# Semi-Supervised Inference: General Theory and Estimation of Means

**Anru Zhang**
Department of Statistics
University of Wisconsin-Madison

Workshop in Honor of Larry Brown

Joint work with Larry Brown and Tony Cai

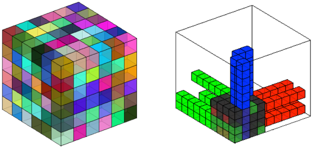
Nov 30, 2018

# In Memory of Larry

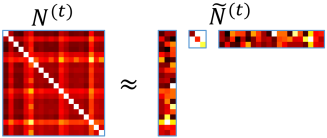

Figure: Anru's PhD Thesis Defense, April, 2015

# My Recent Research

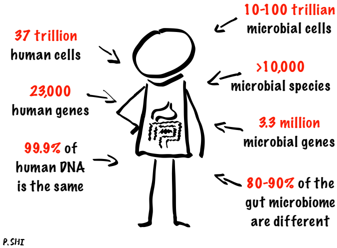
- Tensor Data Analysis



- Singular Subspace Analysis, PCA

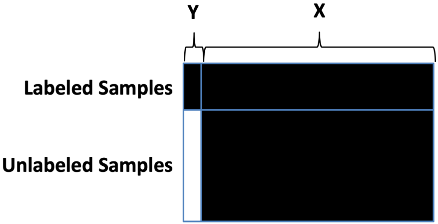



- Human Microbiome Studies



# Semi-supervised Inference

- Semi-supervised settings often appear in machine learning and statistics.




- Possible situations: labels are more difficult or expensive to acquire than unlabeled data.

- Example:
  - Survey sampling
  - Electronic health record
  - Imaging classification
  - ...

# An "Assumption Lean" Framework

- Assume $Y$ is label, $X = (X_1, \ldots, X_p)$ is $p$-dimensional covariate,

$$(Y, X_1, \ldots, X_p) \sim P = P(dy, dx_1, \ldots, dx_p).$$

  No specific assumption on the relationship between $Y$ and $X$.

- Observations:

  → $n$ "labeled" samples from joint distribution $P$,

$$[\boldsymbol{Y}, \boldsymbol{X}] = \left\{Y_k, X_{k1}, \ldots, X_{kp}\right\}_{k=1}^{n};$$

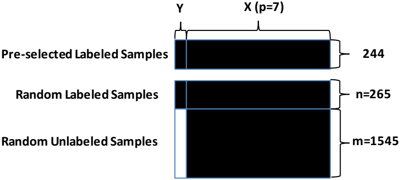
  → $m$ "unlabeled" samples from marginal distribution $P_X$,

$$\boldsymbol{X}_{add} = \left\{X_{k1}, \ldots, X_{kp}\right\}_{k=n+1}^{n+m}.$$

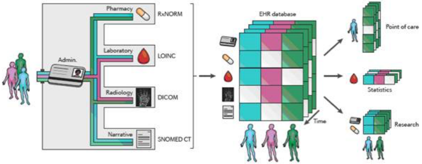
- Goal: statistical inference for $\theta = \mathbb{E}Y$.

# Motivations

- Consensus of Homeless




- Electronic Health Records: prevalence of certain disease




Picture source: Jensen PB, Jensen LJ, and Brunak S. Nature Reviews, 2012

# $m = \infty$: Ideal Semi-Supervised Inference

- $m = \infty$, infinitely many unlabeled samples.
- Baseline estimator: sample mean $\bar{\boldsymbol{Y}}$.
- Least square estimator:

$$\hat{\theta}_{LS} = \bar{\boldsymbol{Y}} - \hat{\beta}_{(2)}^{\top}(\bar{\boldsymbol{X}} - \mu).$$

  - $\mu = \mathbb{E}X$ is known;
  - $\bar{\boldsymbol{Y}} = \frac{1}{n}\sum_{k=1}^{n} Y_k$, $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{k=1}^{n} X_k$;
  - $\hat{\beta} = \left(\vec{\boldsymbol{X}}^{\top}\vec{\boldsymbol{X}}\right)^{-1}\vec{\boldsymbol{X}}^{\top}\boldsymbol{Y}$ is the least square estimator, $\hat{\beta} = [\hat{\beta}_1\ \hat{\beta}_{(2)}^{\top}]^{\top}$;

$$\vec{\boldsymbol{X}} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$$

is the prediction matrix with intercepts;

# $m < \infty$: Ordinary Semi-Supervised Inference

- $m < \infty$: finitely many unlabeled samples; $P_X$ is partially known.
- Semi-supervised least squared estimator

$$\hat{\theta}_{SSLS} = \bar{\boldsymbol{Y}} - \hat{\beta}_{(2)}^{\top}(\bar{\boldsymbol{X}} - \hat{\mu}), \quad \hat{\mu} = \frac{1}{n+m} \sum_{k=1}^{n+m} X_k.$$

- When $m = 0$, i.e., no unlabeled samples,

$$\hat{\theta}_{SSLS} = \bar{\boldsymbol{Y}};$$

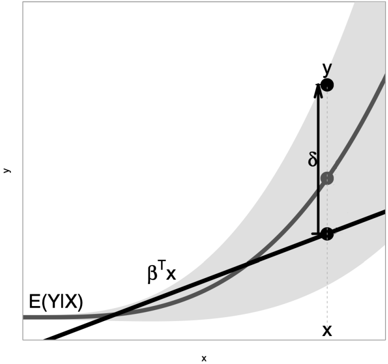
When $m = \infty$, i.e., infinitely many unlabeled samples,

$$\hat{\theta}_{SSLS} = \hat{\theta}_{LS}.$$

# Interpretation: An Assumption-Lean Framework

Define

- population slopes: $\beta = \text{argmin}_\gamma \mathbb{E}(Y - \vec{X}^\top \gamma)^2$;
- linear deviations $\delta = Y - \beta^\top \vec{X}$, $\tau^2 = \mathbb{E}\delta^2$.




Picture source: Buja, Berk, Brown, George, Pitkin, Traskin, Zhao, and Zhang, Statistical Science, 2017.

# Interpretation: An Assumption-Lean Framework

- Facts:

$$\theta = \beta_1 + \mu^\top \beta_{(2)}, \quad \hat{\theta}_{LS} = \hat{\beta}_1 + \mu^\top \hat{\beta}_{(2)}, \quad \hat{\theta}_{SSLS} = \hat{\beta}_1 + \hat{\mu}^\top \hat{\beta}_{(2)}.$$

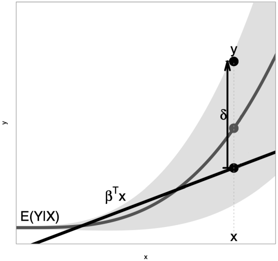
- Thus, $\hat{\theta}_{LS}$ and $\hat{\theta}_{SSLS}$ can be seen as "plug-in" estimators:

$$\beta = \operatorname*{argmin}_{\gamma} \mathbb{E}(Y - \vec{X}^\top \gamma)^2, \quad \hat{\beta} = \operatorname*{argmin}_{\gamma} \sum_{k=1}^{n} (Y_k - \vec{X}_k \gamma)^2,$$

$$\mu = \mathbb{E}X, \quad \hat{\mu} = \frac{1}{n+m} \sum_{k=1}^{n+m} X_k.$$

# Theory: $\ell_2$ risks

- Recall
  - population slopes $\beta = \mathrm{argmin}_\gamma \mathbb{E}(Y - \vec{X}^\top \gamma)^2$, $\beta = [\beta_1 \beta_{(2)}^\top]^\top$;
  - Linear deviations $\delta = Y - \beta^\top \vec{X}$;
  - $\tau^2 = \mathbb{E}\delta^2$, $\mu = \mathbb{E}X$, $\Sigma = \mathrm{Cov}(X)$.




## Proposition ($\ell_2$ risk of $\bar{Y}$)

$$n\mathbb{E}(\bar{Y} - \theta)^2 = \tau^2 + \beta_{(2)}^\top \Sigma \beta_{(2)}.$$

# Theory: $\ell_2$ risks

**Theorem ($\ell_2$ risk of $\hat{\theta}_{LS}$)**

*Suppose we observe $n$ labeled samples and know $P_X$, $p = o(n^{1/2})$, $\hat{\theta}^1_{LS}$ is a truncation version of $\hat{\theta}_{LS}$. Under finite moment conditions, we have*

$$n\mathbb{E}\left(\hat{\theta}^1_{LS} - \theta\right)^2 = \tau^2 + s_n, \quad s_n = O(p^2/n).$$

**Theorem ($\ell_2$ risk of $\hat{\theta}_{SSLS}$)**

*Suppose we observe $n$ labeled samples $\{Y_k, X_k\}_{k=1}^n$ and $m$ unlabeled samples $\{X_k\}_{k=n+1}^{n+m}$, $p = o(n^{1/2})$, $\hat{\theta}^1_{SSLS}$ is a truncation version of $\hat{\theta}_{SSLS}$. Under finite moment conditions, we have*

$$n\mathbb{E}\left(\hat{\theta}^1_{SSLS} - \theta\right)^2 = \tau^2 + \frac{n}{n+m}\beta_{(2)}^\top \Sigma \beta_{(2)} + s_{n,m}, \quad s_{n,m} = O(p^2/n).$$

# Remark: $\ell_2$ Risk Theory

$$n\mathbb{E}\left(\bar{Y}-\theta\right)^2 = \tau^2 + \beta_{(2)}^\top \Sigma \beta_{(2)},$$
$$n\mathbb{E}\left(\hat{\theta}_{LS}^1-\theta\right)^2 = \tau^2 + O(p^2/n),$$
$$n\mathbb{E}\left(\hat{\theta}_{SSLS}^1-\theta\right)^2 = \tau^2 + \frac{n}{n+m}\beta_{(2)}^\top \Sigma \beta_{(2)} + O(p^2/n).$$

Remark

- $$\mathbb{E}\left(\hat{\theta}_{SSLS}^1-\theta\right)^2 \approx \frac{n}{n+m}\mathbb{E}(\bar{Y}-\theta)^2 + \frac{m}{n+m}\mathbb{E}\left(\hat{\theta}_{LS}^1-\theta\right)^2.$$
- $\hat{\theta}_{LS}^1, \hat{\theta}_{SSLS}^1$ are asymptotically better than $\bar{Y}$ in $\ell_2$ risk, if $\beta_{(2)}^\top \Sigma \beta_{(2)} > 0$, i.e., $\mathbb{E}(Y|X)$ is significantly correlated with $X$.

# Asymptotic Distribution of $\hat{\theta}_{LS}$

**Theorem (Fixed $p$ growing $n$ asymptotics of $\hat{\theta}_{LS}$)**

*Assume $(Y, X) \sim P$. $P$ is fixed, has finite and non-degenerate second moments, $\tau^2 > 0$. Based on $n$ labeled samples, we have*

$$\frac{\hat{\theta}_{LS} - \theta}{\tau / \sqrt{n}} \xrightarrow{d} N(0,1), \quad MSE/\tau^2 \xrightarrow{d} 1 \quad \text{as} \quad n \to \infty,$$

$$\text{where} \quad MSE := \frac{\sum_{i=1}^{n} (Y_i - \vec{X}_i^\top \hat{\beta})^2}{n - p - 1}, \quad \tau^2 = \mathbb{E}(Y - \vec{X}^\top \beta)^2.$$

- Essen-Berry-type CLT: let the cdf of $\frac{\hat{\theta}_{LS} - \theta}{\tau/\sqrt{n}}$ be $F_n$,
  $\to |F_n(x) - \Phi(x)| \leq C n^{-1/4}$;
- Under $p = p_n = o(\sqrt{n})$ and other moment conditions,
  $\to$ asymptotic results still hold.

# Asymptotic Distribution of $\hat{\theta}_{SSLS}$

**Theorem (Fixed $p$ growing $n$ Asymptotics of $\hat{\theta}_{SSLS}$)**

*Assume $(Y, X) \sim P$, $P$ is fixed, $P$ has finite and non-degenerate second moments, $\tau^2 > 0$. Based on $n$ labeled samples and $m$ unlabeled samples,*

$$\frac{\hat{\theta}_{SSLS} - \theta}{\nu / \sqrt{n}} \xrightarrow{d} N(0, 1), \quad \hat{\nu}/\nu^2 \xrightarrow{d} 1, \quad \text{as} \quad n \to \infty,$$

*where* $$\hat{\nu} = \frac{m}{m+n} MSE + \frac{n}{m+n} \hat{\sigma}_Y^2, \quad \nu^2 = \tau^2 + \frac{n}{n+m} \beta_{(2)}^\top \Sigma \beta_{(2)},$$

$$MSE = \frac{1}{n-p-1} \sum_{k=1}^{n} (Y_i - \vec{X}_k^\top \hat{\beta})^2, \quad \hat{\sigma}_Y^2 = \frac{1}{n-1} \sum_{k=1}^{n} (Y_i - \bar{\mathbf{Y}})^2.$$

# Inference for $\theta$

- When $p = p_n = o(\sqrt{n})$, $(1-\alpha)$-level confidence interval for $\theta$:

(Ideal semi-supervised) $$\left[\hat{\theta}_{LS} \pm z_{1-\alpha/2}\sqrt{\frac{MSE}{n}}\right],$$

(Ordinary semi-supervised) $$\left[\hat{\theta}_{SSLS} \pm z_{1-\alpha/2}\sqrt{\frac{\frac{m}{m+n}MSE + \frac{n}{m+n}\hat{\sigma}_Y^2}{n}}\right].$$

- Traditional $z$-interval,

$$\left[\bar{Y} - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_Y^2}{n}}, \bar{Y} + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_Y^2}{n}}\right]$$

- Since

$$MSE \xrightarrow{d} \tau^2 \qquad < \qquad \hat{\sigma}_Y^2 \xrightarrow{d} \tau^2 + \beta_{(2)}^\top \Sigma \beta_{(2)}.$$

**LS-confidence intervals are asymptotically shorter!**

# Further Improvement

- $\hat{\theta}_{LS}, \hat{\theta}_{SSLS}$ explore linear relationship between $Y$ and $X$.
- Further improvement: add non-linear covariates

$$X_k^{\bullet} = \left(X_{k1}, \ldots, X_{kp}, g_1(X_k), \ldots, g_q(X_k)\right).$$

Semi-supervised least squared estimator:

$$\hat{\theta}_{LS}^{\bullet} = \bar{\boldsymbol{Y}} - (\hat{\beta}_{(2)}^{\bullet})^\top (\bar{\boldsymbol{X}}^{\bullet} - \mu^{\bullet}), \quad \hat{\beta}^{\bullet} = \left((\vec{\boldsymbol{X}}^{\bullet})^\top \vec{\boldsymbol{X}}^{\bullet}\right)^{-1} (\vec{\boldsymbol{X}}^{\bullet})^\top \boldsymbol{Y}.$$

$$\hat{\theta}_{SSLS}^{\bullet} = \bar{\boldsymbol{Y}} - (\hat{\beta}_{(2)}^{\bullet})^\top (\bar{\boldsymbol{X}}^{\bullet} - \hat{\mu}^{\bullet}), \quad \hat{\mu}^{\bullet} = \frac{1}{n+m} \sum_{k=1}^{n+m} \vec{X}_k^{\bullet}.$$

- Let $q$ grows slowly ($q = o(n^{1/2})$), one can establish **semiparametric efficiency** and **oracle optimality** for $\hat{\theta}_{LS}^{\bullet}$ and $\hat{\theta}_{SSLS}^{\bullet}$.

# Summary

- We introduced an “assumption lean” framework for semi-supervised inference and focus on $\theta = \mathbb{E}Y$.

- Ideal semi-supervised setting: $\hat{\theta}_{LS} = \bar{\boldsymbol{Y}} - \hat{\beta}_{(2)}^{\top}(\boldsymbol{X} - \mu)$.
  Ordinary semi-supervised setting: $\hat{\theta}_{SSLS} = \bar{\boldsymbol{Y}} - \hat{\beta}_{(2)}^{\top}(\boldsymbol{X} - \hat{\mu})$

- Further improvement to semiparametric efficient estimators $\hat{\theta}^{\bullet}_{LS}, \hat{\theta}^{\bullet}_{SSLS}$.

- Future Works:
  - $p$ significantly grows beyond $o\left(n^{1/2}\right)$
    $\rightarrow$ high-dimensional setting.
  - Other problems in semi-supervised settings
    $\rightarrow$ classification, regression, covariance estimation, PCA, CNN, ...

# References


- Zhang, A., Brown, L. and Cai, T. (2018). Semi-supervised inference: General theory and estimation of means. *Annals of Statistics*, to appear.

# In Memory of Larry