Two-sample testing in data integration

Subhabrata Majumdar

Abstract:

Keywords:

1 Model

We have data $\mathcal{Z} = \{\mathcal{Z}^1, \ldots, \mathcal{Z}^K\}; \mathcal{Z}^k = (\mathbf{Y}^k, \mathbf{X}^k)$ where $\mathbf{Y}^k \in \mathbb{R}^{n \times q}, \mathbf{X}^k \in \mathbb{R}^{n \times p}$ for $1 \leq k \leq K$.

$$\mathbf{X}^k = (\mathbf{X}^k_1, \ldots, \mathbf{X}^k_p)^T \sim \mathcal{N}(0, \Sigma^k_x) \tag{1.1}$$

$$\mathbf{Y}^k = \mathbf{X}^k \mathbf{B}^k + \mathbf{E}^k; \quad \mathbf{E}^k = (\mathbf{E}^k_1, \ldots, \mathbf{E}^k_p)^T \sim \mathcal{N}(0, \Sigma^k_y) \tag{1.2}$$

$$\Omega^k_x = (\Sigma^k_x)^{-1}; \quad \Omega^k_y = (\Sigma^k_y)^{-1} \tag{1.3}$$

Want to estimate $\{(\Omega^k_x, \Omega^k_y, \mathbf{B}^k); 1 \leq k \leq K\}$ in presence of known grouping structures $\mathcal{G}_x, \mathcal{G}_y, \mathcal{H}$ respectively.

Notation: Denote 3-dimensional array objects as elements of $\mathbb{T}(a, b, c)$, the set of all $a \times b \times c$ tensors. Define $\mathcal{S}^x = (\Omega^k_x), \mathcal{S}^y = (\Omega^k_y), \mathcal{B} = (\mathbf{B}^k)$

Estimation of $\{\Omega^k_x\}$ done using JSEM. For the other part, we use the following two-step procedure:

1. Run neighborhood selection on y-network incorporating effects of x-data and an additional block-wise group penalty:

$$\min_{\mathcal{B}, \Theta} \left\{ \sum_{i=1}^p \frac{1}{n_k} \left[\sum_{k=1}^K \| \mathbf{Y}^k_i - \mathbf{Y}^k_{-i} \boldsymbol{\theta}^k_i - \mathbf{X}^k \mathbf{B}^k_i \|^2 + 2 \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}_y} \lambda^g_{ij} \| \boldsymbol{\theta}^{[g]}_{ij} \| \right] + 2 \sum_{b \in \mathcal{G}_x \times \mathcal{G}_y \times \mathcal{H}} \eta^b \| \mathbf{B}^{[b]} \| \right\} \tag{1.4}$$

$$= \min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta) + P(\Theta) + Q(\mathcal{B}) \right\} \tag{1.5}$$

   where $\Theta = \{\Theta_i\}, \mathcal{B} = \{\mathbf{B}^k\}, \mathcal{Y} = \{\mathbf{Y}^k\}, \mathcal{X} = \{\mathbf{X}^k\}, \mathcal{E} = \{\mathbf{E}^k\}$.

   This estimates $\mathcal{B}$ **(possibly refit and/or within-group threshold)** .

2. Step I part 2 and step II of JSEM (see 15-656 pg 6) follows to estimate $\{\Omega^k_y\}$.

The objective function is bi-convex, so we are going to do the following in step 1-

- Start with initial estimates of $\mathcal{B}$ and Θ, say $\mathcal{B}^{(0)}, \Theta^{(0)}$.

- Iterate:

$$\Theta^{(t+1)} = \arg\min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}^{(t)}, \Theta^{(t)}) + P(\Theta^{(t)}) \right\} \tag{1.6}$$

$$\mathcal{B}^{(t+1)} = \arg\min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}^{(t)}, \Theta^{(t+1)}) + Q(\mathcal{B}^{(t)}) \right\} \tag{1.7}$$

- Continue till convergence.

2 Two-sample testing

Suppose there are two disease subtypes: $k = 1, 2$, and we are interested in testing whether the downstream effect of a predictor is X-data is same across both subtypes, i.e. if $\mathbf{b}_i^1 = \mathbf{b}_i^2$ for some $i \in \{1, \ldots, p\}$. For this we consider the modified optimization problem:

$$\min_{\mathcal{B}, \Theta} \frac{1}{n} \left\{ \sum_{j=1}^{q} \sum_{k=1}^{2} \|\mathbf{Y}_j^k - \mathbf{Y}_{-j}^k \boldsymbol{\theta}_j^k - \mathbf{X}^k \mathbf{b}_j^k\|^2 + \sum_{j \neq j'} \lambda_{jj'} \|\boldsymbol{\theta}_{jj'}^*\| + \sum_{i=1}^{p} \eta_i \|\mathbf{B}_{i*}^*\| \right\} \tag{2.1}$$

$$= \min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta) + P(\Theta) + Q(\mathcal{B}) \right\} \tag{2.2}$$

with $n_1 = n_2 = n$ for simplicity; and $\mathbf{B}^k = (\mathbf{b}_1^k, \ldots, \mathbf{b}_q^k), (\mathbf{B}_{i*}^*) \in \mathbb{R}^{q \times K}$

3 Conditions

Conditions A1, A2, A3 from JSEM paper.

4 Results

Define

$$\hat{\Theta}^i = \arg\min_{\Theta_i} \left\{ \frac{1}{n_k} \sum_{k=1}^{K} \|\mathbf{Y}_i^k - \mathbf{Y}_{-i}^k \boldsymbol{\theta}_i^k - \mathbf{X}^k \hat{\mathbf{B}}_i^k\|^2 + 2 \sum_{j \neq i} \sum_{g \in \mathcal{G}_y^{ij}} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\| \right\} \tag{4.1}$$

Theorem 4.1. *Assume fixed $\mathcal{X}, \mathcal{E}$ and deterministic $\hat{\mathcal{B}} = \{\mathbf{B}^k\}$. Also*

(T1) $\|\hat{\mathbf{B}}_i^k - \mathbf{B}_i^k\| \leq v_\beta$*;*

(T2) $\|\mathbf{X}^k(\hat{\mathbf{B}}_i^k - \mathbf{B}_i^k)\| \leq c(v_\beta)$ *for some non-negative function $c(.)$;*

Group uniform IC.

Then

(I) Estimation consistency

(II) Direction consistency

Proof of Theorem 4.1. Part I. Follows proof of thm 1 in 15-656. The proof has 3 parts: consistency of neighborhood regression, selection of edge sets, and finally the refitting step.

For any $g \in \mathcal{G}^{ij}, k \in g$, and $j \neq i$, let

$$\hat{\boldsymbol{\epsilon}}_i^k = \mathbf{Y}_i^k - \mathbf{Y}_{-i}^k \boldsymbol{\theta}_{0,i}^k - \mathbf{X}^k \hat{\mathbf{B}}_i^k; \quad \hat{\zeta}_{ij}^k = \frac{(\hat{\boldsymbol{\epsilon}}_i^k)^T \mathbf{Y}_j^k}{n}; \quad \hat{\boldsymbol{\zeta}}_{ij}^{[g]} = (\hat{\zeta}_{ij}^k)_{k \in g}$$

Consider the random event $\mathcal{A} = \bigcap_{i,j \neq i, g} \mathcal{A}_{ij}^g$ with $\mathcal{A}_{ij}^g = \{2 \|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\| \leq \lambda_{ij}^g\}$.

Proposition 4.2. *Given that λ_{ij}^g are chosen as*

$$\lambda_{ij}^g \geq \max_{k \in g} \frac{2}{\sqrt{n \omega_{ii}^k}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} + \sqrt{c(v_\beta)} \right)$$

we shall have $\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}$ for some $q > 1$.

Proof of Proposition 4.2. We follow the proof of Lemma E.2 in 15-656, with $\mathbf{Y}_j^k, \hat{\boldsymbol{\epsilon}}_i^k, \hat{\zeta}_{ij}^k, \hat{\boldsymbol{\zeta}}_{ij}^{[g]}$ in place of $\mathbf{X}_j^k, \boldsymbol{\epsilon}_i^k, \zeta_{ij}^k, \boldsymbol{\zeta}_{ij}^{[g]}$ respectively. Proceeding in a similar fashion we get

$$\|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\|^2 = \frac{1}{n} (\|\mathbf{Z}^{[g]}\|^2 + 2 \sum_{k \in g} Z^k (\mathbf{Q}_j^k)^T \boldsymbol{\delta}_i^k + \|(\mathbf{Q}_j^k)^T \boldsymbol{\delta}_i^k\|^2)$$

where $\mathbf{Z}^{[g]} = (Z^k)_{k \in g}; Z^k = (\mathbf{Q}_j^k)^T \boldsymbol{\epsilon}_i^k$ with $\boldsymbol{\epsilon}_i^k := \mathbf{Y}_i^k - \mathbf{Y}_{-i}^k \boldsymbol{\theta}_{0,i}^k - \mathbf{X}^k \mathbf{B}_{0,i}^k$, $\mathbf{Q}_j^k$ is the first eigenvector of $\mathbf{Y}_j^k (\mathbf{Y}_j^k)^T / n$, and $\boldsymbol{\delta}_i^k := \mathbf{X}^k (\mathbf{B}_{0,i}^k - \hat{\mathbf{B}}_i^k)$. Applying Cauchy-schwarz inequality to right side and by assumption (T2),

$$\|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\| \leq \frac{1}{\sqrt{n}} (\|\mathbf{Z}^{[g]} + \sqrt{c(v_\beta)})$$

thus

$$\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) = \mathbb{P}\left(\|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\| > \frac{\lambda_{ij}^g}{2} \right) \leq \mathbb{P}\left(\|\mathbf{Z}^{[g]}\| > \frac{\sqrt{n} \lambda_{ij}^g}{2} - \sqrt{c(v_\beta)} \right)$$

We now proceed through the proof of Lemma E.2 in 15-656 to end up with the choice of λ_{ij}^g. $\square$

All subsequent derivations in the theorem go through with the new choice of λ_{ij}^g.

Part II. Proof of Thm 2 in 15-656 follows. We only need a new bound for $Var(\mathbf{Y}_i^k | \mathbf{Y}_{-i}^k, \mathbf{X}^k, \hat{\mathbf{B}}_i^k)$. For this we have

$$Var(\mathbf{Y}_i^k | \mathbf{Y}_{-i}^k, \mathbf{X}^k, \hat{\mathbf{B}}_i^k) = \mathbb{E}(\hat{\boldsymbol{\epsilon}}_i^k)^2 = \mathbb{E}(\boldsymbol{\epsilon}_i^k + \boldsymbol{\delta}_i^k)^2 \leq \left(\frac{1}{d_0} + \frac{c(v_\beta)}{n} \right)^2$$

applying cauchy-schwarz inequality followed by assumption (A2). Now Replace $1/\sqrt{nd_0}$ in choice of λ, α_n in Thm 2 statement with $1/\sqrt{n}(\sqrt{1/d_0} + \sqrt{c(v_\beta)/n})$.

$\square$

Proposition 4.3. *Given fixed $\hat{\mathcal{B}}$, prediction errors follow bound in T2 with high enough probability.*

Now concentrate on the k-population estimation problem. We want to obtain

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{pqK}}{\arg\min} \{ -2\boldsymbol{\beta}\hat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^T \boldsymbol{\Gamma} \boldsymbol{\beta} + \| \boldsymbol{\beta} \|_{2,g} \}$$

with

$$\boldsymbol{\beta} = \begin{bmatrix} \text{vec}(\mathbf{B}^1) \\ \vdots \\ \text{vec}(\mathbf{B}^K) \end{bmatrix}; \quad \boldsymbol{\Gamma} = \begin{bmatrix} I_q \otimes (\mathbf{X}^1) T X^1/n) & & \\ & \ddots & \\ & & I_q \otimes (\mathbf{X}^K)^T X^K/n) \end{bmatrix}$$

Theorem 4.4.