Environment Invariant Linear Least Squares, Fan *et al.* (2023)

Xiaohui Yin

University of Connecticut

October 7, 2024

Motivating Example: Risks of Using Spurious Variables

A task to classify cows and camels based on extracted hierarchical features:

- There is a dataset $\mathcal{D}$ containing 10k images of cows and camels from the Internet. When Using 70% to train the classifier, top two features are the back shape x_1, and the background color x_2, and it performs well on the remaining 30% testing set. It is contemplated that cows often appear on the grass while most camels appear on the sand from this dataset.

Figure: Camel in brown

Figure: Cow in green

Motivating Example (Cond)

- However, introducing x_2 is not what we expected: the classifier can not work well in a place farming camels and cows, in which the background color is fixed.
- If we have another dataset $\widetilde{\mathcal{D}}$, in which the association between the background color and object label still exists yet slightly perturbs. Intuitively, we may infer that x_2 may be a "spurious" variable for prediction or causation.

Figure: Camel in green

Figure: Cow in brown

Some Definition

This paper propose the multiple-environment version of linear least squares. It utilizes the heterogeneity across datasets. It combines the linear least squares solutions across different datasets and uses their differences to determine the true parameter and important variables in a completely data-driven manner.

Here are some definitions needed:

- response variable $y \in \mathbb{R}$
- explanatory variable $\mathbf{x} \in \mathbb{R}^p$
- set of environments $\mathcal{E}$
- observations from environment $e \in \mathcal{E}$: $(\mathbf{x}_1^{(e)}, y_1^{(e)}), \ldots, (\mathbf{x}_n^{(e)}, y_n^{(e)}) \sim \mu^{(e)}$

Model Formulation

The true linear model is assumed as

$$y^{(e)} = (\beta^*_{S^*})^\top \mathbf{x}^{(e)}_{S^*} + \varepsilon^{(e)} \qquad \text{with} \qquad \mathbb{E}[\varepsilon^{(e)} | \mathbf{x}^{(e)}_{S^*}] \equiv 0, \tag{1}$$

where the unknown set of important variables $S^* = \{j : \beta^*_j \neq 0\}$ and the model parameters β^* are the same across different environments, while $\mu^{(e)}$, the distribution of $(\mathbf{x}^{(e)}, y^{(e)})$, may vary.

The aim is to estimate β^* and S^* using the $n \cdot |\mathcal{E}|$ data $\{(\mathbf{x}^{(e)}_i, y^{(e)}_i)\}_{e \in \mathcal{E}, i \in \{1,\ldots,n\}}$.

Challenges

- $\boldsymbol{\beta}^*$ is mot the best linear predictor for each single environment.
- A smaller mean squared error by incorporating the linear spurious variables $\boldsymbol{x}_{S^{*\perp}}$ defined outside the set of important variables for all environments. $\boldsymbol{x}_{S^{*\perp}}^{(e)}$ can capture information in $\varepsilon^{(e)}$.
- Such spurious variables are not stable when generalized to other environments, because the association between these variables and y are not stable. (background color in the motivating example)

Loss Function

The population-level EILLS objective is

$$\mathrm{Q}_\gamma(\boldsymbol{\beta}) = \underbrace{\sum_{e\in\mathcal{E}} \mathbb{E}\left[|y^{(e)} - \boldsymbol{\beta}^\top \mathbf{x}^{(e)}|^2\right]}_{\mathrm{R}(\boldsymbol{\beta})} + \underbrace{\gamma \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\} \times \sum_{e\in\mathcal{E}} \left|\mathbb{E}[(y^{(e)} - \boldsymbol{\beta}^\top \mathbf{x}^{(e)}) x_j^{(e)}]\right|^2}_{\mathrm{J}(\boldsymbol{\beta})}, \tag{2}$$

where γ is a tuning parameter.

- $\mathrm{R}(\boldsymbol{\beta})$ requires a good overall solution in each single environment.
- $\mathrm{J}(\boldsymbol{\beta})$ discourages selecting variables that have strong correlation with the fitted residuals in some environments.
- From computational perspective, we have

$$\mathrm{J}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\} \sum_{e\in\mathcal{E}} \frac{1}{4} |\nabla_j \mathrm{R}^{(e)}(\boldsymbol{\beta})|^2, \tag{3}$$

Empirical Loss Function

$$\hat{\mathsf{J}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\} \sum_{e \in \mathcal{E}} \left| \hat{\mathbb{E}}[x_j^{(e)}(y^{(e)} - \boldsymbol{\beta}^\top \mathbf{x}^{(e)})] \right|^2$$

$$\hat{\mathsf{R}}(\boldsymbol{\beta}) = \sum_{e \in \mathcal{E}} \frac{1}{n^{(e)}} \sum_{i=1}^{n^{(e)}} \{y_i^{(e)} - \boldsymbol{\beta}^\top \mathbf{x}_i^{(e)}\}^2$$

$$\hat{\mathsf{Q}}(\boldsymbol{\beta}, \gamma) = \hat{\mathsf{R}}(\boldsymbol{\beta}) + \lambda \hat{\mathsf{J}}(\boldsymbol{\beta})$$

Some linear exogenous variables which do not contribute to explaining y can be eliminated further by introducing an l_0 penalty as

$$\hat{\mathsf{L}}(\boldsymbol{\beta}, \gamma, \lambda) = \hat{\mathsf{R}}(\boldsymbol{\beta}) + \lambda \hat{\mathsf{J}}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_0.$$

An Illustration

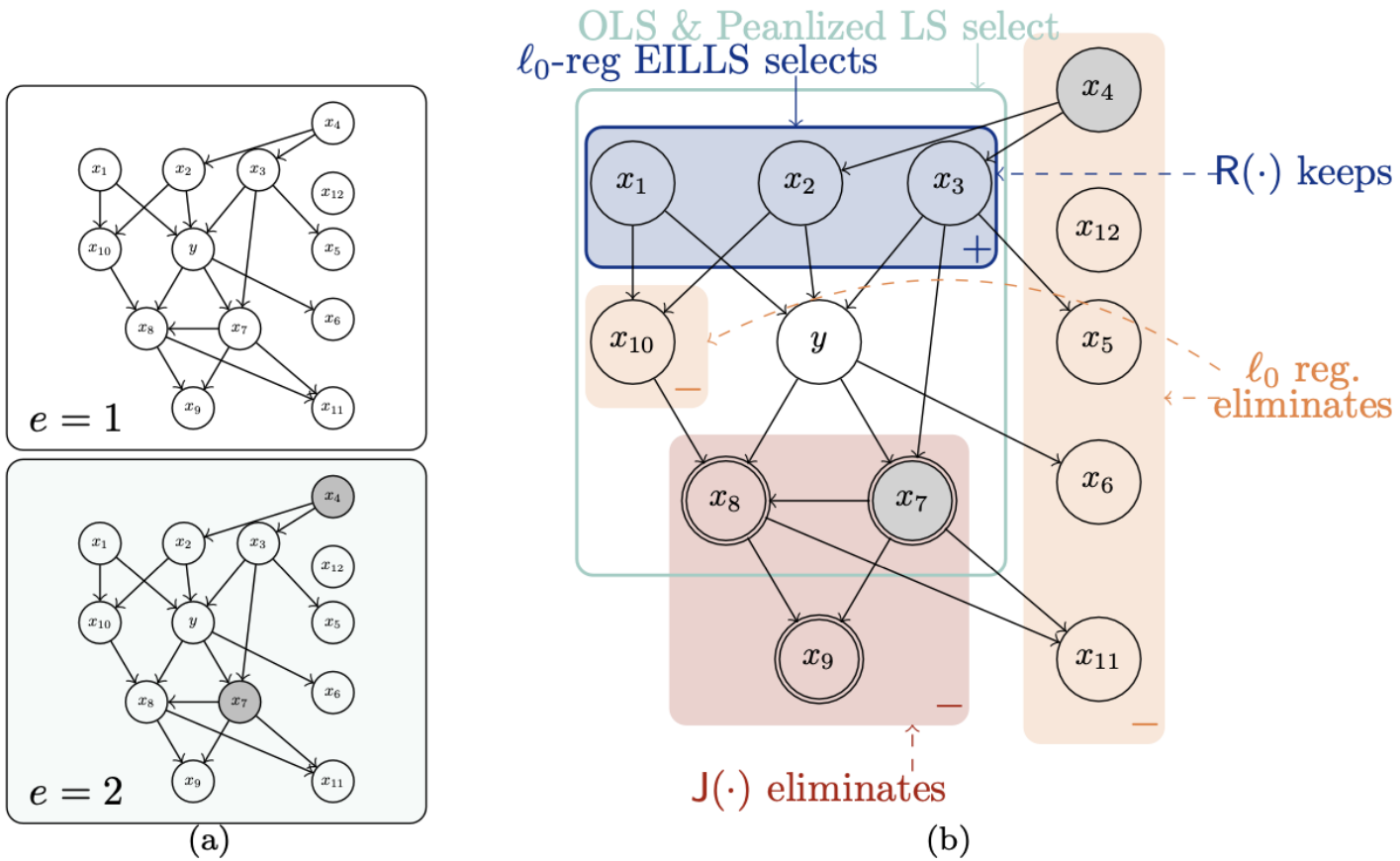

Figure: An Example for Illustration, figure from Fan *et al.* (2023)

References I

Fan, J., Fang, C., Gu, Y. and Zhang, T. (2023) Environment invariant linear least squares. URL https://arxiv.org/abs/2303.03092.