Domain Adaptation under Target and Conditional Shift

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, Zhikun Wang

Dept. Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany

Summary: Why and how to correct for target/conditional shift?

- Problem: predicting Y from X, under $P^{tr}_{Y|X} \neq P^{te}_{Y|X}$ and $P^{tr}(X) \neq P^{te}(X)$, but it is plausible to assume
  - ⋆ **Target shift (TarS)**: $P^{tr}_{X|Y} = P^{te}_{X|Y}$ and $P^{tr}_Y \neq P^{te}_Y$,
  - ⋆ **Conditional shift (ConS)**: $P^{tr}_{X|Y} \neq P^{te}_{X|Y}$ and $P^{tr}_Y = P^{te}_Y$, and
  - ⋆ **Generalized target shift (GeTarS)**: $P^{tr}_{X|Y} \neq P^{te}_{X|Y}$ and $P^{tr}_Y \neq P^{te}_Y$.
- Causal interpretations
- Efficient methods to correct for ConS and GeTarS with **kernel mean matching**

Possible situations for domain adaptation

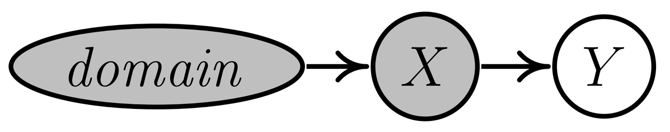


Figure 1: Covariate shift

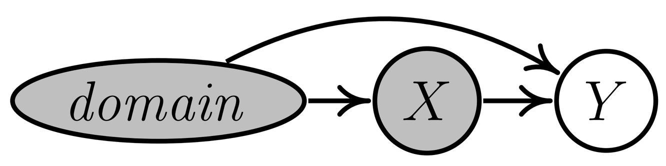


Figure 2: Both P_X and $P_{Y|X}$ change: What to do?

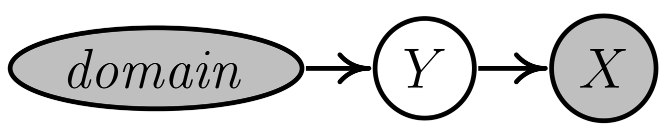


Figure 3: Target shift (or prior probability shift)

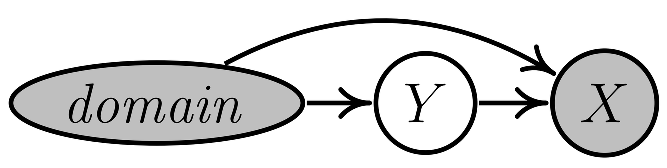


Figure 4: GeTarS (Both P_X and $P_{Y|X}$ change)

↖ P^{te}_X *helps predict* Y ↗

Distribution shift correction by data transformation/reweighting

- **Problem**: Given training data $\mathbf{D}^{tr} = \{x_i, y_i\}_{i=1}^m$, find the regressor (e.g., KRR) or classifier (e.g., SVM) $f(x)$ that works well on test data $\mathbf{D}^{te} = \{x_i\}_{i=1}^n$.
- **Importance reweighting**: Minimize the expected loss on test data:

$$R[P^{te},\theta,l(x,y,\theta)] = \mathbb{E}_{(X,Y)\sim P^{te}_{XY}}[l(x,y,\theta)] = \mathbb{E}_{(X,Y)\sim P^{tr}_{XY}} \cdot \underbrace{P^{te}_Y/P^{tr}_Y}_{\triangleq\beta^*(y)} \cdot \underbrace{P^{te}_{X|Y}/P^{tr}_{X|Y}}_{\triangleq\gamma^*(y)\equiv 1 \text{ for TarS}} \cdot l(x,y,\theta)dxdy.$$

- ⋆ assumes the support of P^{te}_{XY} is contained by that of P^{tr}_{XY}
- ⋆ factorize P_{XY} as $P_Y P_{X|Y}$ instead of $P_X P_{Y|X}$.
- ⋆ empirical version: $\widehat{R}[P^{te},\theta,l(x,y;\theta)] = \frac{1}{m}\sum_{i=1}^m \beta^*(y_i^{tr})\gamma^*(x_i^{tr},y_i^{tr})l(x_i^{tr},y_i^{tr},\theta)$.
- **Sample transformation and reweighting**: find transformation $\mathcal{T}$ such that the conditional distribution of $X^{new} = \mathcal{T}(X^{tr},Y^{tr})$ satisfies $P^{new}_{X|Y} = P^{te}_{X|Y}$; the expected loss on the test domain is

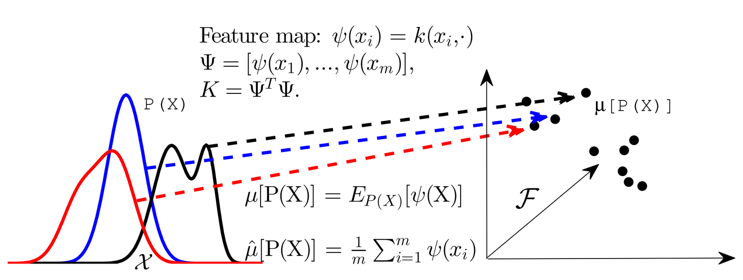
$$R[P^{te},\theta,l(x,y;\theta)] = \mathbb{E}_{P^{te}_{XY}}[l(x,y;\theta)] = \int P^{tr}_Y \cdot\beta^*(y)\cdot P^{te}_{X|Y}\cdot l(x,y;\theta)dxdy = \mathbb{E}_{(X,Y)\sim P^{tr}_Y P^{new}_{X|Y}}[\beta^*(y)\cdot l(x,y;\theta)].$$

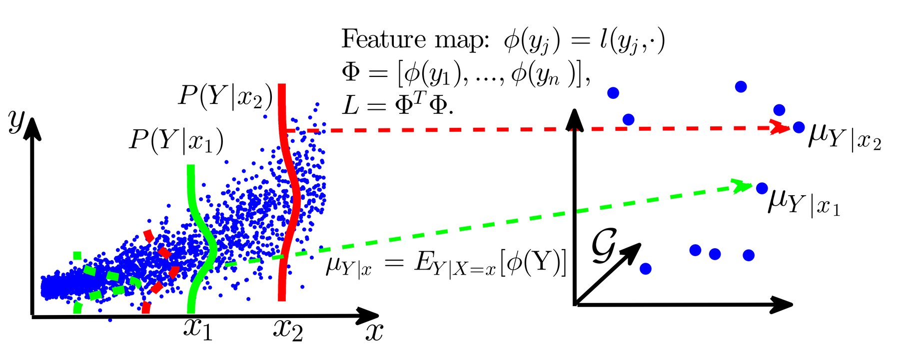
- ⋆ empirical version: $\widehat{R}[P^{te},\theta,l(x,y;\theta)] = \frac{1}{m}\sum_{i=1}^m \beta^*(y_i^{tr})l(x_i^{new},y_i^{tr};\theta)$.
- ⋆ consider $(\mathbf{x}^{new},\mathbf{y}^{tr})$ as new training data and learn under TarS.
- Will be used to correct for GeTarS. Problem: How to find $\beta^*(y)$ and/or $\mathcal{T}$?

Correction for target shift (Fig. 3)

- Aim to find $\beta^*(y) = P^{te}_Y/P^{tr}_Y$ under **TarS**: $P^{te}_{X|Y} = P^{tr}_{X|Y}$ but $P^{te}_Y \neq P^{tr}_Y$, and additional assumptions.
  - ⋆ **Richness** of traning data: the support of $P^{tr}(Y)$ contains that of $P^{te}(Y)$.
  - ⋆ **Invertibility**: only one distribution of Y, together with $P^{tr}_{X|Y}$, leads to P^{te}_X.
  - ⋆ Kernels k (for X) and l (for Y) are **characteristic**.
- Traditionally difficult, but very convenient with kernel mean matching.

- ⋆ $P(X)$ has a unique embedding $\mu[P(X)]$ with characteristic kernels.
- ⋆ Avoid explicit estimation of $P(X)$.
- ⋆ Conditional embedding is an operator from $\mathcal{F}$ to $\mathcal{G}$: $\mathcal{U}(Y|X) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$; $\mathcal{C}_{YX}$ and $\mathcal{C}_{XX}$ are *uncentered* cross- and auto-covariance operators.
- ⋆ $\mu[P(Y)] = \mathcal{U}_{Y|X}\cdot\mu[P(X)]$.
- ⋆ $\hat{\mathcal{U}}_{Y|X} = \mathbf{\Phi}(K+\lambda I)^{-1}\mathbf{\Psi}^\intercal$.





- Let $P^{new}_Y = \beta(y)P^{tr}_Y$. We find $\beta^*(y)$ by matching P^{new}_X (corresponding to P^{new}_Y and $P^{tr}_{X|Y}$) with P^{te}_X:

$$\beta^* = \arg\min_\beta \Big|\Big|\mu[P^{new}(X)] - \mu[P^{te}(X)]\Big|\Big| = \Big|\Big|\mathcal{U}[P^{tr}(X|Y)]\mathbb{E}_{Y\sim P^{tr}(Y)}[\beta(y)\phi(y)] - \mu[P^{te}(X)]\Big|\Big|,$$

whose empirical version is (K^c is the "cross" kernel matrix of X between $\mathbf{D}^{tr}$ and $\mathbf{D}^{te}$):

$$\Big|\Big|\hat{\mathcal{U}}_{X|Y}\cdot\frac{1}{m}\sum_{i=1}^m \beta_i\phi(y_i^{tr}) - \frac{1}{n}\sum_{i=1}^n \psi(x_i^{te})\Big|\Big|^2$$
$$= \frac{1}{m^2}\beta^\intercal\underbrace{L(L+\lambda_m I)^{-1}K(L+\lambda_m I)^{-1}L}_{\triangleq J}\beta - \frac{2}{mn}\underbrace{\mathbf{1}^\intercal K^c(L+\lambda_m I)^{-1}L}_{\triangleq M}\beta + \text{const.}$$

- As in the covariate shift case [1], $\beta^*(\mathbf{y}^{tr})$ can be estimated by solving a constrained **QP** problem:

$$\min.\ \frac{1}{2}\beta^\intercal J\beta - \frac{m}{n}M\beta,\quad \text{s.t. } \beta_i \in [0,B] \text{ and } \Big|\sum_{i=1}^m \beta_i - m\Big| \leq m\epsilon;\quad B \text{ and } \epsilon \text{ are parameters.}$$

Location-scale generalized target shift (Fig. 4)

- **Assumption**: Both P_Y and $P_{X|Y}$ change, but $P_{X|Y}$ changes only in the location and scale:

  i.e., $\exists\ \mathbf{w}(Y^{tr}) = \text{diag}[w_1(Y^{tr}),...,w_d(Y^{tr})]$ and $\mathbf{b}(Y^{tr}) = [b_1(Y^{tr}),...,b_d(Y^{tr})]^\intercal$ such that $X^{new} \triangleq \mathbf{w}(Y^{tr})X^{tr} + \mathbf{b}(Y^{tr})$ satisfies $P_{X^{new}|Y^{tr}} = P^{te}_{X|Y}$.
- **Identifiability**: Under certain conditions on $P^{tr}_{X|Y}(x|y_i)$, $P^{te}_{X|Y}$ and P^{te}_Y uniquely recovered by *reweighting and transoforming traning data to reproduce* P^{te}_X, i.e., by minimizing

$$\Big|\Big|\mu[P^{new}_X] - \mu[P^{te}_X]\Big|\Big|,$$

where $\mu[P^{new}_X] = \mathcal{U}[P^{new}_{X|Y}]\mu[P^{new}_Y]$, $P^{new}_Y = \beta P^{tr}_Y$, and $P^{new}_{X|Y}(x|y_i) = P^{(\mathbf{w}_i,\mathbf{b}_i)}_{X|Y}(x|y_i)$, the LS-transformed $P^{tr}_{X|Y}$.
- **Objective function**: its empirical version

$$J = \Big|\Big|\hat{\mu}[P^{new}_X] - \hat{\mu}[P^{te}_X]\Big|\Big|^2 = \frac{1}{m^2}\beta^\intercal\Omega\tilde{K}\beta - \frac{2}{mn}\mathbf{1}_n^\intercal\tilde{K}^c\beta,$$

where $\Omega \triangleq L(L+\lambda I)^{-1}$, and $\tilde{K}$ is the kernel matrix of $\mathbf{x}^{new}$.
- **Optimization**: Alternate between QP w.r.t. β and SCG optimization w.r.t. LS parameters $\{\mathbf{W},\mathbf{B}\}$.

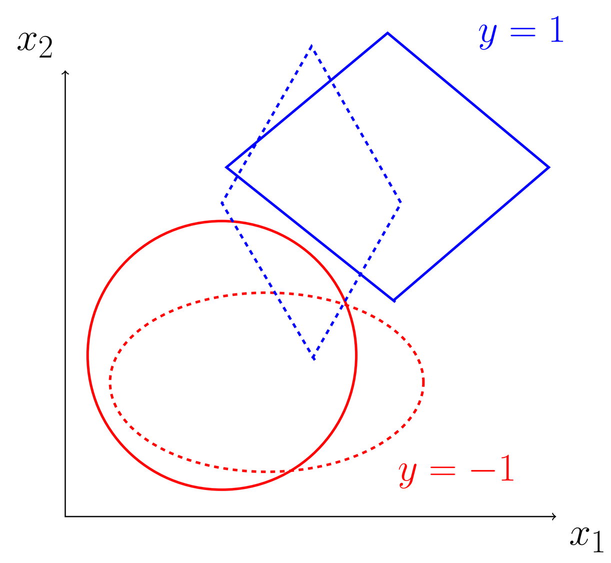
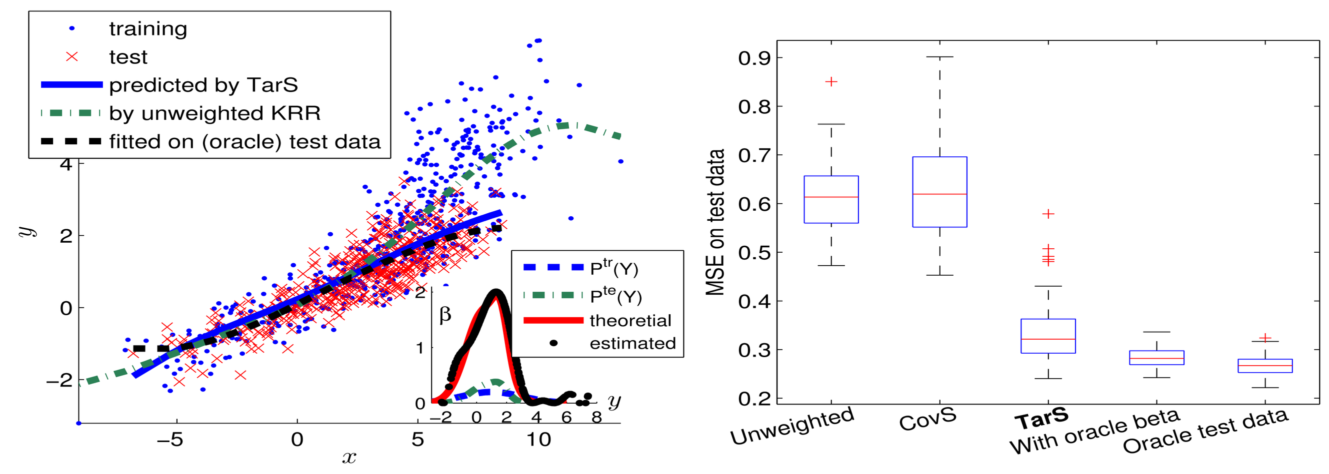


Figure 5: An illustration of LS-ConS where Y is binary and X is two-dimensional. Red and blue lines are contours of $P_{X|Y}(x|y=-1)$ and $P_{X|Y}(x|y=1)$. Solid and dashed lines represent the contours on the training and test domains.

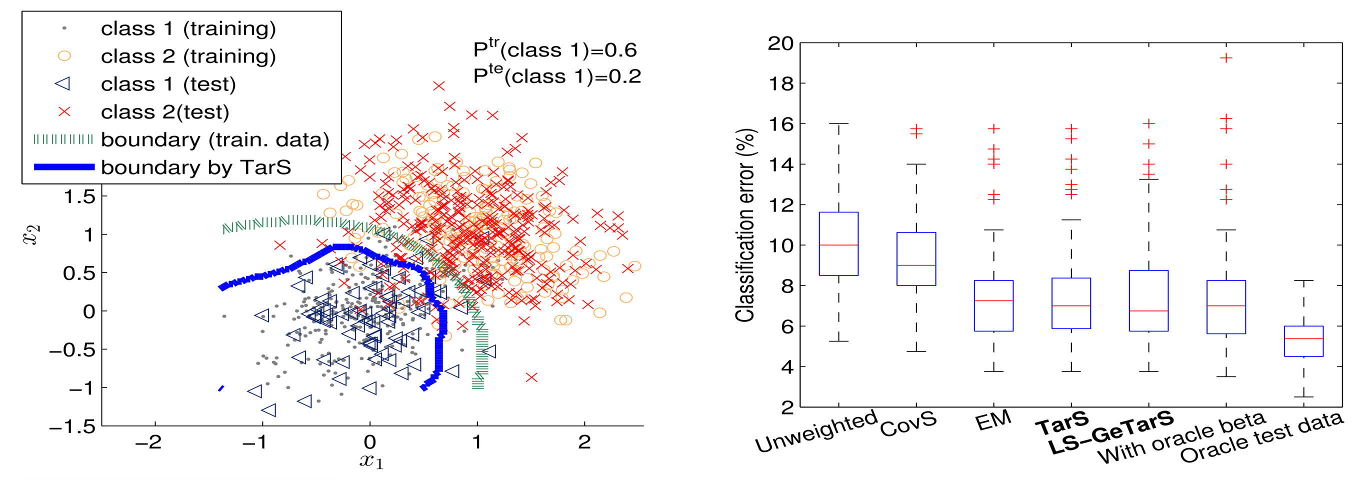
- **Regularization** on $\{\mathbf{W},\mathbf{B}\}$ for stability.

Simulations

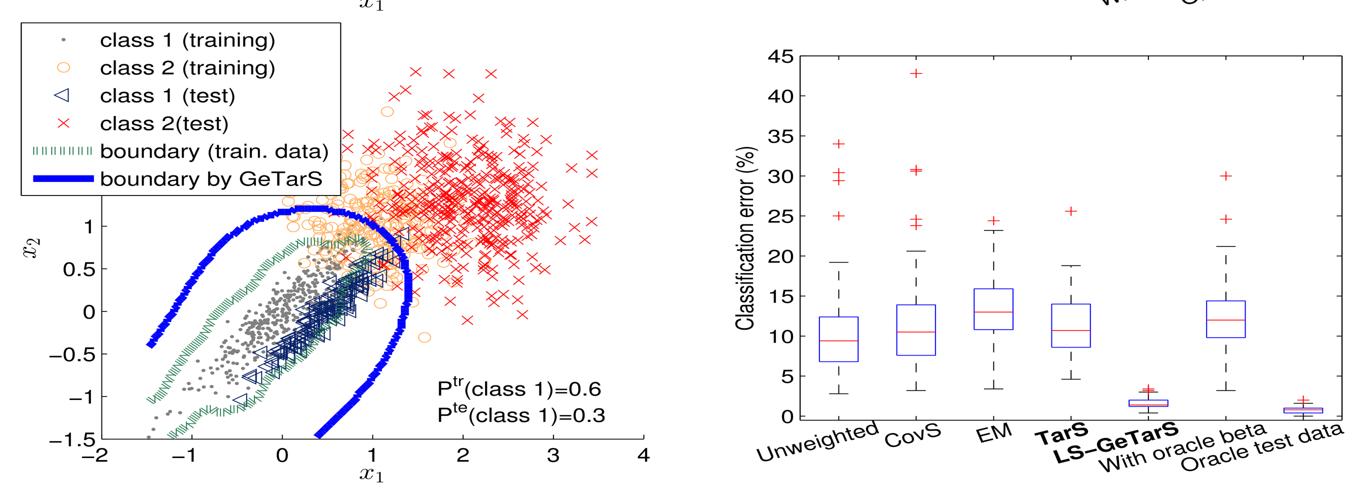
(a) Regression under TarS, where X depends on Y non-linearly.



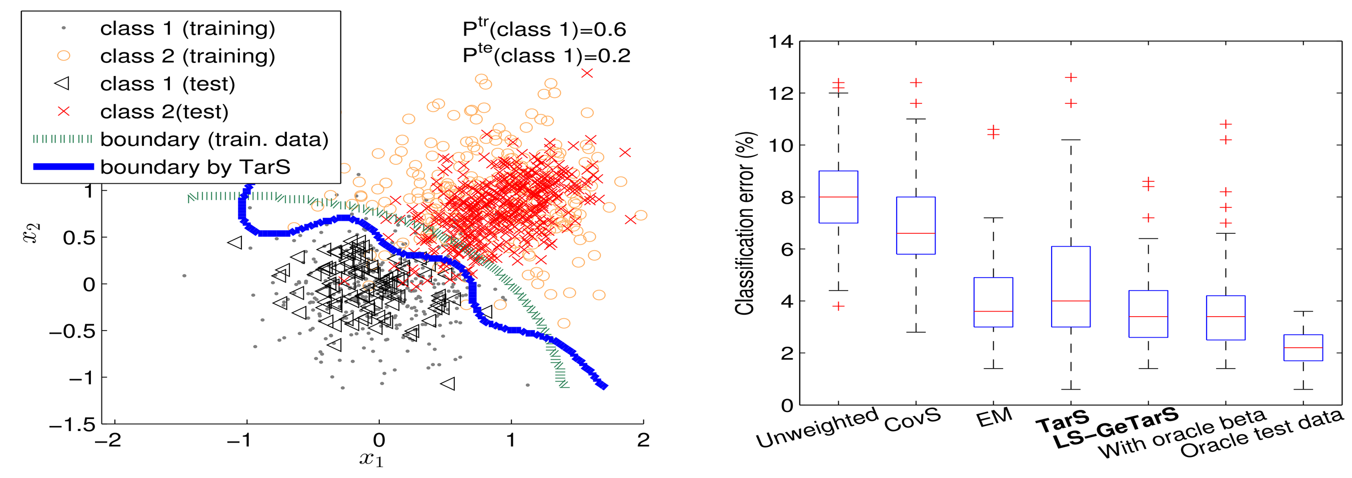
(b) Classification under TarS



(c) Classification under LS-GeTarS

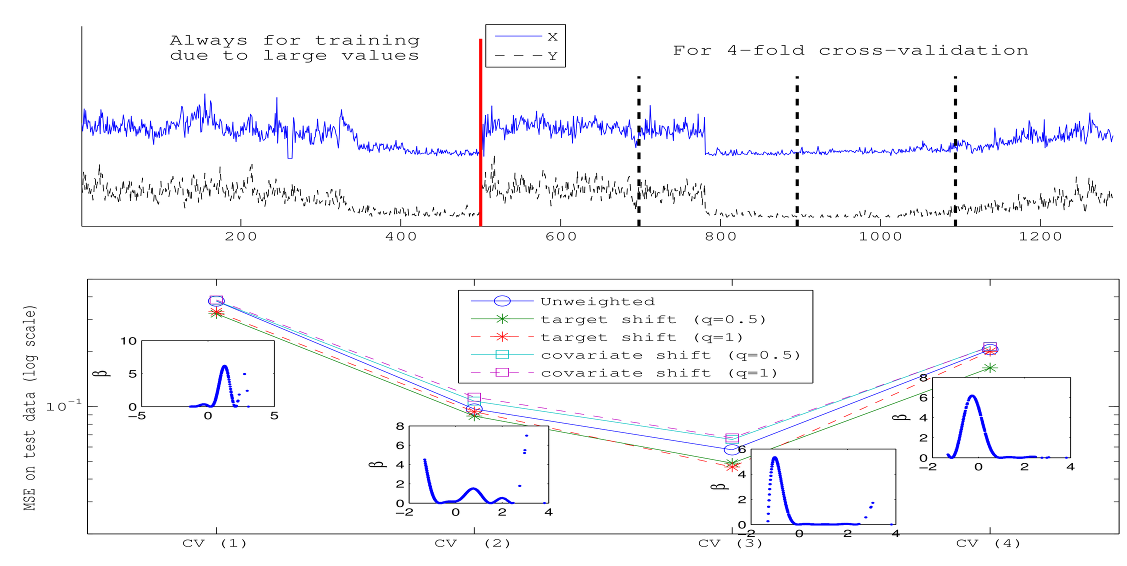


(d) Classification under non-location-scale GeTarS



Real-world problems

- **Regression under TarS**:
  - ⋆ Cause-effect pair *48*: time series Y (# open http connections) → X (# bytes sent by the computer), with a strong dependence.
  - ⋆ Correcting TarS improves prediction performance for Y. ◡̈
  - ⋆ No improvement for predicting X from Y.



- **Remote sensing image classification**:

  Two data sets collected on two different and spatially disjoint areas; the sample on each area was partitioned into TR and TS.

Figure 6: A misclassification rate on remote sensing data set with different distribution shift correction schemes.

Problem	Unweight	CovS	TarS	LS-GeTarS
$TR_1 \to TS_2$	20.73%	20.73%	20.41%	**11.96%**
$TR_2 \to TS_1$	26.36%	25.32%	26.28%	**13.56%**

Conclusions

- TarS and GeTarS: a convenient way to deal with the situation where both conditional and marginal distributions change across domains; why prefer $P_{XY} = P_Y P_{X|Y}$?
- Background (causal) information helps learning: compact description of how distributions change. ◡̈

Refenrence: [1] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf, Correcting sample selection bias by unlabeled data. In *NIPS 19*, 2008.