Linear regression without correspondence

Daniel Hsu†, Kevin Shi†, Xiaorui Sun♯,
†Columbia University, ♯Simons Institute for the Theory of Computing

Problem definition

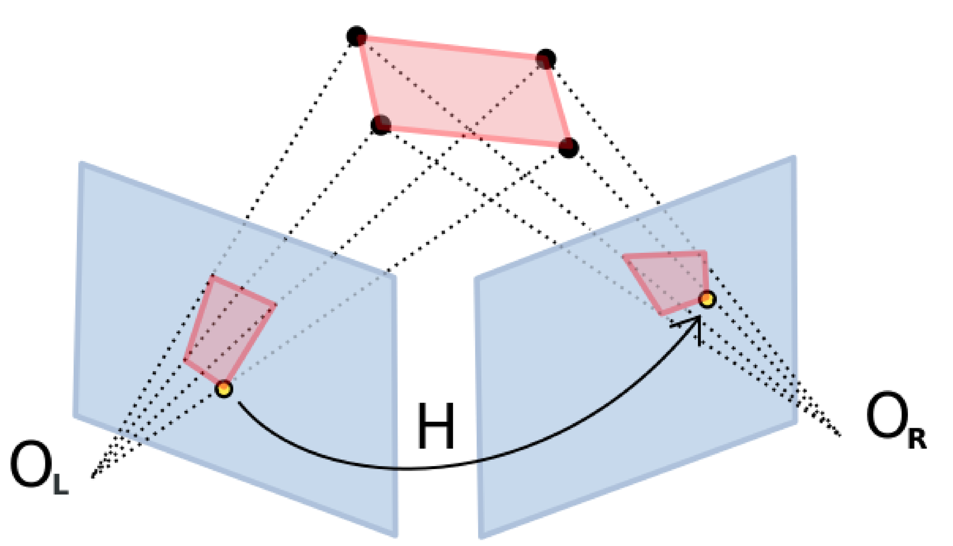
- **Covariate vectors**: $x_1, x_2, \dots, x_n \in \mathbb{R}^d$
- **Responses**: $y_1, y_2, \dots, y_n \in \mathbb{R}$
- **Model**:

$$y_i \;=\; \bar{w}^\top x_{\bar{\pi}(i)} + \varepsilon_i\,, \quad i \in [n]$$

- Unknown linear function: $\bar{w} \in \mathbb{R}^d$
- Unknown permutation: $\bar{\pi} \in S_n$
- Measurement errors: $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \in \mathbb{R}$ e.g., $(\varepsilon_i)_{i=1}^n$ iid from $\mathrm{N}(0, \sigma^2)$)

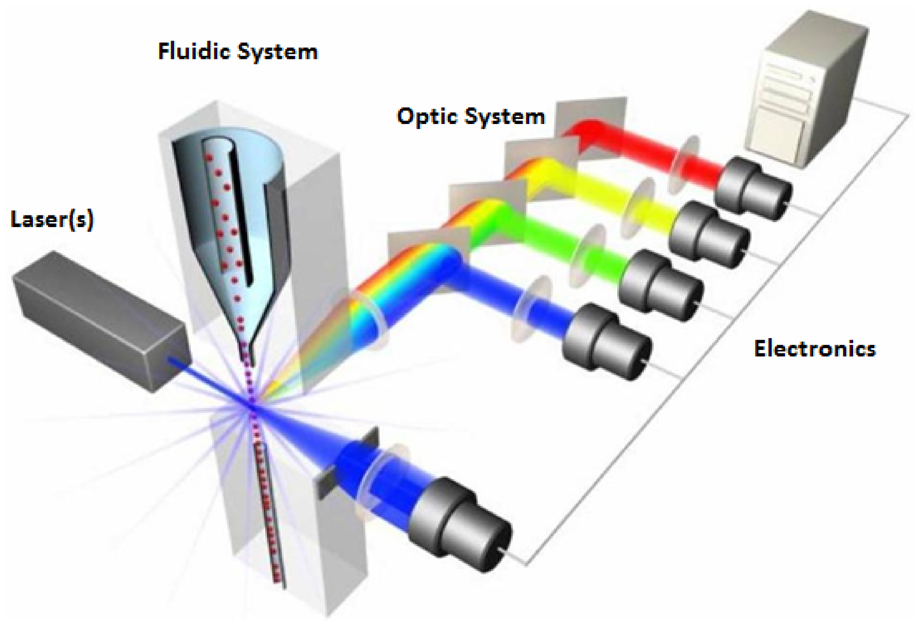
Examples

Multi-view geometry



- Unknown correspondence between keypoints

Flow cytometry



- Observe the entire emission spectrum at once

Strong NP-hardness

Definition 1 (Permuted Linear System). Given $X \in \mathbb{Z}^{n\times d}, Y \in \mathbb{Q}^n$, decide if there exists a vector $w \in \mathbb{Q}^d$ and a permutation $\pi \in S_n$ such that $Xw = Y_\pi$

Proposition 1. *Permuted Linear System is strongly NP-complete by a reduction from 3-Partition.*

Approximation guarantee for least-squares

Definition 2 (Least-squares recovery). Given $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ from $\mathbb{R}$, find

$$(\hat{w}_{mle}, \hat{\pi}_{mle}) := \operatorname*{arg\,min}_{w\in\mathbb{R},\pi\in S_n} \sum_{i=1}^n \left(y_i - w x_{\pi(i)}\right)^2$$

Theorem 1. *There is an algorithm that given any inputs $(x_i)_{i=1}^n$, $(y_i)_{i=1}^n$, and $\epsilon \in (0,1)$, returns a $(1+\epsilon)$-approximate solution to the least squares problem in time $(n/\epsilon)^{O(k)} + \mathrm{poly}(n,d)$, where $k = \dim(\mathrm{span}(x_i)_{i=1}^n)$. ;*

Approximation algorithm

This uses the following coreset result for linear systems:

Proposition 2 (Boutsidis, Drineas, Magdon-Ismail). *Given a matrix $A \in \mathbb{R}^{n\times k}$, there exists a weighted subset of $4k$ rows determined by a matrix $S \in \mathbb{R}^{4k\times n}$ such that for any b, every minimizer of the subsampled linear system*

$$w' \in \operatorname*{arg\,min}_{w} \|S(Aw - b)\|_2^2$$

also satisfies

$$\|Aw' - b\|_2^2 \le c$$

for $c = O(n/k)$. Morever, there exists an efficient algorithm which returns a matrix S in time poly(n,k).

Algorithm 1 Approximation algorithm

input Covariate matrix $X = [x_1|x_2|\cdots|x_n]^\top \in \mathbb{R}^{n\times k}$; response vector $y = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^n$; approximation parameter $\epsilon \in (0,1)$.

Compute the matrix $S \in \mathbb{R}^{r\times n}$ from input matrix X.
Let $\mathcal{B}$ be the set of all permutations of y
Let $c := 1 + 4(1+\sqrt{n/(4k)})^2$.
for each $b \in \mathcal{B}$ **do**
  Compute $\tilde{w}_b \in \arg\min_{w\in\mathbb{R}^k} \|[\|0]S(Xw - b)_2^2$, and let $r_b := \min_{\Pi\in\mathcal{P}_n} \|[\|0]X\tilde{w}_b - \Pi^\top y_2^2$.
  Construct a $\sqrt{\epsilon r_b/c}$-net $\mathcal{N}_b$ for the Euclidean ball of radius $\sqrt{c r_b}$ around $\tilde{w}_b$, so that for each $v \in \mathbb{R}^k$ with $\|[\|0]v - \tilde{w}_{b2} \le \sqrt{c r_b}$, there exists $v' \in \mathcal{N}_b$ such that $\|[\|0]v - v'_2 \le \sqrt{\epsilon r_b/c}$.
end for
return

$$\hat{w} \in \operatorname*{arg\,min}_{w\in\bigcup_{b\in\mathcal{B}}\mathcal{N}_b} \min_{\Pi\in\mathcal{P}_n} \|Xw - \Pi^\top y\|_2^2$$

and

$$\hat{\Pi} \in \operatorname*{arg\,min}_{\Pi\in\mathcal{P}_n} \|X\hat{w} - \Pi^\top y\|_2^2$$

Polynomial time recovery in the random setting

Theorem 2. *Fix any $\bar{w} \in \mathbb{R}^d$ and $\bar{\pi} \in S_n$, and assume $n \ge d$. Suppose $(x_i)_{i=0}^n$ are drawn iid from $\mathrm{N}(0, I_d)$, and $(y_i)_{i=0}^n$ satisfy*

$$y_0 \;=\; \bar{w}^\top x_0\,; \qquad y_i \;=\; \bar{w}^\top x_{\bar{\pi}(i)}\,, \quad i \in [n]\,.$$

There is an algorithm that, given inputs $(x_i)_{i=0}^n$ and $(y_i)_{i=0}^n$, returns $\bar{\pi}$ and $\bar{w}$ with high probability.

Reduction to (random) subset sum

Given $d+1$ measurements and one correspondence $y_0 = \bar{w}^T x_0$, for orthogonal $(x_i)_{i=0}^n$, can write:

$$y_0 = \sum_{j=1}^d (\bar{w}^\top x_j)\,(x_j^\top x_0) = \sum_{j=1}^d y_{\bar{\pi}^{-1}(j)}\,(x_j^\top x_0)$$

$$= \sum_{i=1}^d \sum_{j=1}^d \mathbb{1}\{\bar{\pi}(i) = j\} \cdot \underbrace{y_i\,(x_j^\top x_0)}_{c_{i,j}}$$

- $\{c_{i,j}\}$ and y_0 define a subset sum problem whose solution recovers the underlying correspondence.
- In general $(x_i)_{i=0}^n$ are close to orthogonal; use the Moore-Penrose pseudoinverse.
- The one given correspondence can be brute-forced, creating $d+1$ subset sum instances of which only one has a solution

Solving random subset-sum instances

Proposition 3 (Lagarias and Odlyzko). *Random instances of subset sum are efficiently solvable when the $c_{i,j}$'s are independently and uniformly distributed over a large enough subinterval of $\mathbb{Z}$.*

This relies on the following inequality which lower bounds the closeness to the target sum of incorrect solutions.

Lemma 1. *For any vector $(z_{i,j})$ which is not the correct correspondence,*

$$\left| y_0 - \sum_{i,j} z_{i,j} c_{i,j} \right| \ge \frac{1}{2^{poly(d)}} \|\bar{w}\|_2$$

- We show this bound holds under other distributions satisfying general anticoncentration bounds and even if the $c_{i,j}$'s are not independent

Reduction to shortest vector problem

Definition 3 (Shortest vector problem). Given a lattice basis $\mathbf{B} \subset \mathbb{R}^d$, output a lattice vector $\mathbf{B}z \in \Lambda\mathbf{B}$ where

$$z = \operatorname*{arg\,min}_{z\in\mathbb{Z}-\{0\}} \|\mathbf{B}z\|_2^2$$

Lemma 2 (LLL Lattice Basis Reduction). *There is an efficient approximation algorithm for solving the Shortest Vector Problem with*

- *Approximation factor: $2^{d/2}$*
- *Running time: poly$(d, \log\lambda(\mathbf{B}))$*

Algorithm 2 Lattice algorithm for subset sum

input Source numbers $\{c_i\}_{i\in\mathcal{I}} \subset \mathbb{R}$; target sum $t \in \mathbb{R}$; lattice parameter $\beta > 0$.

Construct lattice basis $B \in \mathbb{R}^{(|\mathcal{I}|+2)\times(|\mathcal{I}|+1)}$ where

$$B \;:=\; \begin{bmatrix} I_{|\mathcal{I}|+1} \\ \beta t \,|\, -\beta c_i : i \in \mathcal{I} \end{bmatrix} \;\in\; \mathbb{R}^{(|\mathcal{I}|+2)\times(|\mathcal{I}|+1)}\,.$$

Run LLL Lattice Basis Reduction to find non-zero lattice vector v of length at most $2^{|\mathcal{I}|/2} \cdot \lambda_1(B)$.

Information-theoretic lower bounds on SNR

Definition 4 (Random measurement setting). Observe

$$y_i = \bar{w}^T x_{\bar{\pi}(i)} + \epsilon_i$$

where

- $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ is the measurement noise
- $x_i \overset{iid}{\sim} \mathcal{N}(0, I_d)$ are the covariates

Definition 5 (SNR). The signal-to-noise ratio for this model is $\|\bar{w}\|_2^2/\sigma^2$

Theorem 3. *In the random measurement setting, if for some constant C*

$$SNR \le C \cdot \min\left\{\frac{d}{\log\log n}, 1\right\}$$

then for every estimator $\hat{w}$, there exists a $\bar{w} \in \mathbb{R}^d$ such that

$$\mathbb{E}\left[\|\hat{w} - \bar{w}\|_2\right] \ge \frac{1}{24}\|\bar{w}\|_2$$

- Recall that standard linear regression satisfies the bound $\mathbb{E}\left[\|w_{mle} - \bar{w}\|\right] \le C\sigma\sqrt{d/n}$
- In the low SNR regime, more measurements makes the problem more difficult

djhsu@cs.columbia.edu, kshi@cs.columbia.edu, xiaoruisun@cs.columbia.edu