Estimation in exponential family Regression based on linked data contaminated by mismatch error

Zhenbang Wang

George Mason University
Department of Statistics

joint work with Martin Slawski

October 14, 2020

Record Linkage

- A way of collecting information about the same individual by combining data from multiple distinct sources
- Record Linkage is necessary in the sense that it leads to more efficient data collection and lower participant burden and increased information for correction of participant bias due to missing data, such as Linking an established longitudinal data set to one or more administrative registers
- Official Statistics (e.g. US Census Bureau), electronic health and etc

Two types of Record Linkage

Deterministic record linkage

a pair of records is said to be a link if the two records agree exactly on each element within a collection of identifiers called the match variables

Probabilistic record linkage

The linkage is called probabilistic if a record pair is deemed a link with a certain probability.

File A

ID	Sex	ZIP	Income
1	F	94109	20k
2	M	94703	40k
3	M	94701	70k
4	F	94109	30k
5	M	94701	80k

File B

ID	Sex	ZIP	Age
4	F	94109	26
2	M	?	31
3	?	94701	39
5	M	94701	46
1	F	94109	24

Two types of Record Linkage

Fellegi–Sunter model

- Consider File A and B, Product space $A \times B = M \bigcup U$ where M is the set of true matches and U is the set of non-matches
- $R = \frac{\mathbf{P}(\gamma \in \Gamma | r \in M)}{\mathbf{P}(\gamma \in \Gamma | r \in U)}$ where γ is an arbitrary agreement pattern in a comparison space Γ
- For example, consider the matching variables from above example {Sex, ZIP} and $\{\gamma_i\}_{i=1}^{2}$ representing simple agreement(A) or disagreement(D) on these matching variables then the comparsion space Γ consist of the eight (2^2) possible patterns $\{(A, A), (A, D), (D, A), (D, D)\}$
- Fellegi–Sunter Decision Rule : if R $\geq$ Upper, then match, if R $\in [Lower, Upper]$, potential match, if R $\leq$ lower, non-match

Two types of analysis of Linked data

- Primary Analysis: the data analyst and the linker are the same.
- Secondary Analysis : the linked data set that eventually made available to analysts may not contain matching variables.

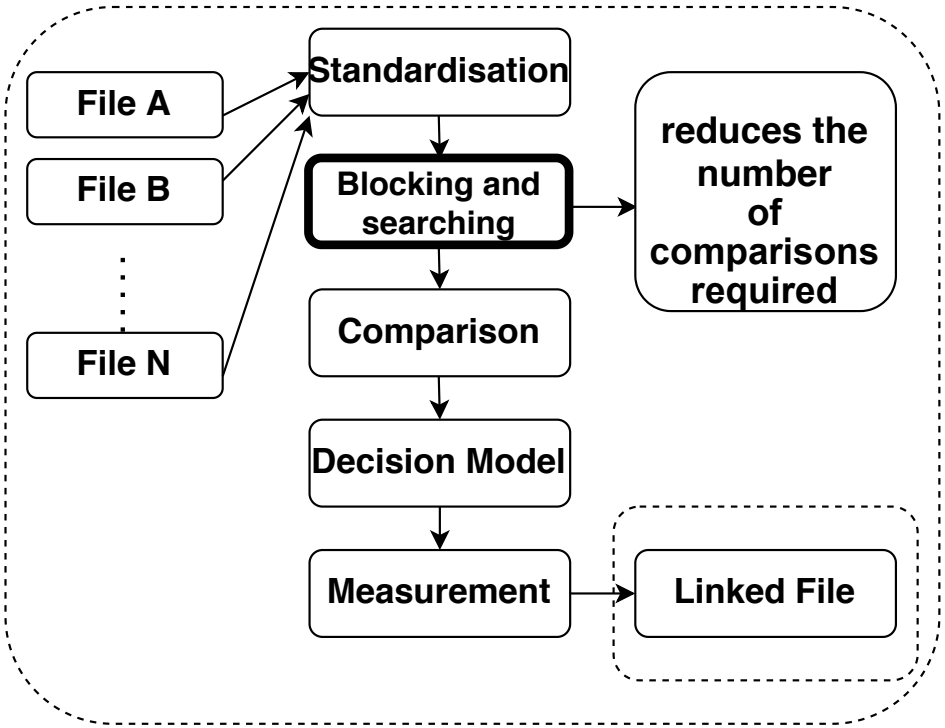

Figure: Large dashed line(Primary Analysis), Small(Secondary Analysis)

Two types of linkage error

Post-linkage Regression Analysis

It is referred as regression analysis after linkage process between response in file A and predict variables in file B. Note that it can be performed under both primary and secondary analysis. The goal is to address bias in the analysis of linked files resulting from a potentially flawed linkage process.

Linkage Error

- Mismatches (false positive): non-matches deemed as matches.
- Missed matches (false negative): matches deemed as non-matches.

We will focus on Mismatches error under Post-linkage Regression Analysis

Set Up and Assumptions

- The response ($\{y_i\}_{i=1}^N \subset \mathbb{R}$) and the predictor variables ($\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$) are contained in two files F_A and F_B.
- The linkage was complete and one-to-one (i.e. n = N) and Record linkage yields a merged file $F_{A \bowtie B} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- Each $\mathbf{x}_i$ is associated with a corresponding latent response variable y_i^*, where $y_i = y^*_{\pi^*(i)}$ .The relationship between the observed samples and the underlying latent sample is given by the following

$$\mathbf{Y} = \Pi^* \mathbf{Y}^*, \text{ where } \Pi^*_{ij} = \begin{cases} 1 & \text{if } y_i = y_j^*, 1 \le i \le n, 1 \le j \le n \\ 0 & \text{otherwise} \end{cases}$$

Example

File $A \bowtie B$ with mismatches

ID	Income	Age
1	30k	24
2	40k	31
3	70k	39
4	20k	26
5	80k	46

File $A \bowtie B$

ID	Income	Age
1	20k	24
2	40k	31
3	70k	39
4	30k	26
5	80k	46

File A⋈B with mismatches **FileA⋈B**

$$\begin{pmatrix} 30k \\ 40k \\ 70k \\ 20k \\ 80k \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 20k \\ 40k \\ 70k \\ 30k \\ 80k \end{pmatrix}$$

$\mathbf{Y}$ Π^* $\mathbf{Y}^*$

Oracle, Naive, Lahiri-Larsen and Chamber's estimator of β^*

Considering the following linear regression model

$$\mathbf{Y} = \Pi^* \mathbf{Y}^*, \quad \mathbf{Y}^* = \mathbf{X}\beta^* + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}_n(\mathbf{0}, I_n)$$

For Post-linkage regression analysis (i.e. only given $(\mathbf{X}, \mathbf{Y})$),

$$\hat{\beta}_{\text{Naive}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} \qquad \hat{\beta}_{\text{Oracle}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}^* \text{ (BLUE)}$$

- $\mathbf{P}(\Pi^*_{ij} = 1 | \Gamma, \mathbf{X}) = \mathbf{Q}_{ij}$
- $\mathbf{E}(\mathbf{Y}|\mathbf{X}, \Gamma) = \mathbf{E}(\Pi^*\mathbf{Y}^*|\mathbf{X}, \Gamma) = \mathbf{E}(\Pi^*|\mathbf{X}, \Gamma)E(\mathbf{Y}^*|\mathbf{X}, \Gamma) = \mathbf{Q}\mathbf{X}\beta^*$

$$\hat{\beta}_{\text{LL}} = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top \mathbf{Y}, \widetilde{\mathbf{X}} = \mathbf{Q}\mathbf{X} \qquad \hat{\beta}_C = (\mathbf{X}^\top \mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$$

MLE and Robust Estimation of Π^* and β^*

Denote $\mathcal{P}(n)$ as the class of permutation on $(1, \ldots, n)$ and $d_{\mathrm{H}}(\cdot, \cdot)$ as Hamming distance.

- $(\hat{\Pi}, \hat{\beta}) = \underset{\Pi \in \mathcal{P}(n), \beta \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{Y} - \Pi \mathbf{X} \beta\|_2^2$
- $(\hat{\Pi}, \hat{\beta}) = \underset{\Pi \in \mathcal{P}(n), \beta \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{Y} - \Pi \mathbf{X} \beta\|_2^2 \quad \text{subject to} \quad d_{\mathrm{H}}(\Pi, I_n) \leq k$
- Define $f = (\Pi - I_n) X \beta$ and consider a convex relaxation

  $$(\widehat{f}, \widehat{\beta}) = \underset{f \in \mathbb{R}^n, \beta \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\beta - f\|_2^2 + \lambda_n \|f\|_1$$

  The ℓ_1- penalty enforce that most of the $\{\xi\}_{i=1}^n$ are set to zero based on the assumption that only a small fraction of the observations is subject to mismatch error

Set Up and Assumptions

We will study Post-linkage Regression Analysis under exponential family(GLM)

- $f(y_i;\vartheta_j) = \exp\left\{\frac{y_i\vartheta_j - \psi(\vartheta_j)}{a(\phi)} + c(y_i,\phi)\right\}$
- GLMs with canonical link, i.e., $\vartheta_i = \eta_i := \mathbf{x}_{\pi(i)}^\top \beta^*$
- The oracle estimator and naive estimator (MLE):

$$\textbf{Oracle} \ : \ \mathbf{X}^\top(\mathbf{Y}^* - \mu(\beta)) = 0 \qquad \textbf{Naive} \ : \ \mathbf{X}^\top(\mathbf{Y} - \mu(\beta)) = 0$$

- Lahiri-Larsen and Chamber's estimator :

$$\textbf{LL} \ : \ \mathbf{X}^\top\mathbf{Q}^\top(\mathbf{Y} - \mathbf{Q}\mu(\beta)) = 0 \qquad \textbf{C} \ : \ \mathbf{X}^\top(\mathbf{Y} - \mathbf{Q}\mu(\beta)) = 0$$

Robust Estimation Approach

Considering the MLE of GLM with constraint on the mismatches,

$$(\widehat{\beta}, \widehat{\Pi}) = \underset{\beta \in \mathbb{R}^d, \Pi \in \mathcal{P}(n)}{\operatorname{argmin}} \; -\langle \Pi \mathbf{X}\beta, \mathbf{Y} \rangle + \sum_{i=1}^{n} \psi(\mathbf{x}_{\pi(i)}^{\top}\beta) \;\text{ subject to }\; d_{\mathrm{H}}(\Pi, I_n) \leq k,$$

Similar to linear regression, we consider the convex relaxation given by

$$(\widehat{\beta}, \widehat{\xi}) = \underset{\beta \in \mathbb{R}^d, \xi \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{n} \left\{ -\langle \mathbf{X}\beta + \sqrt{n}\xi, \mathbf{Y} \rangle + \sum_{i=1}^{n} \psi(\mathbf{x}_i^{\top}\beta + \sqrt{n}\xi_i) \right\} + \lambda \|\xi\|_1,$$

where $\sqrt{n}\xi = (\Pi - I_n)X\beta$

Block coordinate descent algorithm

We adopt the "alternating" strategy for estimation, denote

$$\ell_{\text{pen}}(\xi, \beta) = \frac{1}{n}\left\{-\langle \mathbf{X}\beta + \sqrt{n}\xi, \mathbf{Y}\rangle + \sum_{i=1}^{n} \psi(\mathbf{x}_i^\top \beta + \sqrt{n}\xi_i)\right\} + \lambda\|\xi\|_1$$

1. Update for ξ ("Soft-thresholding"): $\hat{\xi} = \operatorname{argmin} \ell_{\mathbf{pen}}(\xi, \hat{\beta})$

$$\widehat{\xi}_i^{(t+1)} \leftarrow \mathbf{1}\left\{|y_i - \hat{\mu}_i^{(t)}| > \sqrt{n}\lambda\right\} \cdot \left((\psi')^{-1}(y_i - s_i\lambda) - \widehat{\eta}_i^{(t)}\right)/\sqrt{n}$$

2. Update for β (One-step IRLS) : $\hat{\beta} = \operatorname{argmin} \ell_{\mathbf{pen}}(\hat{\xi}, \beta)$

$$\widehat{\beta}^{(t+1)} \leftarrow \widehat{\beta}^{(t)} + (\mathbf{X}^\top W^{(t)}\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{Y} - \psi'(\mathbf{X}\widehat{\beta}^{(t)} + \sqrt{n}\widehat{\xi}^{(t+1)}))$$

Inference

Theorem

(Preliminary) Consider the proposed estimator $\widehat{\theta} = (\hat{\beta}^\top, \hat{\xi}^\top)^\top$ *with* $\lambda \propto \sqrt{\frac{\log(n+d)}{n}}$. *Under some regularity condition, it holds that*

$$\|\widehat{\theta} - \theta^*\|_2 \leq C_1 \sqrt{\frac{d+k}{n}}$$

with probability at least $1 - C_2/n$, *where* C_1 *and* C_2 *are some known constant.*

Selection of Regularization Parameters

$$\lambda = C \cdot \sigma_y \cdot \sqrt{\frac{\log(n+d)}{n}}$$

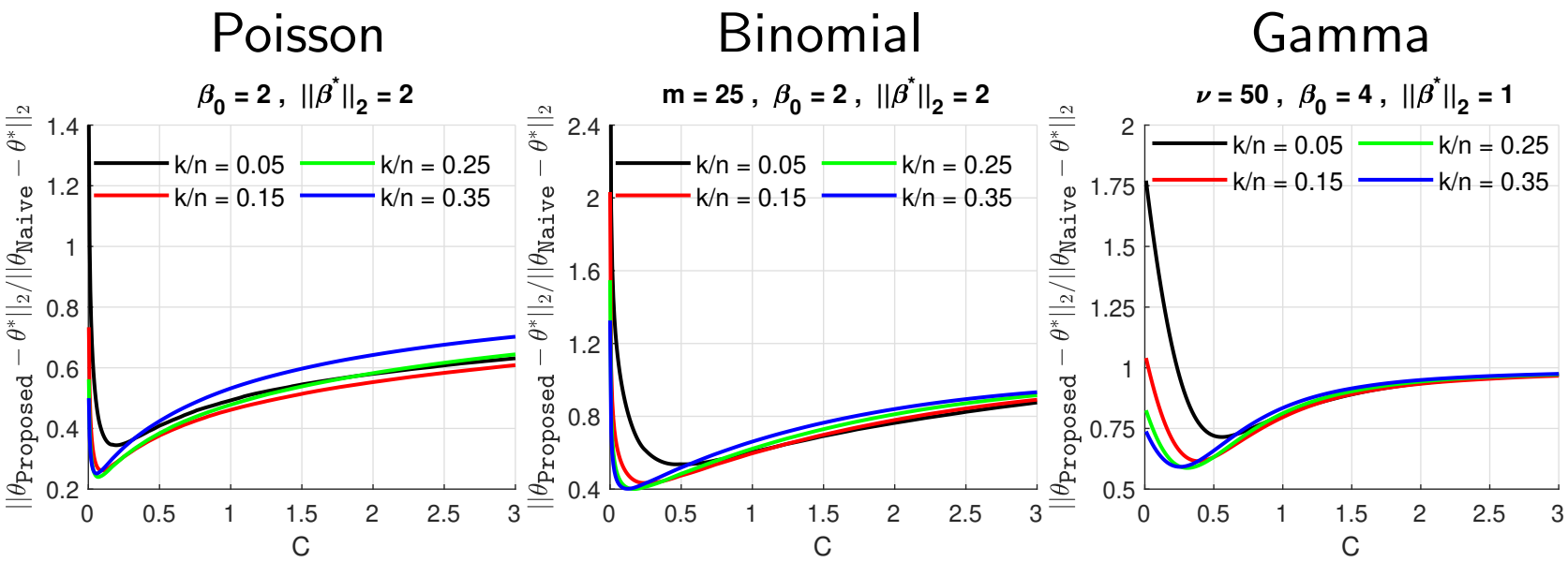

Figure: Estimation error ratios $\|\widehat{\theta} - \theta^*\|_2 / \|\widehat{\theta}^{\text{naive}} - \theta^*\|_2$ in dependence of the pre-factor C appearing in the tuning parameter λ.

Average estimation errors $\|\beta^{\text{est}} - \beta^*\|_2$

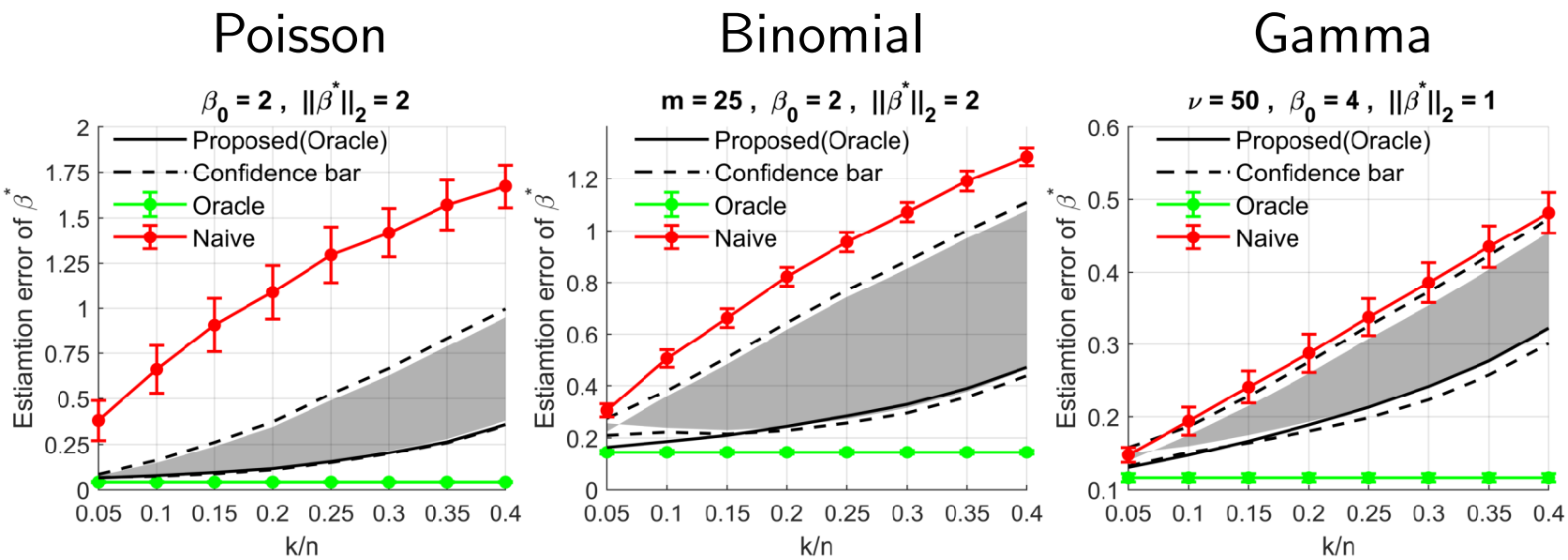

Figure: The lower and upper boundary of the shaded area show the minimum and maximum error over all choices of the pre-factor $C \in [0.1, 2]$, and the corresponding dashed lines represent $\pm 5 \times$ standard error.

Average deviances between μ^* and μ^{est}.

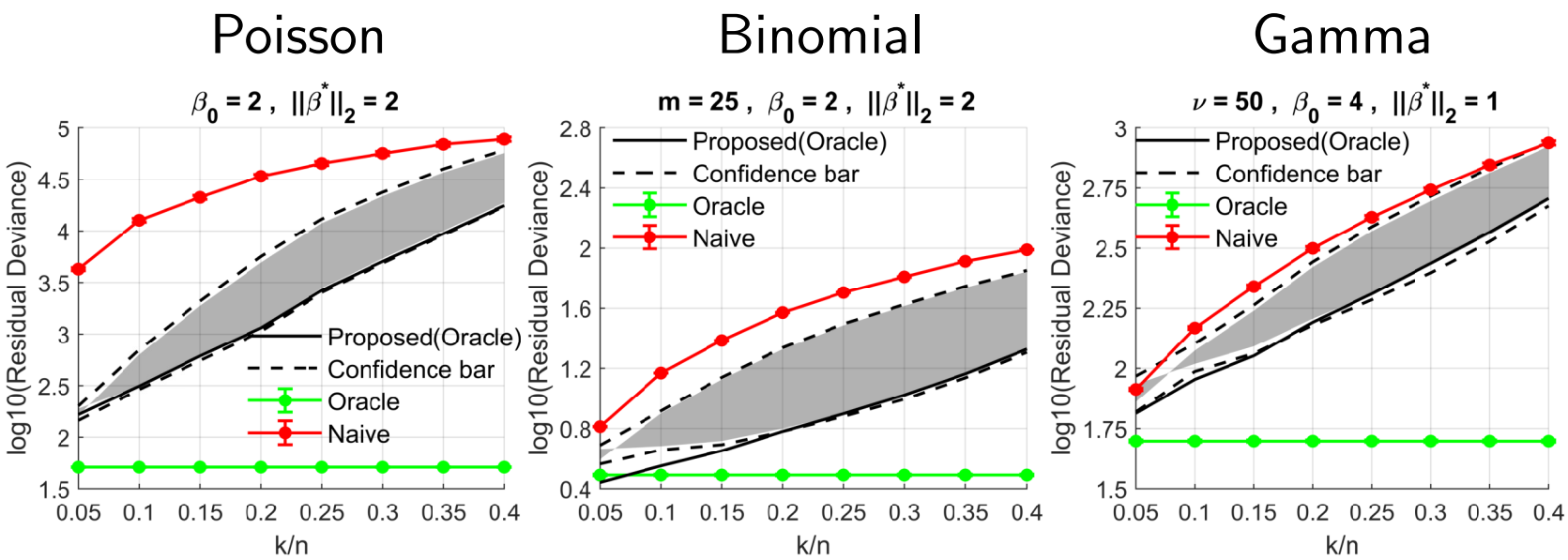

Figure: The lower and upper boundary of the shaded area show the minimum and maximum error over all choices of the pre-factor $C \in [0.1, 2]$, and the corresponding dashed lines represent $\pm 5 \times$ standard error.

- Estimation of Π^*
- Real data analysis of proposed method
- M. Slawski and E. Ben-David, "Linear regression with sparsely permuted data," Electronic Journal of Statistics, vol. 13, pp. 1–36, 2019.
- Wang, Z., Ben-David, E., Slawski, M. (2020). Estimation in exponential family Regression based on linked data contaminated by mismatch error. arXiv preprint arXiv:2010.00181.
- Thank you for your attention.