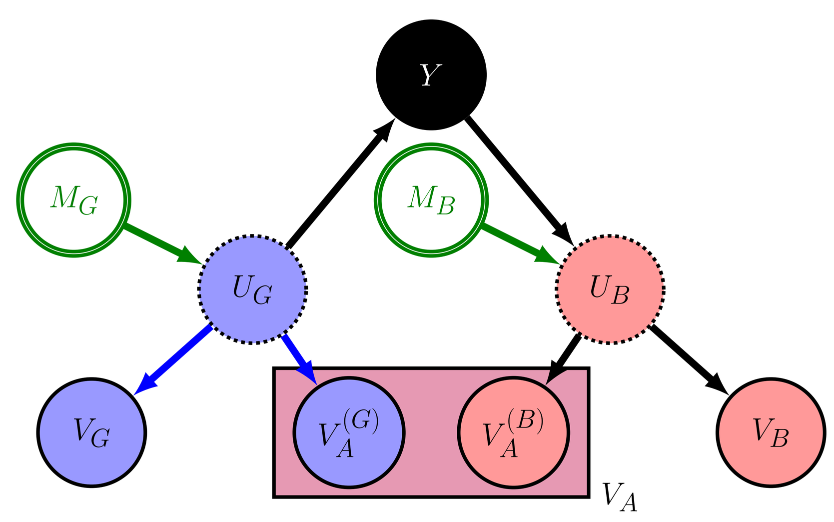




Key Idea: Conditioning on Y d-Separates good proxies and bad proxies.

This allows for *proxy bootstrapping*, which determines good vs bad proxies.

Causal Information Splitting



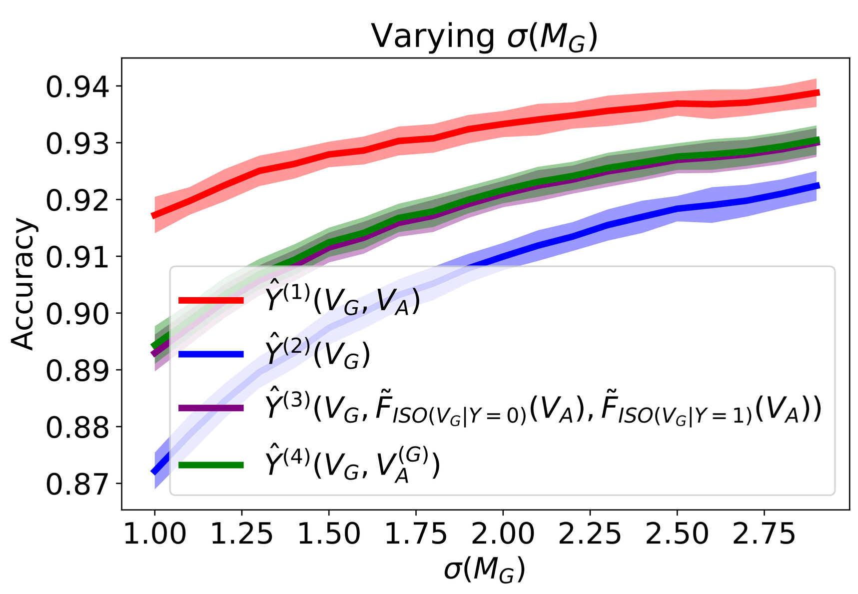
We *want* to use $\mathbf{X} = \{V_G, V_A^{(G)}\}$ but $V_A^{(G)}$ is mixed in as a component of V_A, which is ambiguous.

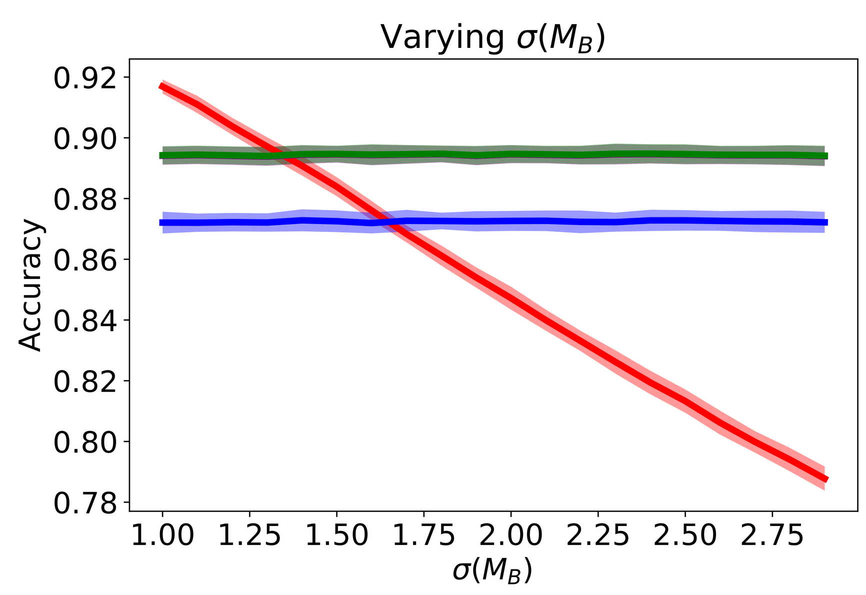
We use *auxiliary training tasks* predicting good proxies using ambiguous proxies.

$\mathbf{X} = \{V_G, \tilde{F}_{\text{ISO}(V_G)}(V_A)\}$ where $\tilde{F}_{\text{ISO}(V_G)}(V_A)$ predicts V_G using V_A under constant Y.

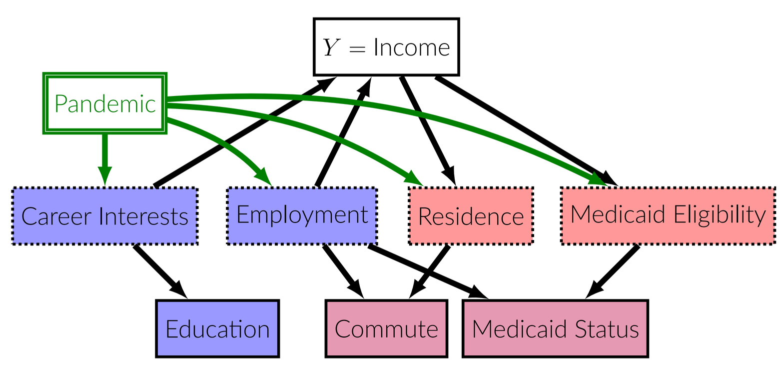
Experimental Results

Synthetic Data





Real World Data



All Features: $\mathbf{X} = \{E, C, MS\}$
Engineered Features: $\mathbf{X} = \{E, \tilde{F}_{\text{ISO}(E)}(C, MS)\}$
Limited Features: $\mathbf{X} = \{E\}$

Table: Comparison of out-of-domain (2021) performance via mean of accuracy.

State	All Features	Engineered Features	Limited Features
CA	**0.712** ± 0.0011	**0.711** ± 0.0014	0.692 ± 0.0014
FL	**0.683** ± 0.0012	0.678 ± 0.0018	0.68 ± 0.0013
GA	0.689 ± 0.0025	**0.707** ± 0.0055	**0.709** ± 0.0029
IL	0.662 ± 0.0026	**0.689** ± 0.0033	0.684 ± 0.0019
NY	**0.707** ± 0.0022	**0.702** ± 0.0025	0.687 ± 0.008
NC	**0.691** ± 0.0031	**0.684** ± 0.0034	**0.683** ± 0.003
OH	0.689 ± 0.0022	**0.703** ± 0.004	**0.696** ± 0.0029
PA	0.672 ± 0.0017	**0.695** ± 0.0023	0.688 ± 0.0022
TX	0.69 ± 0.0029	**0.712** ± 0.0028	**0.712** ± 0.0027
avg	0.688	**0.698**	0.692

References

Magliacane, Sara et al. (2018). ``Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions.'' In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 10869–10879.
Muandet, Krikamol, David Balduzzi, and Bernhard Schölkopf (2013). ``Domain generalization via invariant feature representation.'' In: *International conference on machine learning*. PMLR, pp. 10–18.
Pearl, Judea and Elias Bareinboim (2011). ``Transportability of causal and statistical relations: A formal approach.'' In: *Twenty-fifth AAAI conference on artificial intelligence*.
Rojas-Carulla, Mateo et al. (2018). ``Invariant models for causal transfer learning.'' In: *The Journal of Machine Learning Research* 19.1, pp. 1309–1342.
Shimodaira, Hidetoshi (2000). ``Improving predictive inference under covariate shift by weighting the log-likelihood function.'' In: *Journal of statistical planning and inference* 90.2, pp. 227–244.
Subbaswamy, Adarsh and Suchi Saria (2018). ``Counterfactual Normalization: Proactively Addressing Dataset Shift Using Causal Mechanisms..'' In: *UAI*, pp. 947–957.
Sugiyama, Masashi et al. (2008). ``Direct importance estimation for covariate shift adaptation.'' In: *Annals of the Institute of Statistical Mathematics* 60.4, pp. 699–746.