

From ICSDS @ Nice, France December, 2024

Shifts in distribution, covariates, prior probability, etc

1. Huang MY, Qing J, Huang CY (2024) ``Efficient Data Integration Under Prior Probability Shift,” Biometrics, 2024, Mar 27, 80(2).

<https://pubmed.ncbi.nlm.nih.gov/38768225/>

<https://profiles.ucsf.edu/chiung-yu.huang>

Distribution-free prediction intervals under covariate shift, with an application to causal inference. Journal of the American Statistical Association. 2024. Qin J, Liu Y, Li M, **Huang CY**.

2. Ying Jin, Naoki Egami, and Dominik Rothenhausler (2024) “Beyond Reweighting: On the Predictive Role of Covariate Shift in Effect Generalization”

<https://arxiv.org/pdf/2412.08869>

<https://sites.google.com/view/rothenhaeusler/home>

Some slides for the first talk:

Huang MY, Qing J, **Huang CY** (2024) ``Efficient Data Integration Under Prior Probability Shift,” Biometrics, 2024, Mar 27, 80(2).

The image shows two presentation slides side-by-side. The left slide is titled "Why Study Dataset Shift?" and the right slide is titled "Types of Dataset Shifts".

Why Study Dataset Shift?

- ▶ Models are often trained on specific datasets. When deployed in settings different from the training environment, dataset shifts can significantly degrade model performance.
- ▶ Causes of dataset shift:
 - ▶ Selection bias
 - ▶ Non-stationary environments: temporal or spatial
 - ▶ Real-world dynamics: behavior change, technological breakthroughs

2 / 22

Types of Dataset Shifts

- ▶ Covariate shift
 - $P(Y|X)$ remains constant, but $P(X)$ varies
 - Self-driving car trained on sunny streets, tested in rainy conditions
- ▶ Prior probability shift (label shift)
 - $P(X|Y)$ remains constant, but $P(Y)$ varies
 - Spam filter trained on 50% spam, 50% non-spam; deployed where 10% of emails are spam
- ▶ Concept shift
 - Relationship between X and Y changes
 - Credit default prediction during economic changes affecting income-default relationship

3 / 22

General Dataset Shift Models

- ▶ Notation
 - $(Y_1, X_1) \sim f_1(y, x)$ from training data
 - $(Y_2, X_2) \sim f_2(y, x)$ from testing data
- ▶ A general dataset shift model

$$f_2(y, x) = \frac{w(y, x)f_1(y, x)}{\int \int w(u, v)f_1(u, v)dudv}$$

or, equivalently, $f_2(y, x) \propto w(y, x)f_1(y, x)$

- ▶ $w(y, x)$ can be viewed as sampling weight function
 - Covariate shift: $w(y, x) \equiv w(x) \Rightarrow f_1(y | x) = f_2(y | x)$
 - Prior probability shift: $w(y, x) \equiv w(y) \Rightarrow f_1(x | y) = f_2(x | y)$
 $f_1(y | x) \neq f_2(y | x)$
 - Concept shift: e.g. $w(y, x) = w_1(y)w_2(x)$;
 $w(y, x) = w(y, x; \gamma)$

4 / 22

Data and Model Setup

- ▶ Two datasets
 - $\mathcal{D}_1 = \{(X_{1i}, Y_{1i}) : i = 1, \dots, n_1\}$
 - $\mathcal{D}_2 = \{(X_{2i}, Y_{2i}) : i = 1, \dots, n_2\}$
- ▶ Assume prior probability shift between \mathcal{D}_1 and \mathcal{D}_2 :
 - $f_1(x | y) = f_2(x | y)$
- ▶ Assumed parametric model for \mathcal{D}_1 : $f(y | x; \theta)$
- ▶ Consider efficient estimation of θ under prior probability shift with $\mathcal{D}_1 \cup \mathcal{D}_2$

5 / 22

Likelihood Under Prior Probability Shift

- ▶ Suppose $X_1 \sim dG_1(x)$ and $Y_2 \sim F_2(y)$
- ▶ Apply Bayes Rule and under prior probability shift

$$f_1(x | y) = \frac{f_1(y | x; \theta)dG_1(x)}{\int f_1(y | u; \theta)dG_1(u)}$$

$$\begin{aligned} f_2(y, x) &= f_2(x | y)dF_2(y) \\ &= \frac{f_1(y | x; \theta)dG_1(x)}{\int f_1(y | u; \theta)dG_1(u)}dF_2(y) \end{aligned}$$

6 / 22

Maximum Likelihood Estimation

- ▶ The MLE of F_2 is the empirical distribution, so $\hat{q}_i = n_2^{-1}$.
- ▶ Profile likelihood: for each fixed θ , write

$$\hat{p}(\theta) = \operatorname{argmax}_{p \in \mathcal{P}_n} \ell(\theta, p, \hat{q})$$

- ▶ Lagrange function:

$$\begin{aligned} \mathcal{L}(p, \eta) &= \sum_{i=1}^n \log f(Y_i | X_i; \theta) + \sum_{i=1}^n \log p_i \\ &\quad - \sum_{i=1}^{n_2} \log \sum_{j=1}^n f(Y_{2i} | X_j; \theta)p_j + \eta \left(\sum_{i=1}^n p_i - 1 \right), \end{aligned}$$

η is the Lagrange multiplier.

9 / 22

Profile Likelihood Estimation

- ▶ For fixed θ , $\hat{p}(\theta)$ solves

$$p_i = \left\{ n_1 + \sum_{j=1}^{n_2} \frac{f(Y_{2j} | X_i; \theta)}{\sum_{k=1}^n f(Y_{2j} | X_k; \theta)p_k} \right\}^{-1}, \quad i = 1, \dots, n.$$

- ▶ The equations can be rewritten as $p = T(p)$;
- ▶ Following a standard argument of fixed-point theory, the root can be obtained by iteratively computing $T(p)$ with a proper initial point.

Prediction

Y_k^{new} : outcome corresponding to a new subject with X_k^{new}

- ▶ Predict continuous response:

$$\hat{Y}_1 = \int y f(y | X_1^{\text{new}}; \hat{\theta}) dy \quad \text{and} \quad \hat{Y}_2 = \frac{\int y f(y | X_2^{\text{new}}; \hat{\theta}) d\hat{R}(y)}{\int f(y | X_2^{\text{new}}; \hat{\theta}) d\hat{R}(y)},$$

where $d\hat{R}(y) = d\hat{F}_2(y) / \int f(y | x; \hat{\theta}) d\hat{G}_1(x)$.

- ▶ Predict discrete response:

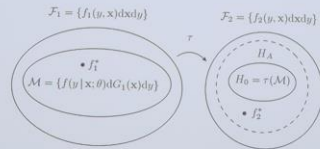
$$\begin{aligned} \hat{Y}_1 &= \operatorname{argmax}_{y \in \{y_1, \dots, y_K\}} f(y | X_1^{\text{new}}; \hat{\theta}) \quad \text{and} \\ \hat{Y}_2 &= \operatorname{argmax}_{y \in \{y_1, \dots, y_K\}} \frac{f(y | X_2^{\text{new}}; \hat{\theta}) \Delta \hat{R}(y)}{\sum_{k=1}^K f(y_k | X_2^{\text{new}}; \hat{\theta}) \Delta \hat{R}(y_k)}, \end{aligned}$$

where $\Delta \hat{R}(y) = \Delta \hat{F}_2(y) / \int f(y | x; \hat{\theta}) d\hat{G}_1(x)$.

12 / 22

Testing Prior Probability Shift

- ▶ $H_0 : f_1(\mathbf{x} | y) \equiv f_2(\mathbf{x} | y)$
- ▶ Under H_0 , the joint density of (Y_2, \mathbf{X}_2) is given by $f_2(y, \mathbf{x})d\mathbf{y} = f_1(y, \mathbf{x})dF_2(y)/\int f_1(y, \mathbf{u})d\mathbf{u}$.
- ▶ The prior probability shift assumption defines a map τ from \mathcal{F}_1 to \mathcal{F}_2 , where $\tau(f_1) = f_1(y, \mathbf{x})dF_2(y)/\int f_1(y, \mathbf{u})d\mathbf{u}$.



13 / 22

Methods for Comparison

- ▶ MLE using $\mathcal{D}_1 \cup \mathcal{D}_2$, thereby ignoring the shift in distribution
- ▶ The plug-in estimator (Saerens et al. 2002)
 - Using \mathcal{D}_1 to obtain

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i=1}^{n_1} f(Y_{1i} | \mathbf{X}_{1i}; \theta)$$

- Replace θ with $\hat{\theta}$ in

$$\frac{f(y | \mathbf{X}_2; \hat{\theta})dR(y)}{\int f(s | \mathbf{X}_2; \hat{\theta})dR(s)}$$

where $dR(y) = dF_2(y)/\int f(y | \mathbf{x}; \theta)dG_1(\mathbf{x})$ and G_1 is the marginal distribution of \mathbf{X}_1 .

Saerens, Latinne & Decaestecker (Neural Computation, 2002)

18 / 22

Discussions

- ▶ Works for both categorical and continuous outcomes; Efficiently combines information from multiple sources under prior probability shift.
- ▶ LRT for checking prior probability shift.
- ▶ Variable selection can be performed using regularization techniques such as LASSO and SCAD
- ▶ Only parametric models are considered; however, the profile likelihood approach allows accommodation of general semiparametric frameworks, such as single-index models and dimension reduction.

Ref: Huang MY, Qin J, Huang CY. Efficient data integration under prior probability shift. Biometrics. 2024 Mar 27; 80(2).

22 / 22

Some Slides from the 2nd talk

