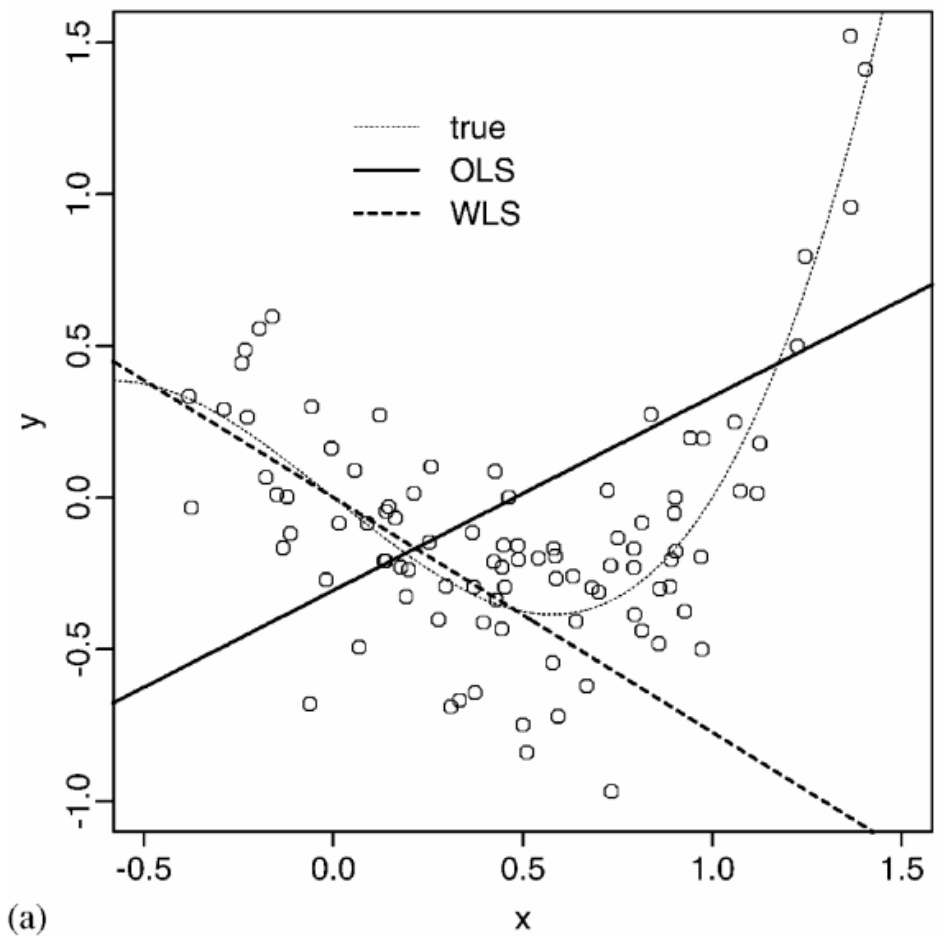
Stable Prediction with Model Misspecification and Agnostic Distribution Shift

Motivation

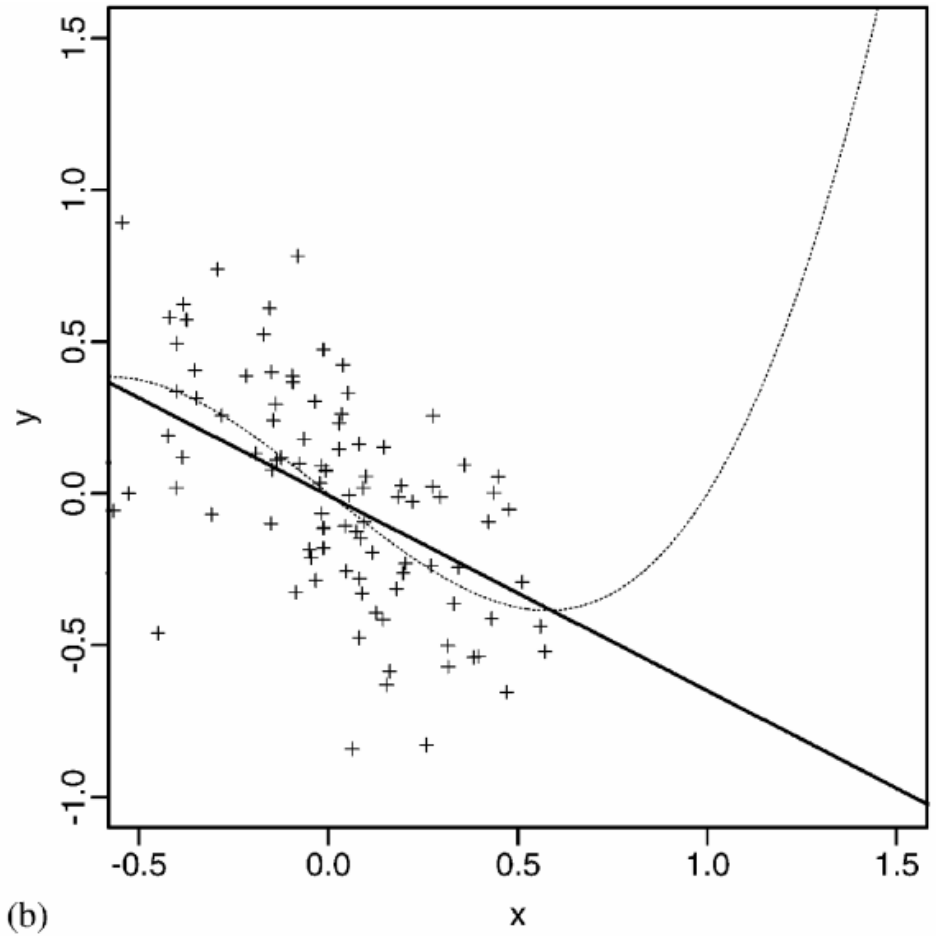
In real applications, however, we rarely know the underlying true model for prediction,
and we cannot guarantee the unknown test data will have the same distribution as the training data

For example

- different geographies, schools, or hospitals may draw from different demographics, and the correlation structure among demographics may also vary (e.g. one ethnic group may be more or less disadvantaged in different geographies)



(a)



(b)

$$\min_{W,\beta} \sum_{i=1}^{n} W_i \cdot (Y_i - \mathbf{X}_i, \beta)^2$$

$$\frac{q_1(x)}{q_0(x)} = \frac{\exp(-(x-\mu_1)^2/2\tau_1^2)/\tau_1}{\exp(-(x-\mu_0)^2/2\tau_0^2)/\tau_0} \propto \exp\left(-\frac{(x-\bar{\mu})^2}{2\bar{\tau}^2}\right),$$

PROBLEM 1. ***(Stable Learning)*** *: Given the target value y and p input variables $x = [x_1, \ldots, x_p] \in \mathbb{R}^p$, the task is to learn a predictive model which can achieve* ***uniformly*** *small error on* ***any*** *data point.*

ASSUMPTION 1. *There exists a decompsition of all the variables* $\mathbf{X} = \{\mathbf{S}, \mathbf{V}\}$, *where* ***S*** *represents the stable variable set and* ***V*** *represents the unstable variable set. Specifically, for all environments* $e \in \mathcal{E}, \mathbb{E}(Y^e|\mathbf{S}^e = s, \mathbf{V}^e = v) = \mathbb{E}(Y^e|\mathbf{S}^e = s) = \mathbb{E}(Y|\mathbf{S} = s)$[1].

$$Y^e = f(\mathbf{S}^e) + \mathbf{V}^e\beta_V + \epsilon^e = \mathbf{S}^e\beta_S + g(\mathbf{S}^e) + \mathbf{V}^e\beta_V + \varepsilon^e.$$

$$\mathcal{L}_{OLS} = \sum_{i=1}^{n} \left(\mathbf{S}_i^T \beta_S + \mathbf{V}_i^T \beta_V - Y_i \right)^2 .$$

$$\begin{aligned}
\hat{\beta}_{V_{OLS}} &= \beta_V + (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{V}_i^T \mathbf{V}_i)^{-1} (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{V}_i^T g(\mathbf{S}_i)) \\
&+ (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{V}_i^T \mathbf{V}_i)^{-1} (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{V}_i^T \mathbf{S}_i)(\beta_S - \hat{\beta}_{S_{OLS}}), \quad (4) \\
\hat{\beta}_{S_{OLS}} &= \beta_S + (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{S}_i^T \mathbf{S}_i)^{-1} (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{S}_i^T g(\mathbf{S}_i)) \\
&+ (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{S}_i^T \mathbf{S}_i)^{-1} (\tfrac{1}{n} \sum_{i=1}^{n} \mathbf{S}_i^T \mathbf{V}_i)(\beta_V - \hat{\beta}_{V_{OLS}}), \quad (5)
\end{aligned}$$

where n is sample size, $\frac{1}{n} \sum_{i=1}^{n} \mathbf{V}_i^T g(\mathbf{S}_i) = \mathbb{E}(\mathbf{V}^T g(\mathbf{S})) + o_p(1)$ and $\frac{1}{n} \sum_{i=1}^{n} \mathbf{V}_i^T \mathbf{S}_i = \mathbb{E}(\mathbf{V}^T \mathbf{S}) + o_p(1)$. To simplify notation, we remove the environment variable e from $\mathbf{X}^e$, $\mathbf{S}^e$, $\mathbf{V}^e$, ε^e.

Proposition 1 *If* $\mathbf{X}$ *are mutually independent with mean 0, then* $\mathbb{E}(\mathbf{V}^T g(\mathbf{S})) = 0$ *and* $\mathbb{E}(\mathbf{V}^T \mathbf{S}) = 0$.

$$\min_{W} \sum_{a=1}^{\infty} \sum_{b=1}^{\infty} \|\mathbb{E}[\mathbf{X}_{,j}^{a^T} \mathbf{\Sigma}_W \mathbf{X}_{,k}^{b}] - \mathbb{E}[\mathbf{X}_{,j}^{a^T} W]\mathbb{E}[\mathbf{X}_{,k}^{b^T} W]\|_2^2,$$

$$W^b = \arg\min_{W} \sum_{j=1}^{p} \left\|\mathbb{E}[\mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}] - \mathbb{E}[\mathbf{X}_{,j}^T W]\mathbb{E}[\mathbf{X}_{,-j}^T W]\right\|_2^2$$

With $\sum_{i=1}^{n} W_i = n$, we can denote the loss in Eq. (7) as:

$$\mathcal{L}_B = \sum_{j=1}^{p} \left\|\mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^T W/n \cdot \mathbf{X}_{,-j}^T W/n\right\|_2^2. \quad (8)$$

$$\hat{W} = \arg\min_{W \in \mathcal{C}} \mathcal{L}_B + \frac{\lambda_3}{n} \sum_{i=1}^{n} W_i^2 + \lambda_4 (\frac{1}{n} \sum_{i=1}^{n} W_i - 1)^2,$$

Theorem 1 *The solution $\hat{W}$ defined in Eq. (10) is unique if $\lambda_3 n \gg p^2 + \lambda_4$, $p^2 \gg \max(\lambda_3, \lambda_4)$ and $|\mathbf{X}_{i,j}| \leq c$ for some constant c.*

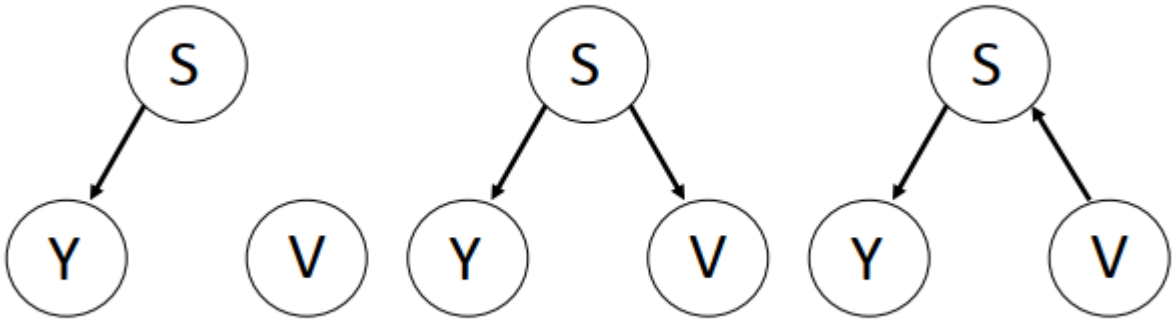
$$\hat{\beta}_{WLS} = \arg\min_{\beta} \sum_{i=1}^{n} \hat{W}_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2.$$

$$\begin{aligned} & \min_{W,\beta} \textstyle\sum_{i=1}^{n} W_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2 \qquad (12) \\ s.t \quad & \textstyle\sum_{j=1}^{p} \left\| \mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^T W/n \cdot \mathbf{X}_{,-j}^T W/n \right\|_2^2 < \lambda_2 \\ & |\beta|_1 < \lambda_1, \ \ \tfrac{1}{n}\textstyle\sum_{i=1}^{n} W_i^2 < \lambda_3, \\ & (\frac{1}{n}\sum_{i=1}^{n} W_i - 1)^2 < \lambda_4, \ \ W \succeq 0, \end{aligned}$$

Algorithm 1 Decorrelated Weighted Regression algorithm

Require: Observed features $\mathbf{X}$ and outcome variable Y.
Ensure: Updated parameters W, β.

Initialize parameters $W^{(0)}$ and $\beta^{(0)}$,
Calculate loss function with parameters $(W^{(0)}, \beta^{(0)})$,
Initialize the iteration variable $t \leftarrow 0$,
repeat
    $t \leftarrow t + 1$,
    Update $W^{(t)}$ with gradient descent by fixing β,
    Update $\beta^{(t)}$ with gradient descent by fixing W,
    Calculate loss function with parameters $(W^{(t)}, \beta^{(t)})$,
until Loss function converges or max iteration is reached.
return W, β.



(a) $\mathbf{S} \perp \mathbf{V}$ (b) $\mathbf{S} \to \mathbf{V}$ (c) $\mathbf{V} \to \mathbf{S}$

$\mathbf{S} \perp \mathbf{V}$:

$$\mathbf{Z}_{,1}, \cdots, \mathbf{Z}_{,p} \overset{iid}{\sim} \mathcal{N}(0,1), \mathbf{V}_{,1}, \cdots, \mathbf{V}_{,p_v} \overset{iid}{\sim} \mathcal{N}(0,1)$$
$$\mathbf{S}_{,i} = 0.8 * \mathbf{Z}_{,i} + 0.2 * \mathbf{Z}_{,i+1}, \ i = 1, 2, \cdots, p_s,$$

$\mathbf{S} \to \mathbf{V}$:

$$\mathbf{V}_{,j} = 0.8 * \mathbf{S}_{,j} + 0.2 * \mathbf{S}_{,j+1} + \mathcal{N}(0,1)$$

$\mathbf{V} \to \mathbf{S}$:

$$\mathbf{S}_{,j} = 0.2 * \mathbf{V}_{,j} + 0.8 * \mathbf{V}_{,j+1} + \mathcal{N}(0,1)$$

Stable Learning via Differentiated Variable Decorrelation

Motivation

- For example, when you want to recognize a dog in image classification task, although the nose, ear and mouth of dog may be represented by different variables, they act as an integrated whole and such correlations are stable across different envinronments.

Algorithm 1 Differentiated Variable Decorrelation (DVD)

Input: Unlabeled heterogeneous data $\mathbf{Z} = [\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^M]$ and labeled homogeneous data $\mathbf{D} = [\mathbf{X}, \mathbf{Y}]$.

Output: Clustering results $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_k$ and sample weight W.

Variable clustering:
Calculate the variable dissimilarity vector F by Equ.9.
Initialize k cluster means $m_1, m_2, \ldots, m_k$.
repeat
    Assignment step: Assign each variable to the cluster with the nearest mean measured by least squared Euclidean distance.
    Update step: Recalculate means for variables assigned to each cluster.
until The assignment result $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_k$ no longer changes.
Variable decorrelation weight learning:
Initialize parameters $W^{(0)}$,
Calculate value of Obj. (11) with parameters $W^{(0)}$ and $\alpha^{(t)}$,
Initialize the iteration variable $q \leftarrow 0$,
repeat
    $q \leftarrow q + 1$,
    Update $W^{(q)}$ by gradient descent,
    Calculate loss function with parameters $W^{(q)}$,
until Loss function converges or max iteration is reached.
return W

Stable Learning via Sample Reweighting

Algorithm 1 Sample Reweighted Decorrelation Operator (SRDO)

Require: Design Matrix $\mathbf{X}$

1: **for** $i = 1 \dots n$ **do**
2: Initialize a new sample $\tilde{x}_i \in \mathbb{R}^p$ with empty vector
3: **for** $j = 1 \dots p$ **do**
4: Draw the j^{th} feature of new sample $\tilde{x}_{i,j}$ from $\mathbf{X}_{,j}$ at random
5: **end for**
6: **end for**
7: Set $\tilde{x}_i$ as positive samples and x_i as negative samples, then train a binary classifier.
8: Set $w(x) = \frac{p(Z=1|x)}{p(Z=0|x)}$ for each sample x_i in $\mathbf{X}$, where $p(Z = 1|x)$ is the probability of sample x been drawn from $\tilde{D}$ estimated by the trained classifier.

Ensure: A set of sample weights $w(x)$ which can deccorelate $\mathbf{X}$