Towards Backwards-Compatible Data with Confounded Domain Adaptation

Calvin McCarter
February 8, 2025

Overview

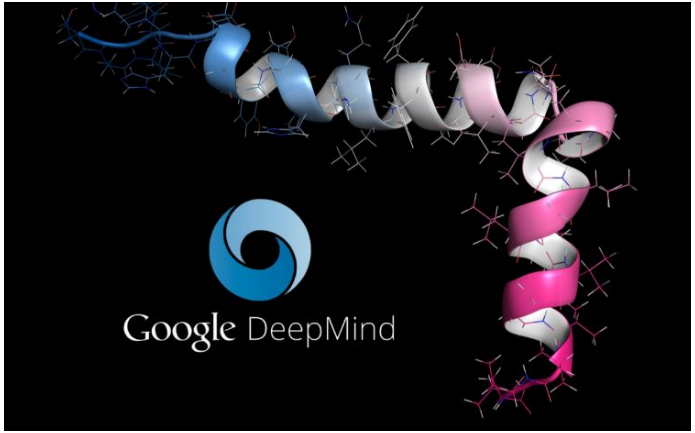
- Background
  - Some motivating examples
  - Previous work on domain adaptation
- Confounded shift
- Confounded domain adaptation
  - Proposed framework
  - Two concrete implementations (based on reverse Gaussian KL divergence and MMD)
- Experiments

McCarter, C. Towards Backwards-Compatible Data with Confounded Domain Adaptation. Transactions on Machine Learning Research. 2024. [paper] [code]

AI for biology: the problem of data heterogeneity

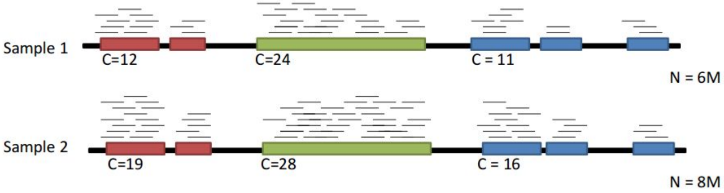
Protein structure prediction:

- homogeneous data
- distance measurements are absolute



Most other biological datasets:

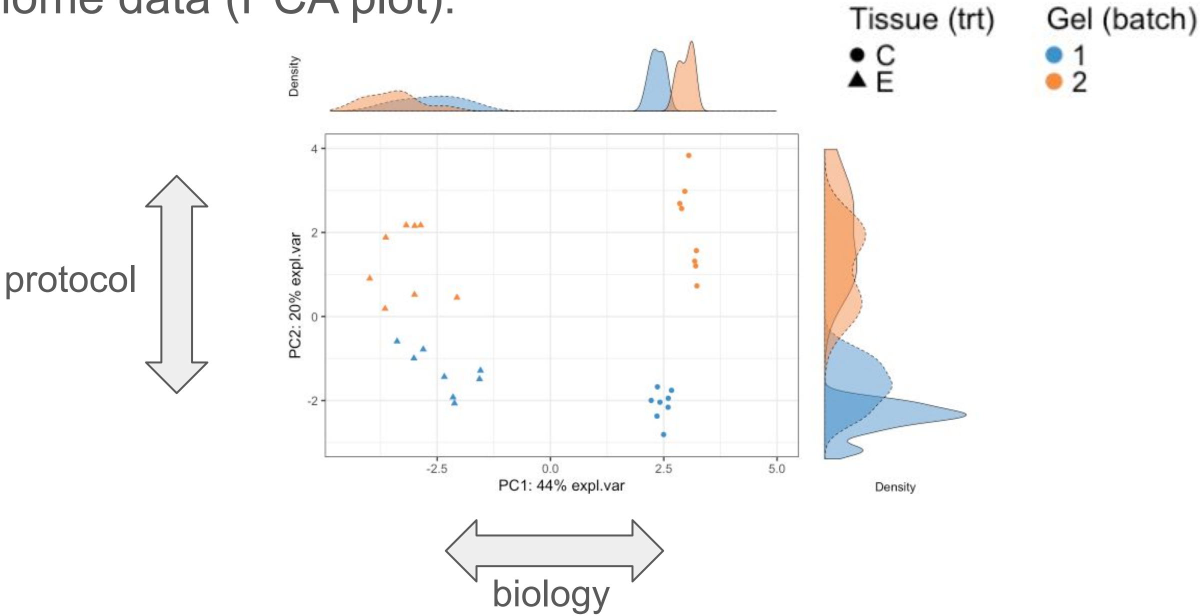
- heterogeneous data
- relative measurements (eg gene expression)



Batch effects

Technical differences among datasets due to *different protocols*

Sponge microbiome data (PCA plot):



protocol

biology

Sacristán-Soriano, Oriol, Bernard Banaigs, Emilio O Casamayor, and Mikel A Becerro. 2011. “Exploring the Links Between Natural Products and Bacterial Assemblages in the Sponge Aplysina Aerophoba.” Applied and Environmental Microbiology 77 (3). Am Soc Microbiol: 862–70.

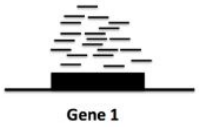
Wang, Y., & LêCao, K. A. (2020). Managing batch effects in microbiome data. Briefings in bioinformatics, 21(6), 1954-1970.

Multi-omics alignment: SNARE-seq

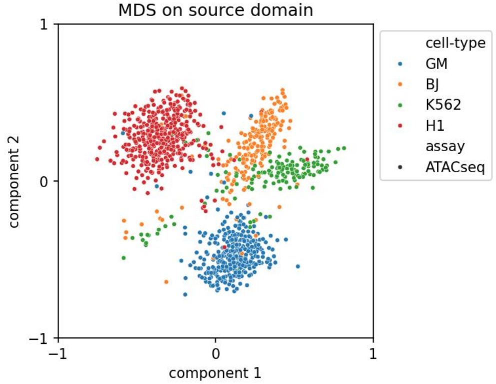
Technical differences among datasets due to *different experiments*

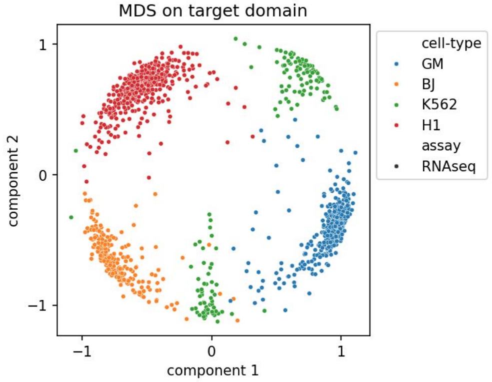
Closed chromatin
Open chromatin

- 19 dimensional ATAC-seq



- 10 dimensional RNA-seq





Chen, Song, Blue B. Lake, and Kun Zhang. "High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell." Nature biotechnology 37, no. 12 (2019): 1452-1457.

Domain adaptation methods

- Model training methods (improving robustness to technical variation)
- **Data transformation methods (matching distributions)**
  - sample reweighting
  - **feature transformation**

Our goal:

Estimate what the features would have looked like, had they been obtained using the same technical process as the reference dataset.

Domain adaptation notation

- features (i.e. covariates): $X \in \mathcal{X}$
- confounding variables: $Z \in \mathcal{Z}$
- other variables to predict, given features: $Y \in \mathcal{Y}$

real-valued vectors: $\boldsymbol{x}_S \in \mathbb{R}^{M_S}, \boldsymbol{x}_T \in \mathbb{R}^{M_T}$

user-specified confounder-space kernel function $k_{\mathcal{Z}}(z^{(n_1)}, z^{(n_2)})$

A joint distribution over covariate space $\mathcal{X}$ and confounder space $\mathcal{Z}$ is called a domain $\mathcal{D}$.

- We consider two domains: source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$.
- $\mathcal{D}_S^X, \mathcal{D}_T^X$ denote the marginal distributions of covariates under the source and target domains.
- $\mathcal{D}_S^Z, \mathcal{D}_T^Z$ denote the corresponding marginal distributions of confounders.

Also assume:

- N_S samples from source, N_T samples from target
- $N = N_S + N_T$, and $N_S >> N_T$

Affine domain adaptation: Gaussian optimal transport

The optimal transport map under the type-2 Wasserstein metric for

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S) \qquad \text{to} \qquad \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)$$

is:

$$\boldsymbol{x} \mapsto \boldsymbol{\mu}_T + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{\mu}_S) = \boldsymbol{A}\boldsymbol{x} + (\boldsymbol{\mu}_T - \boldsymbol{A}\boldsymbol{\mu}_S), \quad \text{where}$$

$$\boldsymbol{A} = \boldsymbol{\Sigma}_S^{-1/2}\left(\boldsymbol{\Sigma}_S^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_S^{1/2}\right)^{1/2}\boldsymbol{\Sigma}_S^{-1/2} = \boldsymbol{A}^\top.$$

Observe:

- All you need are samples to estimate the means and covariances.
- This also minimizes the Gaussian KL divergence.

Affine domain adaptation: MMD

Given the Gaussian kernel for feature-space vectors,

$$k_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(- \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2} \right)$$

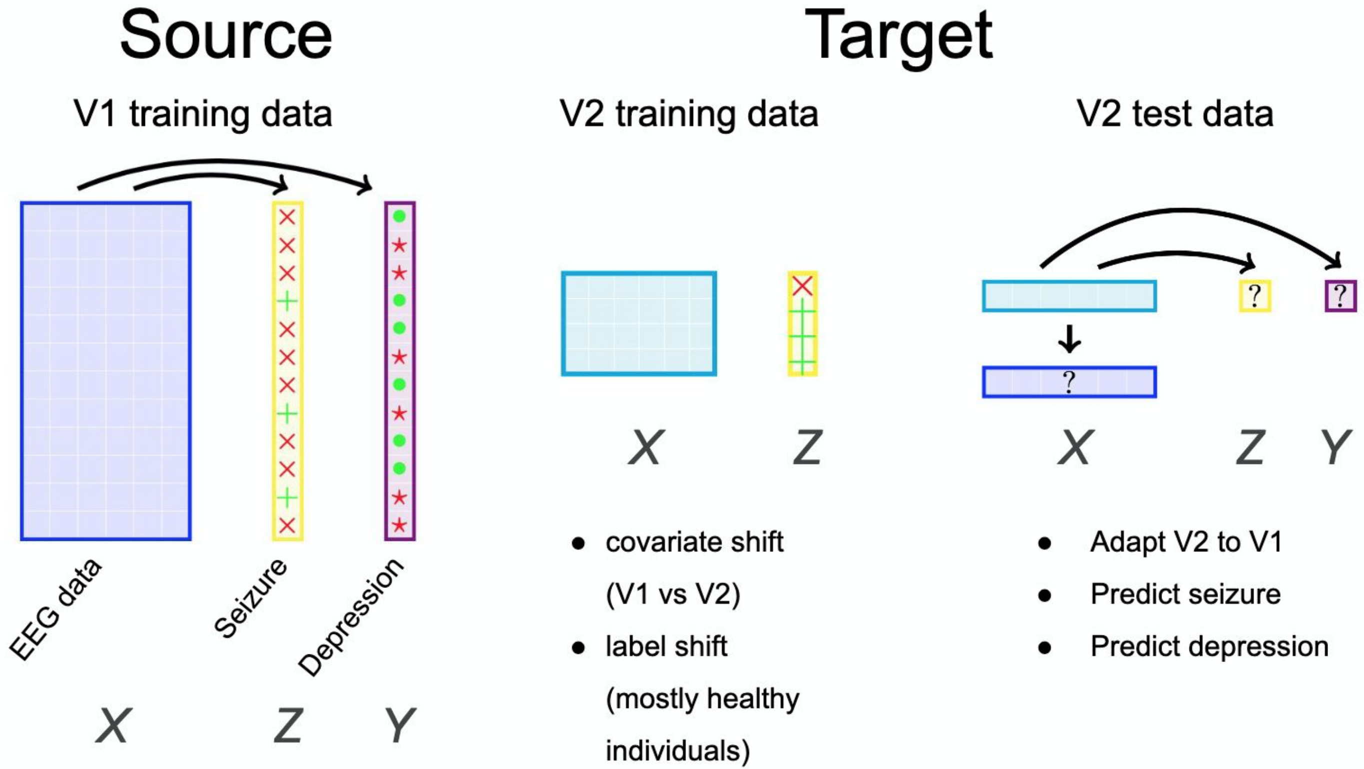
the maximum mean discrepancy (MMD) loss is 0 iff the distributions are identical:

$$\begin{aligned} \mathrm{MMD}^2(\mathcal{D}_T, \mathcal{D}_S) =& \mathbb{E}_{\boldsymbol{x}^{(n_1)}, \boldsymbol{x}^{(n_1)\prime} \sim \mathcal{D}_T} k_{\mathcal{X}}(\boldsymbol{x}^{(n_1)}, \boldsymbol{x}^{(n_1)\prime}) \\ & - 2\mathbb{E}_{\boldsymbol{x}^{(n_1)} \sim \mathcal{D}_T, \boldsymbol{x}^{(n_2)} \sim \mathcal{D}_S} k_{\mathcal{X}}(\boldsymbol{x}^{(n_1)}, \boldsymbol{A}\boldsymbol{x}^{(n_2)} + \boldsymbol{b}) \\ & + \mathbb{E}_{\boldsymbol{x}^{(n_2)}, \boldsymbol{x}^{(n_2)\prime} \sim \mathcal{D}_S} k_{\mathcal{X}}(\boldsymbol{A}\boldsymbol{x}^{(n_2)} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x}^{(n_2)\prime} + \boldsymbol{b}). \end{aligned}$$

Given source and target datasets, you can optimize this via sampling.

The problem of confounding

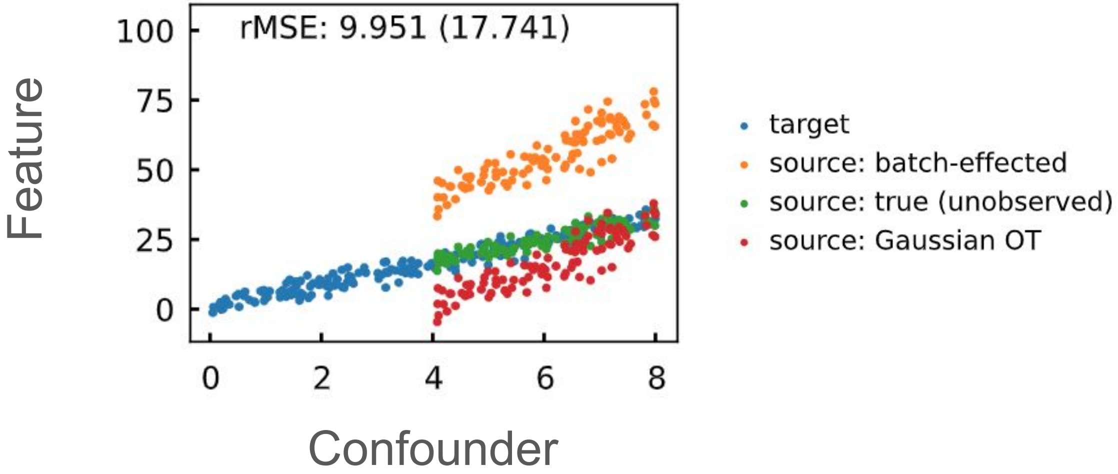
When "what you measure" and "how you measure it" are confounded:



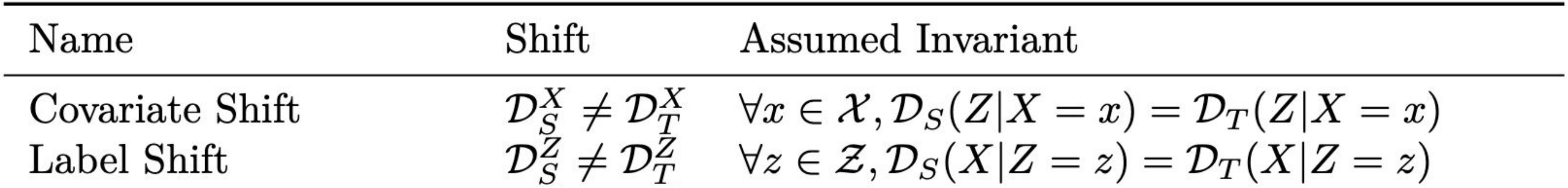
- covariate shift (V1 vs V2)
- label shift (mostly healthy individuals)

- Adapt V2 to V1
- Predict seizure
- Predict depression

How confounding affects domain adaptation



Domain adaptation settings (1)

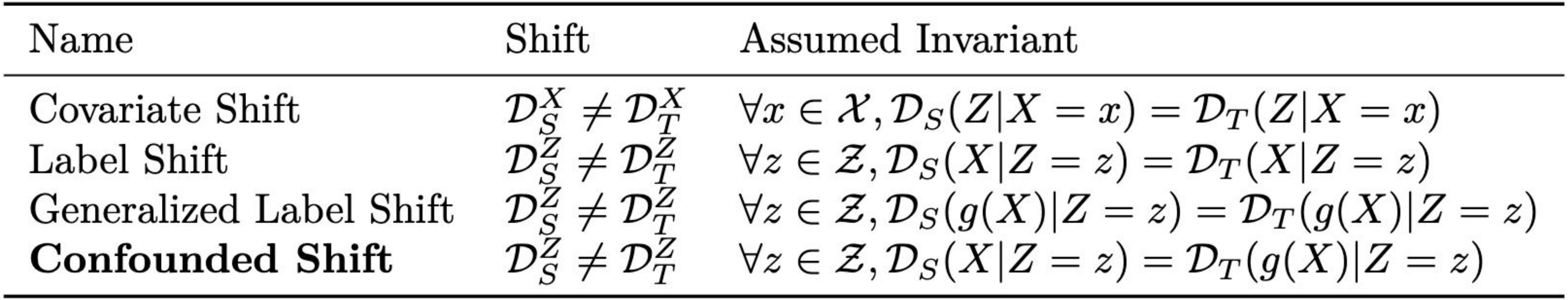


Name	Shift	Assumed Invariant
Covariate Shift	$\mathcal{D}_S^X \neq \mathcal{D}_T^X$	$\forall x \in \mathcal{X}, \mathcal{D}_S(Z\|X = x) = \mathcal{D}_T(Z\|X = x)$
Label Shift	$\mathcal{D}_S^Z \neq \mathcal{D}_T^Z$	$\forall z \in \mathcal{Z}, \mathcal{D}_S(X\|Z = z) = \mathcal{D}_T(X\|Z = z)$

Covariate shift ⇒ feature transformation methods

Label shift ⇒ sample reweighting methods

Domain adaptation settings (2)



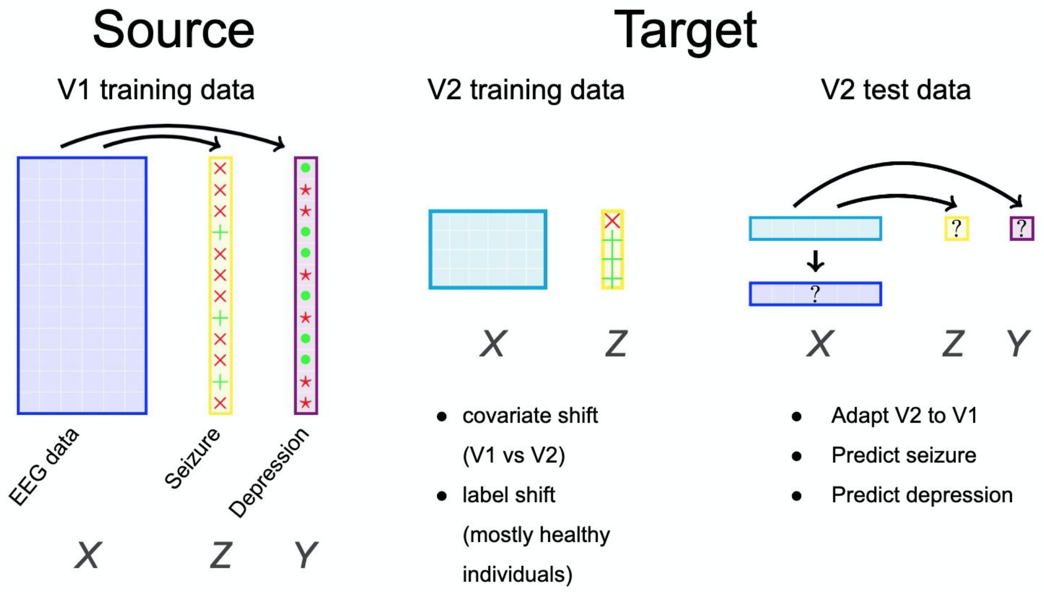
Name	Shift	Assumed Invariant
Covariate Shift	$\mathcal{D}_S^X \neq \mathcal{D}_T^X$	$\forall x \in \mathcal{X}, \mathcal{D}_S(Z\|X = x) = \mathcal{D}_T(Z\|X = x)$
Label Shift	$\mathcal{D}_S^Z \neq \mathcal{D}_T^Z$	$\forall z \in \mathcal{Z}, \mathcal{D}_S(X\|Z = z) = \mathcal{D}_T(X\|Z = z)$
Generalized Label Shift	$\mathcal{D}_S^Z \neq \mathcal{D}_T^Z$	$\forall z \in \mathcal{Z}, \mathcal{D}_S(g(X)\|Z = z) = \mathcal{D}_T(g(X)\|Z = z)$
Confounded Shift	$\mathcal{D}_S^Z \neq \mathcal{D}_T^Z$	$\forall z \in \mathcal{Z}, \mathcal{D}_S(X\|Z = z) = \mathcal{D}_T(g(X)\|Z = z)$

Confounded Shift and Generalized Label Shift coincide with:

$$\tilde{g}(\{X, D\}) = \begin{cases} g(X) & D = T \\ X & D = S \end{cases}$$

Generalized Label Shift:
Tachet des Combes, Remi, et al. "Domain adaptation with conditional distribution matching and generalized label shift." Advances in Neural Information Processing Systems 33 (2020): 19276-19289.

Towards backwards-compatible data



Source

V1 training data

Target

V2 training data

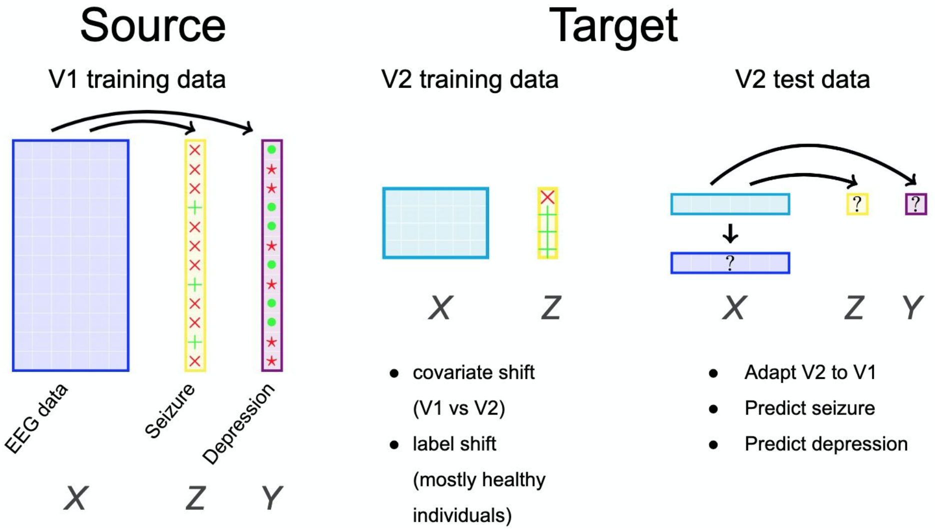
V2 test data

- covariate shift (V1 vs V2)
- label shift (mostly healthy individuals)

- Adapt V2 to V1
- Predict seizure
- Predict depression

- We might be unable to update downstream models.
- Confounder value is possibly unknown at test time.

Using backwards-compatible data



- covariate shift (V1 vs V2)
- label shift (mostly healthy individuals)

- Adapt V2 to V1
- Predict seizure
- Predict depression

Unknown shift: g^{-1}

Assume: $p_{\mathcal{D}_S}(Y|X) = p_{\mathcal{D}_T}(Y|g(X))$

Downstream prediction model: $h : \mathcal{X}_S \rightarrow \mathcal{Y}$

Estimated transformation: $\hat{g} : \mathcal{X}_T \rightarrow \mathcal{X}_S$

$$h \circ \hat{g}$$

Confounded domain adaptation

$$\min_{f_\theta} \mathbb{E}_{z\sim\hat{\mathcal{D}}_Z}\ d\Big(\mathcal{D}_T(\boldsymbol{x}|Z=z), \mathcal{D}_S(f_\theta(\boldsymbol{x})|Z=z)\Big)$$

Minimizes the expected divergence between conditional distributions

Requires 4 ingredients:

- a feature-space transformation $f_\theta : \mathcal{X} \to \mathcal{X}$
- a prior confounder distribution $\hat{\mathcal{D}}_Z$
- a conditional generative model for $\mathcal{D}_\cdot(\boldsymbol{x}|Z=z)$
- a distance/divergence function d

Feature-space transformation

$$\min_{f_\theta} \mathbb{E}_{z\sim\hat{\mathcal{D}}_Z}\; d\Big(\mathcal{D}_T(\boldsymbol{x}|Z=z), \mathcal{D}_S(\boxed{f_\theta(\boldsymbol{x})}|Z=z)\Big)$$

We restrict ourselves to linear transforms in this work:

- affine

  $$\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$$

- location-scale
  (requires same dimensionality)

  $$\boldsymbol{A} = \mathrm{diag}(\boldsymbol{a})$$

Prior confounder distribution (1)

$$\min_{f_\theta} \mathbb{E}_{\boxed{z\sim\hat{\mathcal{D}}_Z}} d\Big(\mathcal{D}_T(\boldsymbol{x}|Z=z), \mathcal{D}_S(f_\theta(\boldsymbol{x})|Z=z)\Big)$$

We can be flexible since $\mathcal{D}_T(X|Z) = \mathcal{D}_S(X|Z) \Rightarrow \mathcal{D}_T(X|Z=z) = \mathcal{D}_S(X|Z=z) \quad \forall z$

Goal: minimize the distance between conditional distributions only where we can estimate them with high accuracy.

Idea: sample from the product of $\mathcal{D}_S^Z$ and $\mathcal{D}_T^Z$

Prior confounder distribution (2)

$$\min_{f_\theta} \mathbb{E}_{\boxed{z\sim\hat{\mathcal{D}}_Z}} \; d\Big(\mathcal{D}_T(\boldsymbol{x}|Z=z), \mathcal{D}_S(f_\theta(\boldsymbol{x})|Z=z)\Big)$$

Compute the kernel density estimators $\hat{\mathcal{D}}^Z_S$ and $\hat{\mathcal{D}}^Z_T$

Choose $\hat{\mathcal{D}}^Z_\times := \hat{\mathcal{D}}^Z_S \times \hat{\mathcal{D}}^Z_T$

Reweight all observed values by $\hat{\mathcal{D}}^Z_\times$:

$$\hat{\mathcal{D}}^Z_\times := \sum_n^{N_S} \boldsymbol{w}_S^{(n)} \delta(z - Z_S^{(n)}) + \sum_n^{N_T} \boldsymbol{w}_T^{(n)} \delta(z - Z_T^{(n)}), \text{ where}$$

$$\boldsymbol{w}_S^{(n)} \propto \frac{\sum_{i=1}^{N_S} k_{\mathcal{Z}}(Z_S^{(i)}, Z_S^{(n)})}{\sum_{j=1}^{N_S}\sum_{i=1}^{N_S} k_{\mathcal{Z}}(Z_S^{(i)}, Z_S^{(j)})} \times \frac{\sum_{i=1}^{N_T} k_{\mathcal{Z}}(Z_T^{(i)}, Z_S^{(n)})}{\sum_{j=1}^{N_T}\sum_{i=1}^{N_T} k_{\mathcal{Z}}(Z_T^{(i)}, Z_T^{(j)})} \text{ and}$$

$$\boldsymbol{w}_T^{(n)} \propto \frac{\sum_{i=1}^{N_S} k_{\mathcal{Z}}(Z_S^{(i)}, Z_T^{(n)})}{\sum_{j=1}^{N_S}\sum_{i=1}^{N_S} k_{\mathcal{Z}}(Z_S^{(i)}, Z_S^{(j)})} \times \frac{\sum_{i=1}^{N_T} k_{\mathcal{Z}}(Z_T^{(i)}, Z_T^{(n)})}{\sum_{j=1}^{N_T}\sum_{i=1}^{N_T} k_{\mathcal{Z}}(Z_T^{(i)}, Z_T^{(j)})}.$$

Sampling from conditional distributions (1)

$$\min_{f_\theta} \mathbb{E}_{z \sim \hat{\mathcal{D}}_Z} \; d\Big(\boxed{\mathcal{D}_T(\boldsymbol{x}|Z=z)}, \boxed{\mathcal{D}_S(f_\theta(\boldsymbol{x})|Z=z)}\Big)$$

- For each given value of z, we generate $K_{\mathcal{X}}$ samples from the conditional distributions for both source and target

- Learn, then sample from, models for features | confounders

- We use MICE-Forest (Wilson, 2022), but you could plug in a conditional diffusion model, language model, etc.

Sampling from conditional distributions (2)

$$\min_{f_\theta} \mathbb{E}_{z\sim\hat{\mathcal{D}}_Z}\ d\left(\boxed{\mathcal{D}_T(\boldsymbol{x}|Z=z)}, \boxed{\mathcal{D}_S(f_\theta(\boldsymbol{x})|Z=z)}\right)$$

- Conditional generative modeling *is* multiple imputation.
- We concatenate the original dataset and a second copy with all features masked and all confounder(s) unmasked.
- We use MICE-Forest imputation (Wilson, 2022)
  - Multiple imputation with chained equations (MICE) (Van Buuren et al, 1999) is a leading method.
  - Gradient-boosted decision trees (Ke et al, 2017) flexibly handle tabular data.

Van Buuren, Stef, Hendriek C. Boshuizen, and Dick L. Knook. "Multiple imputation of missing blood pressure covariates in survival analysis." Statistics in medicine 18.6 (1999): 681-694.

Wilson, Samuel Von, Cebere, Bogdan, Myatt, James, & Wilson, Samuel. 2022 (Dec.). AnotherSamWilson/miceforest: Release for Zenodo DOI.

Ke, Guolin, Meng, Qi, Finley, Thomas, Wang, Taifeng, Chen, Wei, Ma, Weidong, Ye, Qiwei, & Liu, Tie-Yan. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems.

Measuring divergences between distributions

$$\min_{f_\theta} \mathbb{E}_{z\sim\hat{\mathcal{D}}_Z} \boxed{d}\Big(\mathcal{D}_T(\boldsymbol{x}|Z=z), \mathcal{D}_S(f_\theta(\boldsymbol{x})|Z=z)\Big)$$

1. For each ***z***, obtain $K_{\mathcal{X}}$ samples from each of the source and target domains.
2. Return a scalar distance / divergence.

- Gaussian KL divergence
- Maximum mean discrepancy (MMD)
- Others are possible!

Gaussian KL divergence

For each value drawn from the confounder prior, use $K_{\mathcal{X}}$ samples to estimate the mean and covariance matrix of features at that value.

Forward KLD: $d(P,Q) := d_{KL}(P||Q)$

Reverse KLD: $d(P,Q) := d_{KL}(Q||P)$

$$\min_{\boldsymbol{A},\boldsymbol{b}} -2\log\left(|\det(\boldsymbol{A})|\right) + \sum_{n=1}^{N} \boldsymbol{w}_n * \Big[\operatorname{tr}\left({\boldsymbol{\Sigma}_T^{(n)}}^{-1} \boldsymbol{A}\boldsymbol{\Sigma}_S^{(n)}\boldsymbol{A}^\top\right) \\ +\left(\boldsymbol{A}\boldsymbol{\mu}_S^{(n)} + \boldsymbol{b} - \boldsymbol{\mu}_T^{(n)}\right)^\top {\boldsymbol{\Sigma}_T^{(n)}}^{-1}\left(\boldsymbol{A}\boldsymbol{\mu}_S^{(n)} + \boldsymbol{b} - \boldsymbol{\mu}_T^{(n)}\right)\Big]$$

Benefits of reverse KLD:

- Preserves sign of the determinant of *A*
- Requires only a single matrix inversion per sample
- Closed-form solution for location-scale transformation

Conditional maximum mean discrepancy

For a particular *z*, the conditional MMD loss is:

$$\begin{aligned} d\Big(\mathcal{D}_T(\cdot|Z=z), \mathcal{D}_S(\cdot|Z=z)\Big) :=& \mathrm{MMD}^2(\mathcal{D}_T(\cdot|Z=z), \mathcal{D}_S(\cdot|Z=z)) \\ =& \mathbb{E}_{\boldsymbol{x}^{(n_1)}, \boldsymbol{x}^{(n_1)'} \sim \mathcal{D}_T(\cdot|Z=z)} k_{\mathcal{X}}(\boldsymbol{x}^{(n_1)}, \boldsymbol{x}^{(n_1)'}) \\ & - 2\mathbb{E}_{\boldsymbol{x}^{(n_1)} \sim \mathcal{D}_T(\cdot|Z=z), \boldsymbol{x}^{(n_2)} \sim \mathcal{D}_S(\cdot|Z=z)} k_{\mathcal{X}}(\boldsymbol{x}^{(n_1)}, \boldsymbol{A}\boldsymbol{x}^{(n_2)} + \boldsymbol{b}) \\ & + \mathbb{E}_{\boldsymbol{x}^{(n_2)}, \boldsymbol{x}^{(n_2)'} \sim \mathcal{D}_S(\cdot|Z=z)} k_{\mathcal{X}}(\boldsymbol{A}\boldsymbol{x}^{(n_2)} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{x}^{(n_2)'} + \boldsymbol{b}). \end{aligned}$$

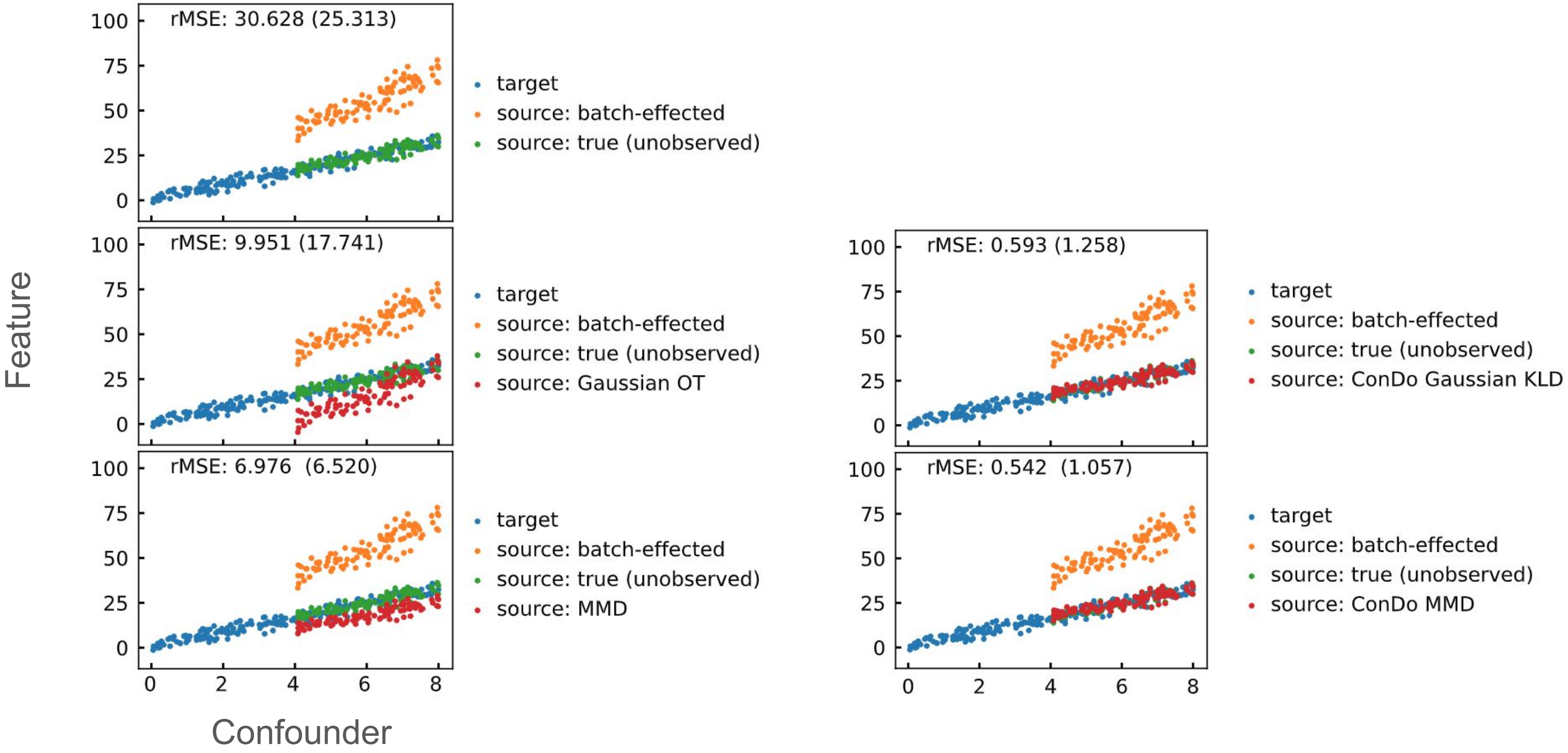
Stochastic optimization:

- Sample $K_{\mathcal{Z}}$ values from the confounder prior with replacement.
- For each *z* value, sample $K_{\mathcal{X}}$ vectors to estimate the conditional MMD loss.

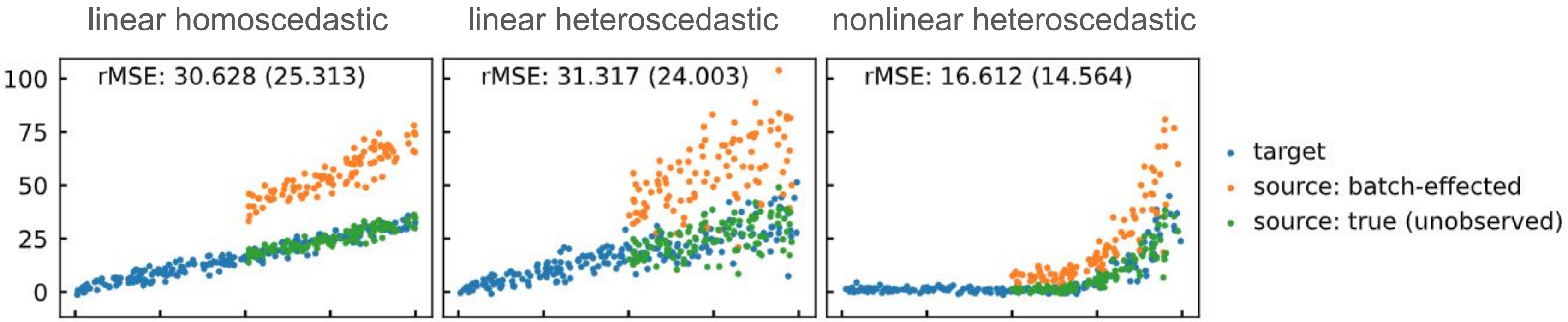
Experiments

- Synthetic data
  - 1d features with 1d continuous confounder
  - 1d features with multi-dimensional continuous confounders
  - 1d and 2d features with categorical confounders
- Hybrid data
  - ANSUR II anthropometric survey data
  - Image color adaptation
- Real data
  - California housing price prediction
  - SNARE-seq multi-omics data
  - Gene expression batch effect correction

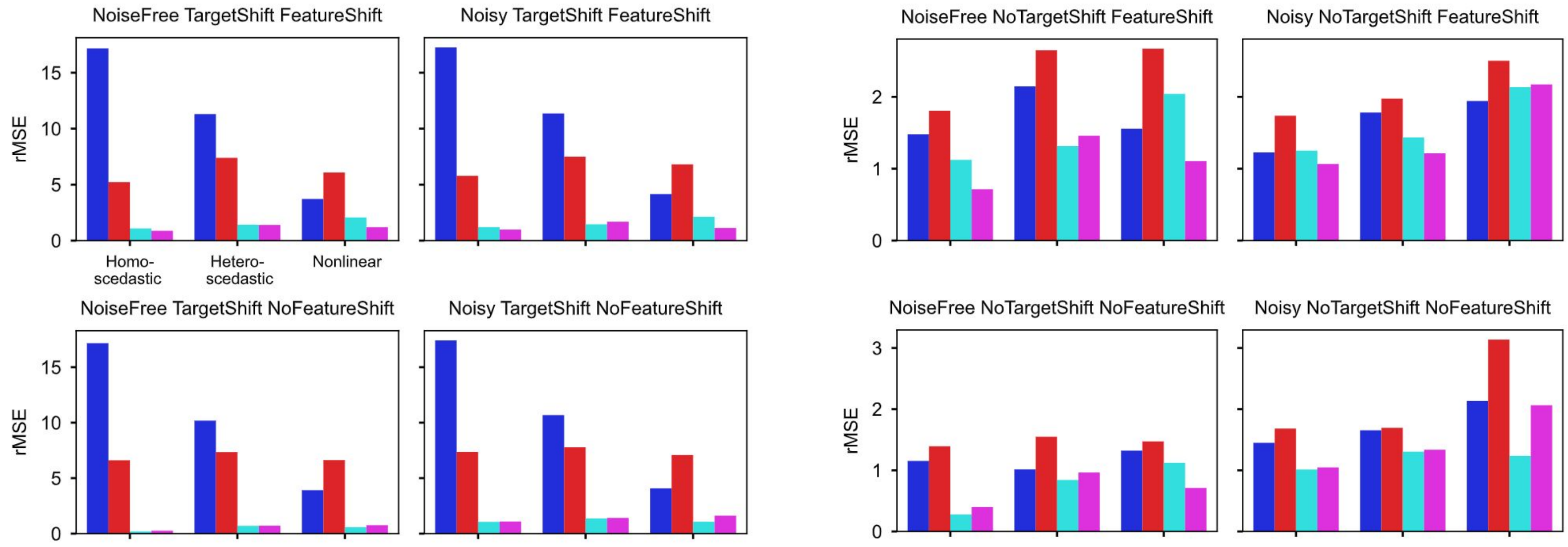
1d feature, 1d continuous confounder



Robustness to multiple types of shift (1)

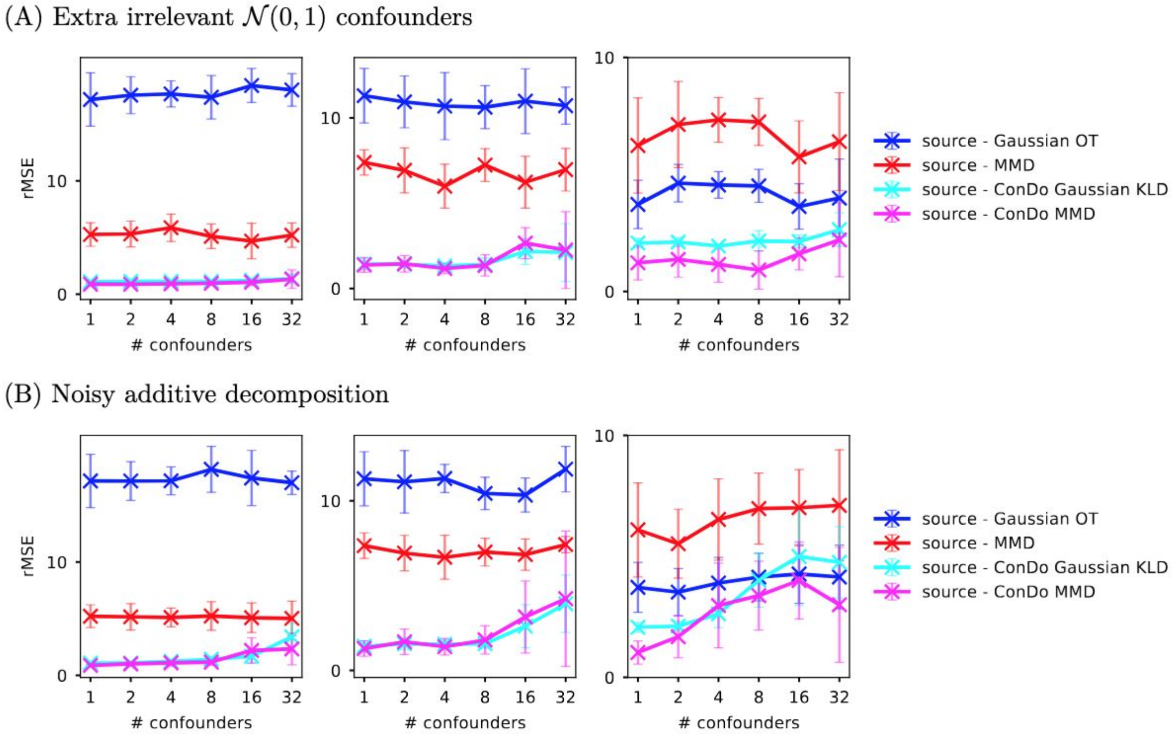


Robustness to multiple types of shift (2)



1d feature, multiple continuous confounders

3 settings: linear homoscedastic, linear heteroscedastic, nonlinear heteroscedastic



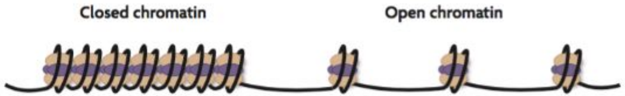
(A) Extra irrelevant $\mathcal{N}(0,1)$ confounders

(B) Noisy additive decomposition

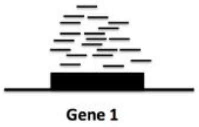
Multi-omics alignment: SNARE-seq revisited

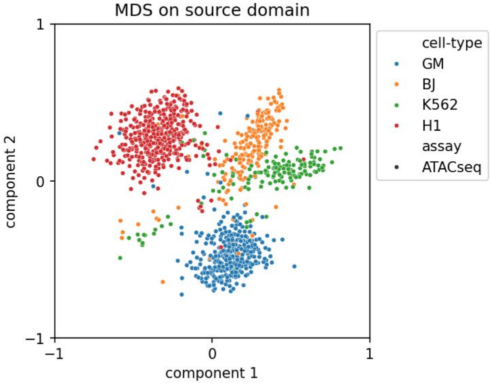
Technical differences among datasets due to *different experiments*

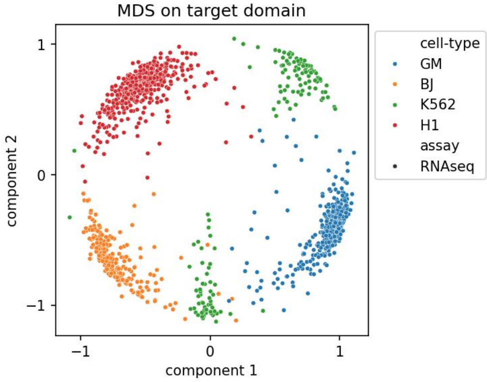
- 19 dimensional ATAC-seq



- 10 dimensional RNA-seq



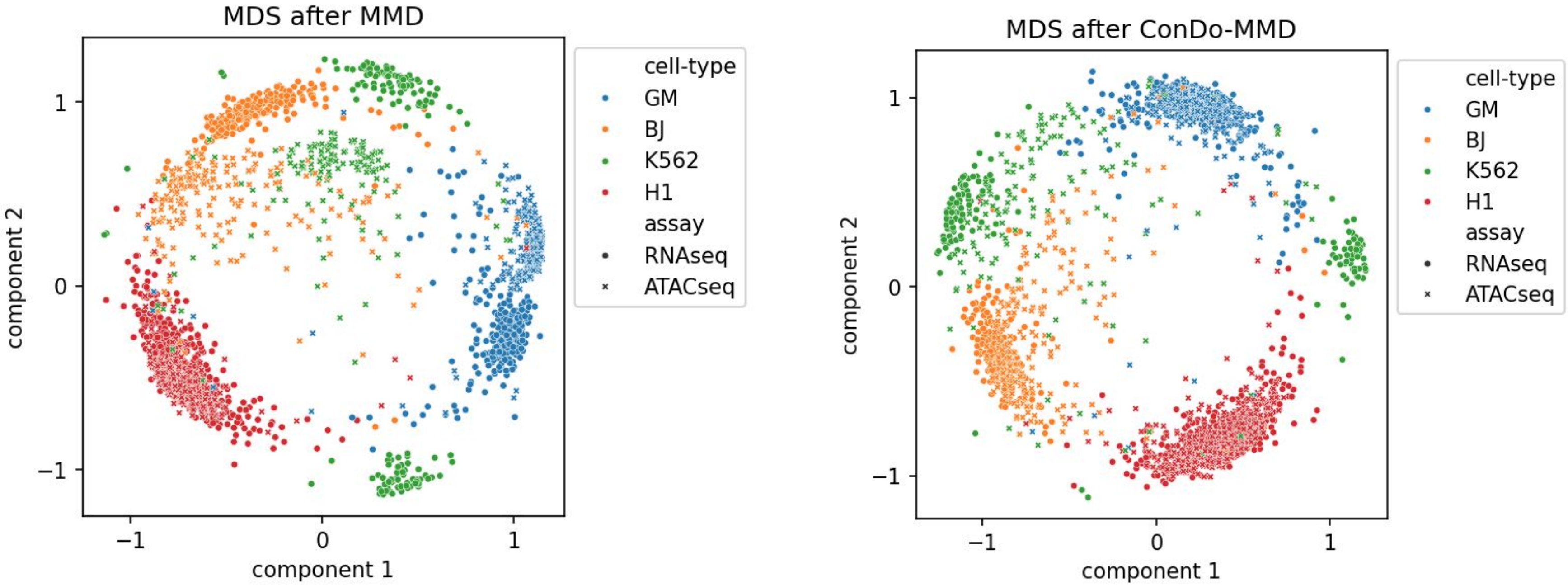




Chen, Song, Blue B. Lake, and Kun Zhang. "High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell." Nature biotechnology 37, no. 12 (2019): 1452-1457.

SNARE-seq after domain adaptation

Cell-type data improves overlap between RNAseq (o) and ATACseq (x):



SNARE-seq after domain adaptation

500 ATAC-seq samples + C (RNA-seq, ATAC-seq) pairs, with $C \in \{5, 10, 20, 50, 100\}$

Train cell-type classifier on ATAC-seq, then evaluate on RNA-seq.

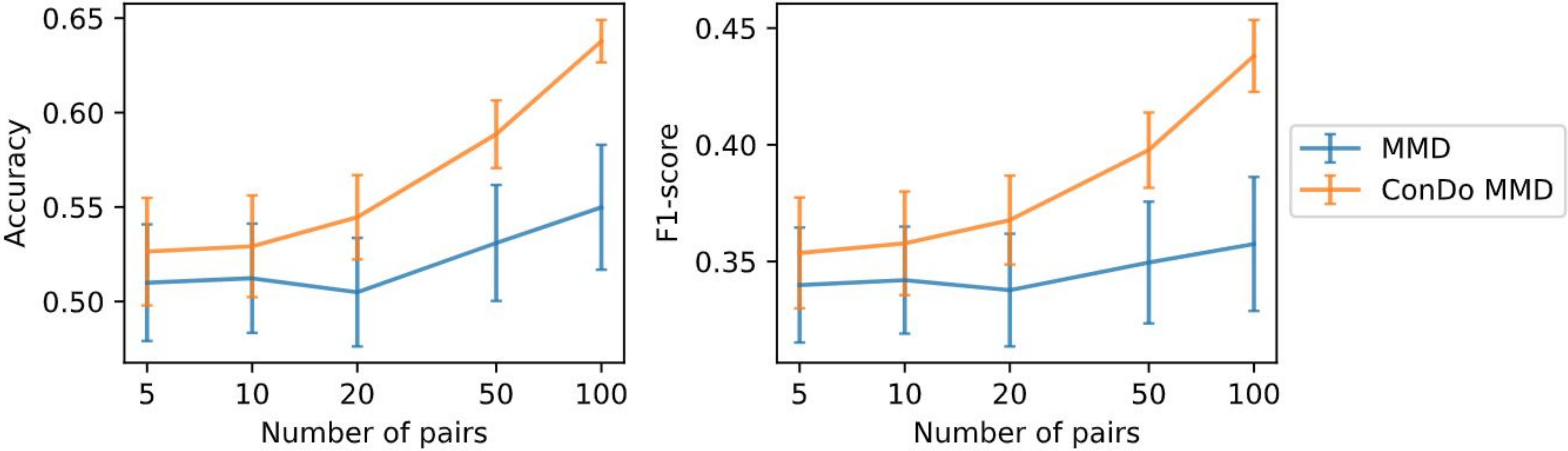
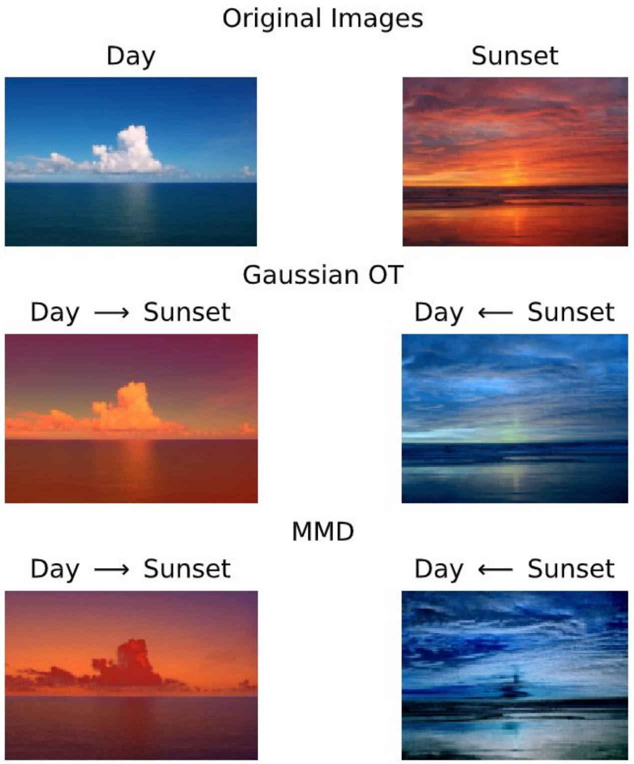


Image color adaptation - no confounding

Treat each image as a (# pixels, 3) dataset



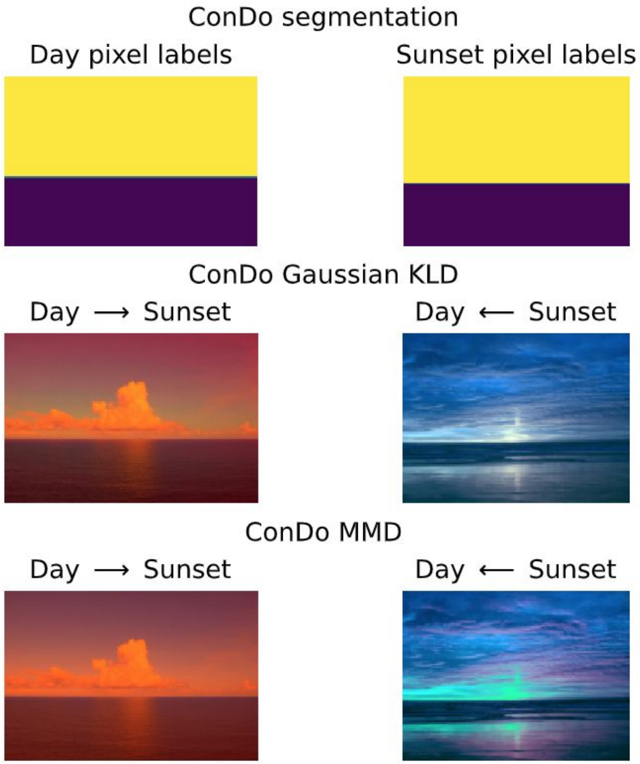
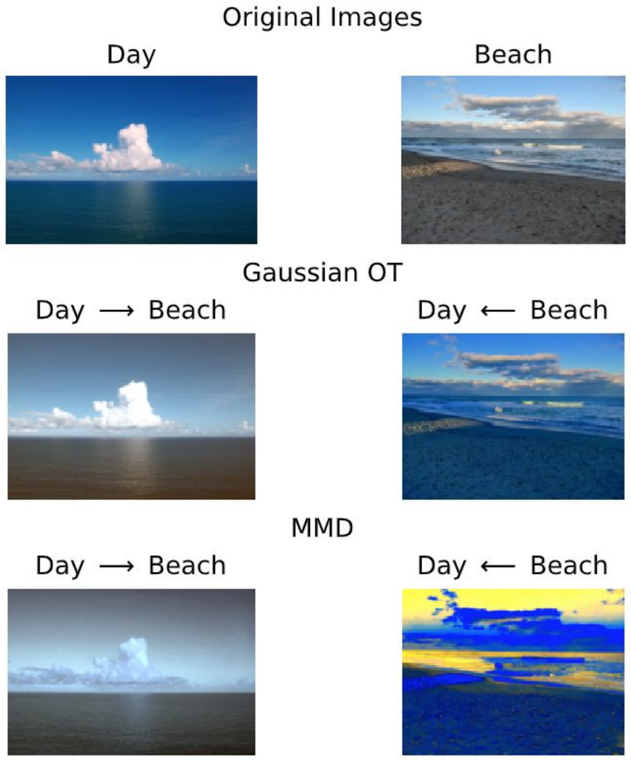
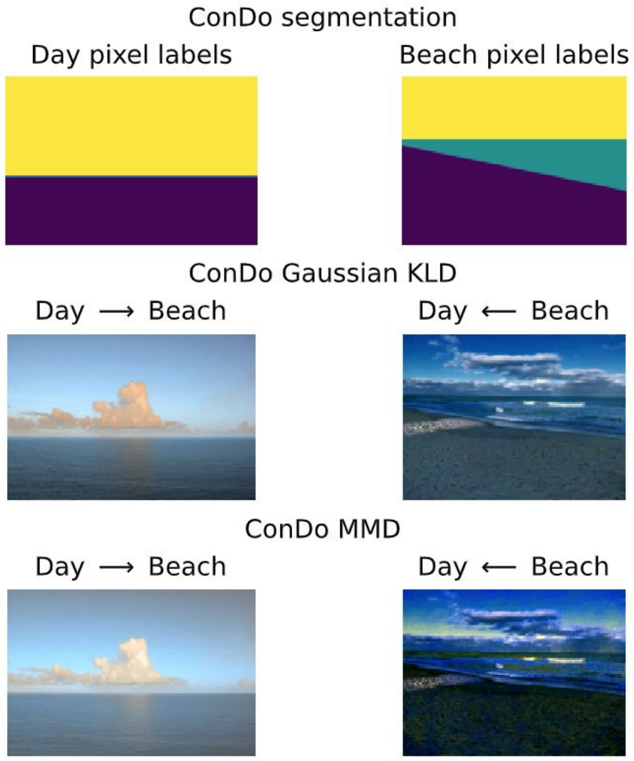


Image color adaptation - confounding

Treat each image as a (# pixels, 3) dataset





ANSUR II anthropometric data (1)

93 anthropometric measurements (e.g. wrist height) from 6068 military personnel

Source: random subsample of 500 with a 75%-25% male-female split

Target: random subsample of 500 with a 25%-75% male-female split

$$\boldsymbol{A} = \boldsymbol{U}\mathrm{diag}(\boldsymbol{d})\boldsymbol{V}^\top \quad \boldsymbol{d}_i \sim \mathrm{Unif}[0.5, 2]$$ *U*, *V* ~ Haar distributed

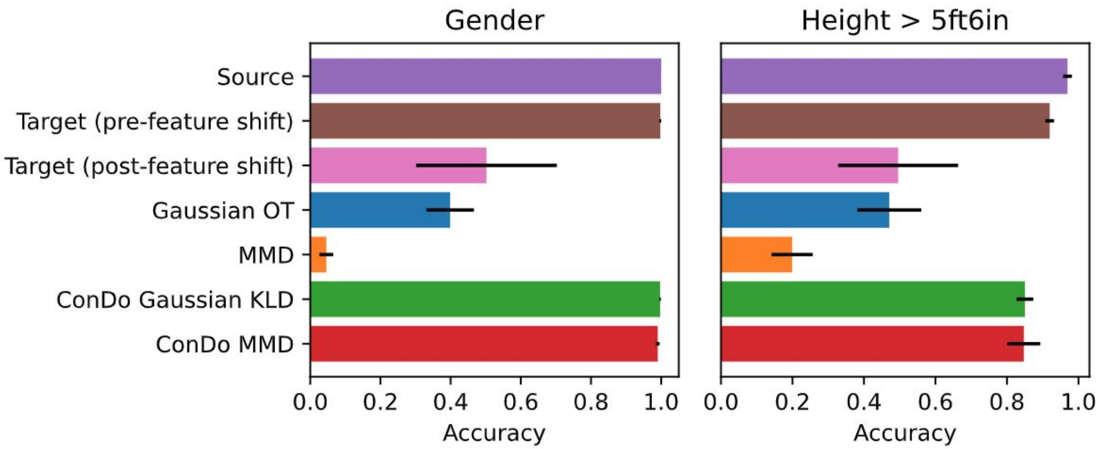
Confounder variable: Male vs Female

Prediction models, trained on source:

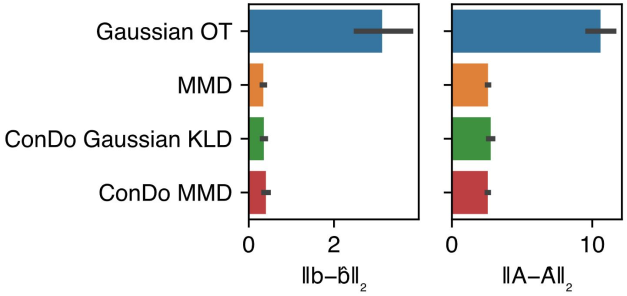
- Male vs Female
- Height greater than median

ANSUR II anthropometric data (2)

Downstream prediction performance:



True mapping parameter recovery:



Limitations

- Assumes access to all confounders at training time
- Assumes a deterministic (and linear) transformation between features
- Despite assumptions, the true transformation is non-identifiable
- Using transformed data for downstream task assumes that conditional distribution of the target variable given features is the same for source and target

Future work

- Optimal transport distance
- Constraints (e.g. non-negative) and regularization
- Nonlinear adaptations parameterized by neural networks
- Theoretical guarantees

Thanks!

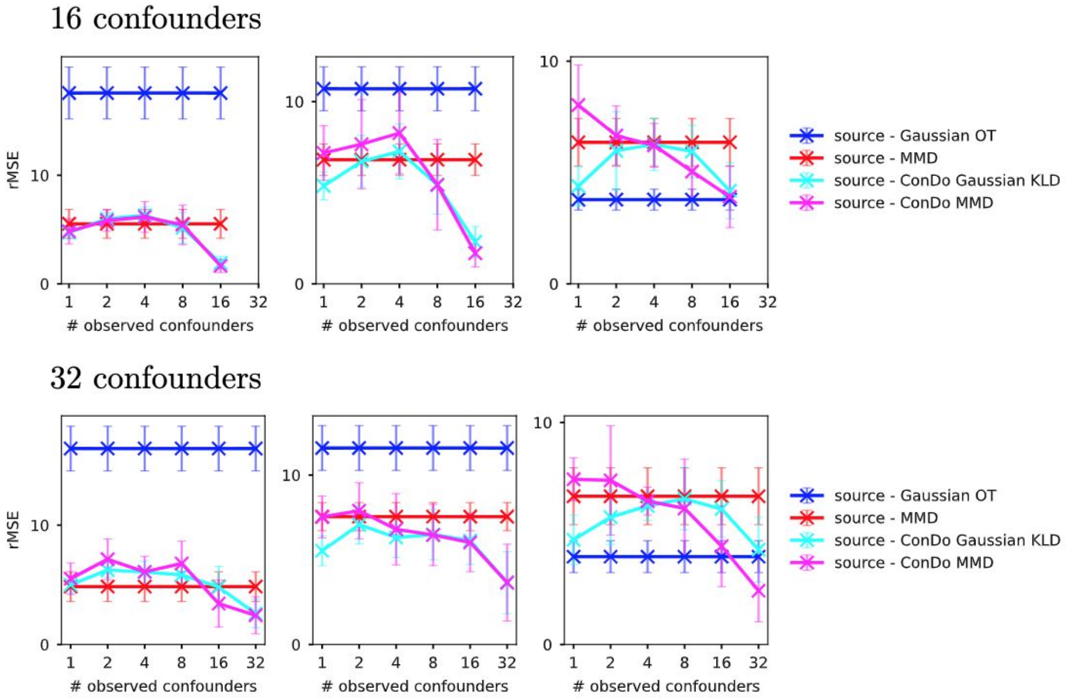
Questions?

Please feel free to reach out: mccarter.calvin@gmail.com

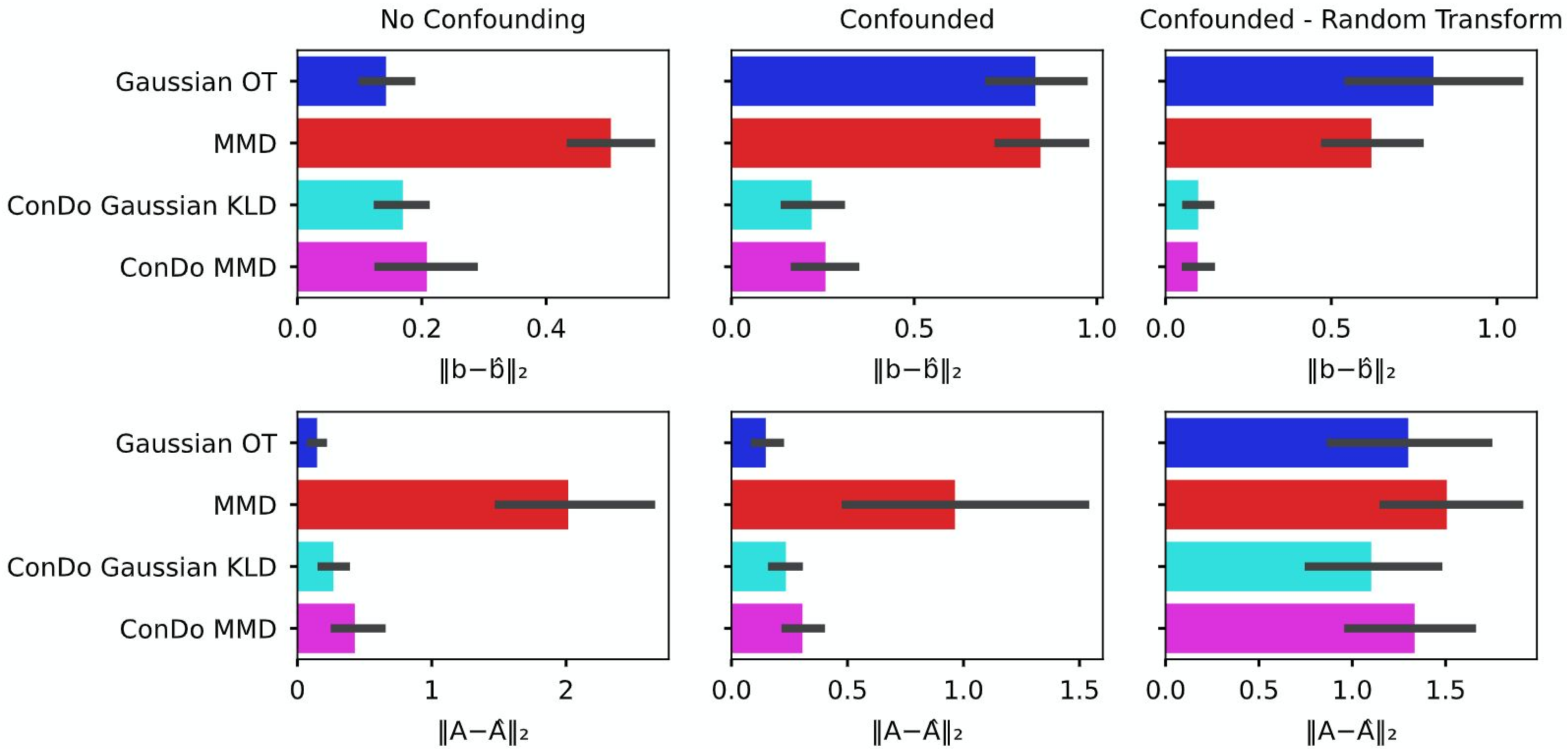
McCarter, C. Towards Backwards-Compatible Data with Confounded Domain Adaptation. Transactions on Machine Learning Research. 2024. [paper] [code]

Partially-observed confounding

Noisy additive decomposition
3 settings: linear homoskedastic, linear heteroscedastic, nonlinear heteroscedastic



True transformation recovery - 2d data, 1d confounder


No Confounding
Confounded
Confounded - Random Transform
Gaussian OT
MMD
ConDo Gaussian KLD
ConDo MMD
0.0
0.2
0.4
0.5
1.0
$\|b-\hat{b}\|_2$
0
1
2
1.5
$\|A-\hat{A}\|_2$