A Note on Multiplicative Noise Perturbation for Privacy Protection

Xiaoyu Zhai

George Washington University

zhaixiaoyu1990@gwu.edu

This is based on my Ph.D. dissertation research under Professor Tapan Nayak's supervision.

February 28, 2018

Randomized Response Background

- **Motivation**: In surveys with sensitive variables, data agencies need to protect respondents' privacy to encourage truthful answers.
- **Privacy** is an individual's freedom from unauthorized intrusion. **Privacy protection** is hiding a respondent's true values from others, including the interviewer.
- **Solution**: Collect individual's randomized response (RR) of his true value for privacy protection.(Warner (1965))

Randomized Response for Quantitative Variables

- Denote Y as quantitative survey variable with mean μ_y and variance σ_y^2, which are unknown.
- Denote V is the noise variable with finite mean θ and variance γ^2.
- V is independent of Y.
- Denote Z as the perturbed version of Y.

Series of Generalized Models

- Pollock and Beck (1976) Model: $Z = YV$
- Bar-lev, Bobovitch and Boukai (2004) Model (BBB Model):

$$Z = \begin{cases} YV & \text{w.p. } 1-p, \\ Y & \text{w.p. } p \end{cases} \tag{1}$$

- Ryu, Kim, Heo and Park (2005) Model:

$$Z = \begin{cases} YV & \text{w.p. } (1-p)(1-\alpha), \\ Y & \text{w.p. } p+(1-p)\alpha, \end{cases} \tag{2}$$

- Singh and Tarray (2016) Model:

$$Z = \begin{cases} Y[(1-m)V + m\theta(\frac{V-\theta}{\gamma})^2] & \text{w.p. } 1-p, \\ Y & \text{w.p. } p, \end{cases} \tag{3}$$

- Tarray and Singh (2017) Model:

$$Z = \begin{cases} Y\left[\frac{aV+b\theta}{a+b}\right] & \text{w.p. } 1-p, \\ Y & \text{w.p. } p, \end{cases} \tag{4}$$

Note: m, a, b, p are known constants and $p \in [0,1]$.

Series of Generalized Models Continued

Truly, these are specializations rather than generalizations. Denote general multiplicative model as $Z = YS$.

- In BBB model, $S \sim p\delta(1) + (1-p)f_V(v)$.
- In Ryu et al. model, $S \sim (p + (1-p)\alpha)\delta(1) + (1-p)(1-\alpha)f_V(v)$.
- In Singh and Tarray model, $S \sim p\delta(1) + (1-p)f_{V_1}(v_1)$ where $V_1 = mV + (1-m)\theta(\frac{V-\theta}{\gamma})^2$.
- In Tarray and Singh model, $S \sim p\delta(1) + (1-p)f_{V_2}(v_2)$ where $V_2 = \frac{aV+b\theta}{a+b}$.

Motivation

Gaps that motivated our work are the following.

- Past papers compared these models by restricting some common features and only compared variance inflation.
- It seems researchers view the value of p as the main privacy measure, which we consider inadequate. No privacy comparison or explicit privacy measures were investigated.
- They considered mostly infinite population, not finite population.

Method of Moment Estimators

Multiplicative Model:

$$Z = YS \tag{5}$$

where S denotes the noise variable with mean μ and variance σ^2. And S is independent of Y.

- An adaptation of Horvitz Thompson estimator of finite population total based on perturbed data is

$$T = \sum_{i=1}^{n} \frac{1}{\pi_i} \frac{Z_i}{\mu}, \tag{6}$$

  where π_i is the inclusion probability for element i. T is an unbiased estimator of population total of Y.
- Consider sampling n *i.i.d.* observations from infinite population, a moment estimator of infinite population mean is

$$\hat{\mu}_Y = \frac{\sum_{i=1}^{n} z_i}{n\mu} \tag{7}$$

.

Trade Off Between Privacy and Data Utility

Theorem

Under model (5), the variance inflation of the unbiased estimator T and that of $\hat{\mu}_Y$ are only determined through $\frac{\sigma^2}{\mu^2}$.

$$Var(T) = \frac{\sigma^2}{\mu^2}\sum_{i\in\Omega}\frac{1}{\pi_i}Y_i^2 + \frac{1}{2}\sum\sum_{i\neq j\in\Omega}(\pi_i\pi_j - \pi_{ij})(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}). \quad (8)$$

$$Var(\hat{\mu}_Y) = \frac{1}{n}(\sigma_y^2 + \frac{\sigma^2}{\mu^2}E(Y^2)) = \frac{1}{n}(\sigma_y^2 + \frac{\sigma^2}{\mu^2}(\sigma_y^2 + \mu_y^2)). \quad (9)$$

In the series of models, comparisons were made based only on $Var(T)$ under common features. For example: Singh and Tarray (2016) showed for some m, Singh and Tarray model is better than BBB model given the common V and p. In fact, V is not needed to construct V_1. Singh and Tarray claimed that given a BBB model, one can construct a better model. In fact, the converse is true!

Trade Off Between Privacy and Data Utility Continued

- Proposition 1: Let $Q = \frac{Z}{\mu}$. Then, the two proposed privacy measures $\rho^2_{QY} = \frac{Var(Y)}{E(Y^2)(\frac{\sigma^2}{\mu^2}+1)-E(Y)^2}$ and $E(Q-Y)^2 = E(Y^2)\frac{\sigma^2}{\mu^2}$ are also only determined through $\frac{\sigma^2}{\mu^2}$.
- **Privacy and statistical efficiency have 1-1 correspondence. There is no room for efficiency optimization for a required privacy level.**

Optimal Choice of K in Modified BBB model

Modified BBB model is

$$Z_i = \begin{cases} Y_i V & \text{w.p. } 1-p, \\ kY_i & \text{w.p. } p, \end{cases} \tag{10}$$

where k is a constant, and V is a positive random variable. V is independent of Y.

- We proved that the optimal k which maximize the estimation efficiency, i.e.,the minimizer of $\frac{\sigma^2}{\mu^2} = \frac{pk^2+(1-p)(\gamma^2+\theta^2)}{(pk+(1-p)\theta)^2} - 1$ is $k_0 = \frac{E_R(V^2)}{E_R(V)} = \theta + \frac{\gamma^2}{\theta}$

Future work

Our overall goal is to define privacy, compare and choose randomization mechanism at a fixed privacy level. Here are a few ideas.

- Suppose S_1 is defined in (10) and S_2 is a continuous variable with with mean μ_i and variance σ_i^2 separately, $i = 1, 2$. If they satisfy $\frac{\sigma_1^2}{\mu_1^2} = \frac{\sigma_2^2}{\mu_2^2}$, Proposition 1 implies the two models provide same privacy protection, which are clearly different from respondents' perspective. For example, if $k = 2$ and $z = 100$ and $y = 50$ has $p = 0.2$ prob. to be the true value. This model exposes the true value easily compared to a continuous random variable hence might lower the respondent's belief for privacy protection. In a nutshell, the existing privacy measures based on first two moments are not satisfactory.

Future Work Continued

- Pick the point mass k outside of the range of continuous noise variable is not a good idea for protecting privacy. Suppose S is $Uniform[2, 5]$, $k = 10$ and $p = 0.4$. If $y = 10$, and $z = 100$ is one possible value of perturbed variable Z. Without prior information, the intruder's may predict $\hat{y}$ as $\{10$ w.p. 0.4 and $(20, 50)$ w.p. $0.6\}$. If the intruder has prior information that the true response is in $(5, 17)$, then the true value 10 will be identified with probability 1.
- We observe that length of confidence interval is not an adequate measure for privacy. New privacy measure is in need. Suppose we change k to 4, which is in the range of the continuous distribution. If $z = 40$, then without prior information, the intruder may predict $\hat{y}$ as $\{10$ w.p. 0.4 and $(8, 20)$ w.p. $0.6\}$. The point mass in the interval is difficult to deal with if we use confidence interval for privacy measure.

Thank you very much.