A literature review on regression with shuffled data

Sungmin Park

November 3, 2022

Regression with shuffled data

Consider a linear regression problem. Given a set of predictors $\{X_1, \ldots, X_n\}$ and responses $\{Y_1, \ldots, Y_n\}$, we implicitly assume that Y_i is related to X_i of the same index through a model

$$Y_i = X_i^T \beta + \epsilon_i.$$

However, suppose the set of predictor is now shuffled and Y_i is related to $X_{\pi(i)}$ for some unknown permutation π as

$$Y_i = X_{\pi(i)}^T \beta + \epsilon_i.$$

Y_1	X_1
Y_2	X_4
Y_3	X_3
Y_4	X_2
Y_5	X_5

Table: Mismatch example

Regression with shuffled data

The model can also be expressed using a permutation matrix Π:

$$Y = \Pi X\beta + \epsilon.$$

Is it possible to i) recover the original ordering of predictors and ii) estimate the regression coefficient β as well? The problem has other names as well:

- regression under mismatch,
- permutation recovery (emphasis on recovering Π),
- unlabeled sensing (emphasis on recovering β),

But why?

If we ignore the permutation, the least squares estimate is biased:

$$\begin{aligned}\mathbf{E}\hat{\beta} &= \mathbf{E}\left\{(\Pi X)^T(\Pi X)\right\}^{-1}(\Pi X)^T Y \\ &= \mathbf{E}(X^T X)^{-1}X^T\Pi Y = (X^T X)^{-1}(X^T\Pi X)\beta.\end{aligned}$$

Consider $Y_i = X_i\beta + \epsilon_i$ with $\beta = 1, n = 10^4$, and $X_i, \epsilon_i \sim N(0,1)$ i.i.d.

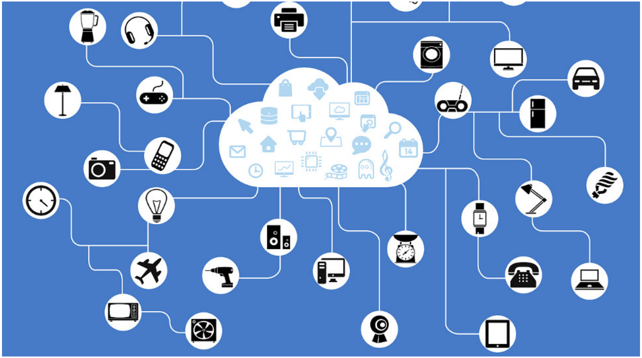
Permutation	$\hat{\beta}$
0%	1.0001
1%	0.9901
10%	0.9000
20%	0.8000
50%	0.4997

Table: Estimated coefficient ignoring mismatch from 1000 simulations

But why? Applications

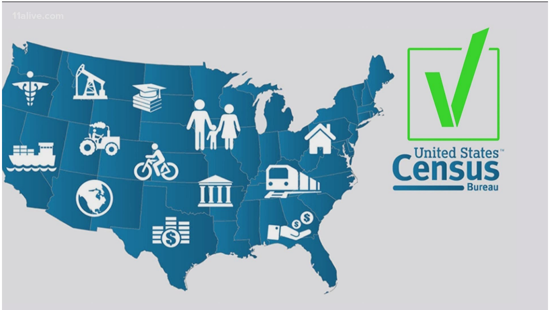
Header-less communication:

- In large sensor networks, the size of data a sensor records and sends to a main server is sometimes exceeded by or comparable to the size of the sensor ID information.
- If measurements are linear, transmit data without sensor ID.



But why? Applications

- Post-linkage data analysis:
  - It is cost-effective to combine data from various sources rather than collection a new data containing all variables of interest.
  - Due to data quality, the links between the data can be error-prone and modeling mismatch using permutation can help.



But why? Applications

- Unsupervised alignment: Align two sets of text embedding for natural language processing.

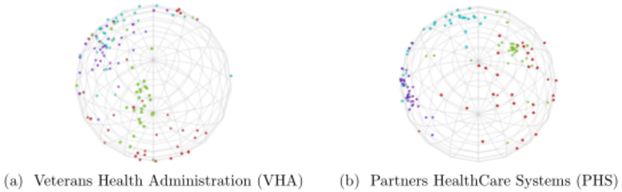


Figure: Examples of word embeddings on a sphere

Regression with shuffled data

- At a higher level, there are two different directions of research:
  - Permutation is random.
  - Permutation is deterministic.
- When permutation is assumed to be random, there is a line of research which has the name *record linkage* problem.
- In this presentation, we will focus more on the recent results in the deterministic case.

Deterministic case

One of the first results in the deterministic case was presented by Pananjady et al. (2018) providing sufficient and necessary conditions for a successful recovery. For model

$$Y = \Pi^* X \beta^* + \epsilon,$$

with

- Response: $Y \in \mathbb{R}^n$,
- Predictor: $X \in \mathbb{R}^{n \times p}$ s.t. $X_{ij} \sim N(0, 1)$ i.i.d.
- Random noise: $\epsilon \in \mathbb{R}^n$ s.t. $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.

define the maximum likelihood estimator:

$$(\hat{\Pi}, \hat{\beta}) = \underset{\Pi, \beta}{\operatorname{argmin}} \|Y - \Pi X \beta\|_2^2.$$

Pananjady et al. (2018)

Theorem 1 (Sufficient condition for successful recovery)

Define signal-to-noise ratio $SNR = \frac{\|\beta^*\|_2^2}{\sigma^2}$*. For any* $p < n$ *and* $\xi < \sqrt{n}$*, if*

$$\log(SNR) \geq \left(c_1 \frac{n}{n-p} + \xi \right) \log n,$$

then $\mathbf{P}(\hat{\Pi} \neq \Pi^*) \leq c_2 n^{-2\xi}$ *where* c_1, c_2 *are absolute constants.*

Theorem 2 (Necessary condition for successful recovery)

For any $\delta \in (0, 2)$*, if*

$$2 + \log(1 + SNR) \leq (2 - \delta) \log n,$$

then $\mathbf{P}(\tilde{\Pi} \neq \Pi^*) \geq 1 - c_3 e^{-c_4 n \delta}$ *for any estimator* $\tilde{\Pi}$ *where* c_3, c_4 *are absolute constants.*

Pananjady et al. (2018)

- Sufficient and necessary conditions of successful recovery is characterized by a threshold on the signal-to-noise ratio.
- A few downsides:
  - (Computation) Solving

$$(\hat{\Pi}, \hat{\beta}) = \underset{\Pi,\beta}{\operatorname{argmin}} \|Y - \Pi X \beta\|_2^2$$

    is generally intractable for $\dim(\beta) > 1$ (quadratic assignment problem).
  - Sufficient condition requires $\mathrm{SNR} = \frac{\|\beta^*\|_2^2}{\sigma^2}$ to be stronger than n^c ($\approx n^5$). Stronger SNR is needed as we acquire more data.
  - Why? The search space is growing exponentially. ($n!$ permutations)

More realistic cases

Do we always need to search the entire set of permutations?

- Mismatch may occur only rarely (sparsity).
- Permutations may only occur within groups (hierarchy in the observations).

Nguyen and Tran (2012)

If mismatch only occurs rarely (sparse), then there are other methods available as well. Nguyen and Tran (2012) provides a method to consistently estimate the regression parameters under a regression model with grossly corrupted observations (errors-in-variables model):

$$Y = X\beta^* + \sqrt{n}e^* + \epsilon$$

where

- Response: $Y \in \mathbb{R}^n$,
- Design matrix: $X \in \mathbb{R}^{n \times p}$,
- Gross-error: $e \in \mathbb{R}^n$ (fixed),
- Random noise: $\epsilon \sim N(0, \sigma^2 I)$.

Nguyen and Tran (2012)

$$Y = X\beta^* + \sqrt{n}e^* + \epsilon$$

- Looking at the model more closely, we have
  1. $\pi(i) = i$, $Y_i = x_i^T\beta^* + \sqrt{n}e_i^* + \epsilon_i = x_i^T\beta^* + 0 + \epsilon_i$,
  2. $\pi(i) \neq i$, $Y_i = x_i^T\beta^* + \sqrt{n}e_i^* + \epsilon_i = x_i^T\beta^* + (x_{\pi(i)} - x_i)^T\beta^* + \epsilon_i$,

  or, $e_i \neq 0 \iff \pi(i) \neq i$.
- Furthermore, assume that β^* and e^* are sparse:
  - $\|\beta^*\|_0 = \sum_i I(\beta_i \neq 0) = k$,
  - $\|e^*\|_0 = s$.

Nguyen and Tran (2012)

Let $(\hat{\beta}, \hat{e})$ be the solution to

$$\min_{\beta, e} \frac{1}{2n} \|Y - X\beta - \sqrt{n}e\|_2^2 + \lambda_{n,\beta}\|\beta\|_1 + \lambda_{n,e}\|e\|_1.$$

Theorem 3

Under some regularity conditions,

$$\|\hat{\beta} - \beta^*\|_2 + \|\hat{e} - e^*\|_2 \leq C\left(\frac{1}{\mu}\sqrt{\frac{\sigma^2 k \log p}{n}} + \sqrt{\frac{\sigma^2 s \log n}{n}}\right)$$

with high probability if regularization parameters are chosen as $\lambda_{n,\beta} = \frac{4}{\mu}\sqrt{\frac{\sigma^2 \log p}{n}}$ *and* $\lambda_{n,e} = 4\sqrt{\frac{\sigma^2 \log n}{n}}$ *for some* $\mu \in \left[\frac{1}{\sqrt{\log n}}, 1\right]$.

Nguyen and Tran (2012)

$$\|\hat{\beta} - \beta^*\|_2 + \|\hat{e} - e^*\|_2 \leq C \left(\frac{1}{\mu} \sqrt{\frac{\sigma^2 k \log p}{n}} + \sqrt{\frac{\sigma^2 s \log n}{n}} \right)$$

Therefore, if

- the number of mismatch $s = o\left(\frac{n}{\log n}\right)$,
- or the ratio of mismatch $\frac{s}{n} = o\left(\frac{1}{\log n}\right)$,

then we can consistently estimate β^* and also the indexes of observations where $\pi(i) \neq i$ in regression with suffled data.

Multivariate case

Going back to the specific case of shuffled data, Zhang et al. (2021) expands the results of Pananjady et al. (2018) to multivariate linear regression

$$Y = \Pi^* X B^* + E$$

or

$$Y_j = \Pi^* X \beta_j^* + \epsilon_j \text{ for } j = 1, \ldots, m$$

where

- Response: $Y = [Y_1, \ldots, Y_m] \in \mathbb{R}^{n \times m}$,
- Predictor: $X \in \mathbb{R}^{n \times p}$, $X_{ij} \sim N(0, 1)$ i.i.d.,
- Regression coefficient: $B^* = [\beta_1^*, \ldots, \beta_m^*] \in \mathbb{R}^{p \times m}$,
- Random noise: $E = [\epsilon_1, \ldots, \epsilon_m] \in \mathbb{R}^{n \times m}$, $E_{ij} \sim N(0, 1)$ i.i.d.

Zhang et al. (2021)

Again, define the maximum likelihood estimator

$$(\hat{\Pi}, \hat{B}) = \underset{\Pi, B}{\operatorname{argmin}} \|Y - \Pi X B\|_2^2.$$

Theorem 4 (Inachievability results)

Let $\mathcal{H}$ be any subset of $\mathcal{P}_n$, the set of all $n \times n$ permutation matrices. Assuming B^* is known, we have

$$\inf_{\tilde{\Pi}} \sup_{\Pi^* \in \mathcal{H}} \mathbf{P}(\tilde{\Pi} \neq \Pi^*) \geq \frac{1}{2} \quad \textit{if} \quad \log\det\left(I + \frac{B^{*T} B^*}{\sigma^2}\right) < \frac{\log(|\mathcal{H}|) - 2}{n}$$

where the expectation is taken w.r.t X and E, and the infimum is over all estimators $\tilde{\Pi}$.

Zhang et al. (2021)

The necessary condition for sufficient recovery

$$\log\det\left(I + \frac{{B^*}^T B^*}{\sigma^2}\right) \geq \frac{\log(|\mathcal{H}|) - 2}{n}$$

- If $m = 1$, $\mathcal{H} = \mathcal{P}_n$, $|\mathcal{H}| = n!$ and $\frac{\log(|\mathcal{H}|-2)}{n} \approx \frac{n\log n}{n} = \log n$ holds showing us a similar bound found in Pananjady et al. (2018)

$$2 + \log\left(1 + \frac{\|\beta^*\|_2^2}{\sigma^2}\right) \leq (2-\delta)\log n$$

- If our search space is smaller than the entire set of permutation matrices (e.g. $|\mathcal{H}|$ is a fixed nonnegative integer), the necessary condition becomes less restrictive.

Zhang et al. (2021)

- Let λ_i's be the singular values of B^*. Then

$$\log\det\left(I + \frac{B^{*T}B^*}{\sigma^2}\right) = \sum_i \log\left(1 + \frac{\lambda_i^2}{\sigma^2}\right)$$

- Suppose the signal energy $\|B^*\|_F^2 = \sum_i \lambda_i^2$ is fixed. In order to maximize $\log\det\left(I + \frac{B^{*T}B^*}{\sigma^2}\right)$, it is favorable to have λ_i's with more or less similar magnitude.

Zhang et al. (2021)

This leads to the notion of *stable rank*:

$$\rho(B^*) := \frac{\|B^*\|_F^2}{\|B^*\|_2^2} = \frac{\sum_i \lambda_i^2}{\max_i \lambda_i^2} \quad \text{for} \quad B^* \neq 0.$$

- (Worst case) B^* is rank 1:

$$\rho(B^*) = \frac{\sum_i \lambda_i^2}{\max_i \lambda_i^2} = \frac{\lambda_1^2}{\lambda_1^2} = 1.$$

- (Best case) B^* is full rank (rank m) with constant singular values:

$$\rho(B^*) = \frac{\sum_i \lambda_i^2}{\max_i \lambda_i^2} = \frac{\sum_{i=1}^m \lambda^2}{\lambda^2} = m.$$

Zhang et al. (2021)

Theorem 5 (Achievability results)

Define Hamming distance $d_H(\Pi_1, \Pi_2) = \sum_{i=1}^{n} I(\pi_1(i) \neq \pi_2(i))$. *Suppose*

1. $d_H(I, \Pi^*) \leq h_{max}$ *with* $h_{max} \times rank(B^*) \leq \frac{n}{8}$,
2. $SNR = \frac{\|B^*\|_F^2}{m\sigma^2} > c_0$,
3. $\rho(B^*) \geq c_1 \log n$,
4. $\log(SNR) \geq \frac{c_2 \log n}{\rho(B^*)} + c_3$,

then ML estimator $\hat{\Pi}$ *equals* Π^* *with probability going to 1, where* $c_0, \ldots, c_3$ *are some positive constants.*

Zhang et al. (2021)

- If
  - $\rho(B^*) = \text{rank}(B^*) \geq O(\log n)$,
  - the number of maximum mismatch $h_{\max} = O\left(\frac{n}{\log n}\right)$,

  then ML estimator consistently estimates Π^* without having a stringent condition on the order of signal-to-noise ratio.
- This result is better than Nguyen and Tran (2012) as they require the number of mismatch to be $o\left(\frac{n}{\log n}\right)$.
- The authors hypothesize that they can further remove the constraint on $h_{\max}$ using more advanced proof techniques.

Testing for the presence of mismatch

While there are numerous results on the estimation problem, there are not as many results on the inference part. Slawski et al. (2019) suggests a method to test the presence of mismatch in the linear regression model.

1. define P_X be the orthogonal projection onto range(X),
2. do SVD on $P_X^{\perp} = I - P_X := UU^T$.

Under the assumption of a linear model with i.i.d. normal error,

$$\xi := U^T Y = U^T(\Pi^* X\beta^* + \epsilon) \overset{H_0}{=} U^T I X\beta + U^T\epsilon \overset{d}{=} \epsilon \sim N(0, \sigma^2 I)$$

Test whether the elements of ξ follow i.i.d. normal by applying Kolmogorov-Smirnov test or Cramer-von-Mises test.

Slawski et al. (2019) also derives an asymptotic distribution under a restricted model setting:

1. True permutation in the data is chosen uniformly from $\mathcal{P}_n(k) = \{\pi \in \mathcal{P}_n : \sum_{i=1}^n I(\pi(i) \neq i) = k\}$, the set of permutations with only k mismatch.
2. Conditional on π^*, the pairs $\{x_{\pi^*(i),y_i}\}_{i=1}^n$ are i.i.d. zero-mean random variables drawn from a joint density $f_{x,y}(x,y)$ s.t.

$$f_{x,y}(x,y) = f_{y|x}(y|x) \times f_x(x)$$
$$f_{y|x} \sim N(x^T\beta^*, \sigma_*^2)$$
$$f_x \sim N(0, I)$$

Pseudo-likelihood based inference

Define indicator $z_i = I(\pi^*(i) \neq i)$. Then from previous assumptions

$$y_i|\{x_i, z_i = 0\} \sim N(x_i^T \beta^*, \sigma_*^2)$$
$$y_i|\{x_i, z_i = 1\} \sim f_y$$

where $f_y \sim N(0, \|\beta^*\|_2^2 + \sigma_*^2)$. Then $y_i|x_i$ can be expressed as a mixture model with mixing parameter $\mathbf{P}(Z_i \neq 1) = \alpha_* = \frac{k}{n}$.

$$f_{y_i|x_i} \sim (1 - \alpha_*)N(x_i^T \beta^*, \sigma_*^2) + \alpha_* N(0, \|\beta^*\|_2^2 + \sigma_*^2)$$

Pseudo-likelihood based inference

Formulate a pseudo-likelihood function of $\theta = (\beta, \sigma^2, \alpha)$

$$L(\theta) = \sum_{i=1}^{n} \log f_{y_i|x_i}(y_i|x_i;\theta).$$

It is a pseudo-likelihood because $\{y_i|x_i\}_{i=1}^n$'s are not independent. Nevertheless, the maximizer $\hat{\theta}$ of $L(\theta)$ enjoys several attractive properties.

Theorem 6

Under the regularity conditions for the theory of pseudo-likelihood,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, H_*^{-1} G^* H_*^{-1})$$

where

$$H_* = \mathbf{E}[-\nabla_\theta^2 \log f(y|x;\theta^*)]$$
$$G^* = \mathbf{E}[\nabla_g \log f(y|x;\theta^*) \nabla_g \log f(y|x;\theta^*)^T]$$

Reference

N. H. Nguyen and T. D. Tran. Robust lasso with missing and grossly corrupted observations. *IEEE transactions on information theory*, 59 (4):2036–2058, 2012.

A. Pananjady, M. J. Wainwright, and T. A. Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 64 (5):3286–3300, 2018. doi: 10.1109/TIT.2017.2776217.

M. Slawski, G. Diao, and E. Ben-David. A pseudo-likelihood approach to linear regression with partially shuffled data, 2019.

H. Zhang, M. Slawski, and P. Li. The benefits of diversity: Permutation recovery in unlabeled sensing from multiple measurement vectors. *IEEE Transactions on Information Theory*, 2021.