Bayesian Randomized Response Technique

Mike K.P. SO[1] and Ray S.W. CHUNG[1]
[1]The Hong Kong University of Science and Technology, Hong Kong
Corresponding author: Ray S.W. CHUNG, email: swchungaa@ust.hk

Abstract

When sensitive attributes are investigated, Randomized Response Technique (RRT) is a popular approach to reduce the bias arisen from untruthful response. Nonetheless, traditional RRT has weakness that it mainly focuses on estimating the moments of univariate random variables but not dependence among multiple random variables. This paper is to introduce a new method to estimate the covariance matrix of random vectors under the framework of RRT. Modified Cholesky decomposition is applied to reparameterize the covariance matrix so that the parameters of the covariance matrix can be expressed as regression on a row-by-row basis. This simplifies the structure of the matrix and ensures the positive definiteness of the estimator. To keep inference nonparametric, moment equations of the randomized realizations are adopted as the quasi-likelihood. Moreover, Bayesian lasso is applied to impose shrinkage effect in estimation. This helps reduce estimation error when the covariance matrix is sparse. An easy-to-implement Gibbs sampling scheme is proposed for the inference. A simulation study is conducted to evaluate the accuracy of estimation. An empirical study related to software piracy behavior is conducted to compare the difference of the estimates under randomized and non-randomized settings.

Keywords: Bayesian lasso analysis; modified Cholesky decomposition; sensitive responses; shrinkage.

1. Introduction

When a sensitive question like “Do you have another lover other than your official spouse?” is asked in a face-to-face survey, respondents may give incorrect response to hide the truth. This induces undesired bias to estimate. To overcome the problem, Warner (1965) introduced Randomized Response Technique (RRT), which imposes noise with known property to distort the response. Interviewers never know the true response to the sensitive question under the randomization, and this is expected to encourage truthful response from the respondents. On the other hand, as the probabilistic property of the noise is well known, the property of the sensitive attributes among the whole population can be inferred statistically even though the exact responses from individuals are not known.

The settings of Warner (1965) can cope with binomial or multinomial distribution only (Abdel-Latif et al., 1967). However, the interested attributes in surveys are not limited to these two distributions. Few years later, Greenberg et al. (1969) proposed an Unrelated Question Approach (UQA) to incorporate RRT in a broader framework. Its settings are given as follows. Assume that the number of pirate software used by people from a certain population is of interest. To impose noise to the response, the interviewers do not directly ask for the response. Instead, they ask respondents to draw a ball from a box with only red balls and blue balls beforehand. The colors of the balls drawn by the respondents are never revealed by the interviewers. The interviewees are informed to answer the question "What is the number of pirate software installed by you?" if a red ball is drawn and answer the question How many family members do you have?" if a blue ball is drawn. Under this procedure, the interviewers do not know which question is answered, so the privacy of the respondents is assured. The randomizing procedure is carried out on two different groups, each with different probability of drawing a red ball. Let $\mathrm{E_i}$ and $\mathrm{p_i}$ be the expected response after randomization and probability of drawing a red ball for Group i, i = 1, 2, respectively. Also, denote the mean responses to the sensitive question and the innocuous question from the interested population are $\mathrm{\mu_Y}$ and $\mathrm{\mu_U}$ respectively. The following moment equations are resulted

$$\begin{cases} \mathrm{E}_1 = p_1\mu_Y + (1-\mathrm{p}_1)\mu_U \\ \mathrm{E}_2 = p_1\mu_Y + (1-\mathrm{p}_1)\mu_U \end{cases}.$$

With the moment equations, μ_Y can be estimated by Method of Moment.

The UQA mainly focuses on estimating the moment of an arbitrary random attribute. As everything relies on moment, no distribution assumption is required, and this enables the UQA to handle different types of data. This nonparametric specification makes the UQA favorable. Please note that for survey problem, the data can be distributed in any form (e.g. binomial, normal, gamma). Thus the nonparametric specification avoids the need of validating the underlying distribution for each attribute. The property is especially important when several sensitive variables are investigated. Nevertheless, the extension of the UQA from univariate settings to multivariate settings is rarely discussed in literature.

One of the most important parameters essential for multivariate analysis is covariance matrix of the investigated random vector. With the covariance matrix, more sophisticated analysis like factor analysis or structural equation modeling can be carried out. Kwan et al. (2009) attempted to extend the UQA to estimate the covariance matrix of a set of sensitive attributes related to software piracy behavior. The extension is simply based on applying Method of Moment to moment equations of higher order. Yet two problems arose. First, the covariance matrix estimated from their method cannot guarantee positive-definiteness. The problem is more obvious when the dimension of the random vector increases. Imposing parametric assumption may be feasible, but a lot of effort is required to determine the underlying distribution of every random element, which is time-consuming. Second, the covariance matrix involved in survey problem is usually sparse, i.e. containing a lot of zero within its entries. As a result, shrinking the estimate of the entries towards zero help reduce the estimation error. Consequently, this paper aims at developing a new method based on UQA to resolve the defect of the method proposed by Kwan et al. (2009) when incorporating UQA into higher dimensional settings.

References

Abul-Ela, A.-L. A., Greenberg, B. G. and Horvitz, D. G. (1967) “A multi-proportions randomized response model,” *Journal of the American Statistical Association*, 62, 990-1008.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R. and Horvitz, D. G. (1969) “The unrelated question randomized response model: theoretical framework,” *Journal of the American Statistical Association*, 64, 520-539.

Kwan, S. S. K., So, M. K. P. and Tam, K. Y. (2010) “Applying the randomized response technique to elicit truthful responses to sensitive questions in IS research: the case of software piracy behavior,” *Information Systems Research*, 21, 941–959.

Warner, S. L. (1965) “Randomized response: a survey technique for eliminating evasive answer bias,” *Journal of the American Statistical Association*, 60, 63-69.