Qualifying Paper Report for Data Fission: Splitting A Single Data Point

Naitong Chen

January 10, 2023

1 Problem Definition

Given a dataset $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \pi_\theta$, with π_θ being a distribution from the exponential family whose parameter θ is of interest. We decompose each X_i to $f(X_i)$ and $g(X_i)$ such that both parts contain information about θ, and there exists some function h such that $X_i = h(f(X_i), g(X_i))$ satisfying either of the two properties:

- (P1): $f(X_i)$ and $g(X_i)$ are independent with known distributions (up to unknown θ);
- (P2): $f(X_i)$ has a known marginal distribution and $g(X_i)$ has a known conditional distribution given $f(X_i)$ (up to unknown θ).

2 Significance

3 Limitations and challenges

- paper does not discuss robustness of the method to the distribution assumptions
- experiments do not cover cases where fissioned data get transformed to follow a different distribution
- paper does not discuss how to choose tuning parameter (controlling amount of information split between $f(X)$ and $g(X)$)
- following the previous point, this paper does not discuss the relations between having a discrete vs. continuous tuning parameter (e.g. the two different ways of fissioning exponentially distributed data in Appendix B of Leiner et al. (2022))

4 Paper-specific project

I noticed that most of the experiments and simulation studies in the paper only cover cases where the distributions of $f(X)$ and $g(X) \mid f(X)$ are in the same family as the original data X (i.e. Gaussians or Poissons with different parameters), and there is minimal discussion on when this is not the case. I would like to therefore focus my paper-specific project on an instance of data fission where $f(X)$ follows a distribution that is not in the same family as X or $g(X) \mid f(X)$.

The particular case that I would like to focus on is to **construct selective CIs in the fixed-design GLM model** where $Y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Exp}(\exp(\theta_i)$ where $\theta_i = \beta^\top x_i)$ (listed under Appendix B of Leiner et al. (2022)).

- For each Y_i with $i \in \{1, \ldots, n\}$, draw $Z_i = (Z_{i1}, \ldots, Z_{iB})$ where each element is i.i.d. $Z_{ib} \sim \mathsf{Poiss}(Y_i)$ with $b \in \{1, \cdots, B\}$.
  Then $f(Y_i) = Z_i$, where each element is i.i.d. $\mathsf{Geom}\left(\frac{\theta_i}{\theta_i + B}\right)$,
  $g(Y_i) \mid f(Y_i)$ has conditional distribution $\mathsf{Gam}(1 + \sum_{b=1}^{B} [f(Y_i)]_b, \theta_i + B)$.
- For each Y_i, draw $Z_i \sim \mathsf{Poiss}(\tau Y_i)$ with $\tau \in (0, \infty)$.
  Then $f(Y_i) = Z$, where each element is i.i.d. $\mathsf{Geom}\left(\frac{\theta_i}{\theta_i + \tau}\right)$,
  $g(Y_i) \mid f(Y_i)$ has conditional distribution $\mathsf{Gam}(1 + f(Y_i), \theta_i + \tau)$.

The proposed procesure is

- Decompose each y_i using one of the two above procedures (assuming we are going with the first one)
- fit $f(y_i)$ by maximizing

$$\sum_{i=1}^{n} \sum_{b=1}^{B} \log \mathsf{Geom}\left(z_{ib} \mid \frac{\beta^\top x_i}{\beta^\top x_i + B}\right) + \lambda \|\beta\|_1$$

  (I've verified that this function is convex in β.)
- Fit a Gamma GLM model using just the selected features from the previous step
- Since Gam is in the exponential dispersion family, we can follow the exact same setup as in Appendix A.5 of Leiner et al. (2022) to use the QMLE procedure to construct CIs in the inference step.

Key areas that I would like to explore (through simulations)

- compare the variables selected using decomposed data following geometric distribution against those using data splitting and the (invalid) approach of using the same dataset twice for both selection and inference. The setup that I have is mostly inline with the simulation studies in Sections 4 and 5 in Leiner et al. (2022), however, I would like

to ensure there is no influential point and all assumptions are met. This way we can better isolate the effect of transforming the original dataset to something that follows a different distribution.

- compare the two data fission procedures laid out above with discrete and continous tuning parameters (B and τ) for deciding how much information is split between $f(Y)$ and $g(Y) \mid f(Y)$. The second version seems like a continuous relaxation (under expectation) of the first data fission method. However, in the case with a discrete tuning parameter, the dimension of $f(Y_i)$ changes and we need to somehow account for that in the selection step. This is not addressed directly in the paper, but I think a natural way to deal with this is to think of stacking the elements of $f(Y_i)$ so that for each particular set of covariates x_i, we have multiple corresponding responses instead of 1. I would like to explore the connection between these two fission processes (probably empirically) in terms of the amount of information allocated in each component of the fissioned data. The same simulation setup from the previous bullet point can be used here.

- check the robustness of this procedure with respect to distributional assumptions. Maybe instead of generating data actually from an exponential distribution, we can generate data using one of a different shape, for example, the log normal distribution.

5 Discussion

References

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *arXiv preprint arXiv:2112.11079 v4*, 2022.