Approximate Bayesian Computation Estimator for Respondent-Driven Sampling

Rui Fu
Group #90
Stanford University
ruifstanford.edu

1 Abstract

Respondent-driven sampling is a network-based technique to collect information and make estimation about behavior and composition of social groups in hidden population. The non-randomly selected samples prohibit the use of the sample mean as a statistically valid estimator. Researchers have proposed several asymptotically unbiased estimators, but many fail to realize that the high variance of these estimators inevitably leads to attenuated performance. We propose to use a Bayesian estimator in the hope of achieving better precision by reducing variance of the estimator.

2 Introduction

The problem of collecting accurate information about hidden populations, such as injection drug users (IDU) and commercial sex workers, arises in many areas of research. The simple random sampling (SRS) and estimation techniques necessarily fails because: 1) these populations are hard to reach due to their sensitive nature, which is why they are often called 'hidden', 2) even if researchers manage to locate some of the target population, the basic assumption of SRS that each unit is drawn with the same probability would most likely be violated. In light of this, a new sampling method, called respondent-driven sampling (RDS), was introduced [1]. The basic idea is that respondents are selected from the social network of the existing members of the sample. The sampling process begins with a small number of seeds selected by researchers to participate in the study. Then these seeds are provided with coupons to recruit their friends to participate in the study. The process of existing sample members recruiting future sample members continues until the desired sample size is reached.

This chain-referral methods have been proved to be effective at penetrating hidden populations. However, the network-based property of the sampling procedure produces samples that are not even close to simple random samples in two senses. First, members are drawn with different probabilities. Obviously the more friends (i.e. higher degree) a member owns, the better chance he/she stands to be recruited into the study. Second, the samples are correlated. Two members of the population are more likely to form affiliation when they share certain characteristics (e.g. religion, age, race). To illustrate, Figure 1 is the recruitment network from a study of drug users in New York, and nodes are colored by race. It can be seen that the mixing of different colored nodes is assorted, rather than homogeneous. Affiliations, in turn, may enhance homogeneity (e.g. spread of beliefs and behaviors in social networks). Both scenarios result in similarity of neighboring nodes in the social network, hence naturally the responses collected are inevitably correlated.

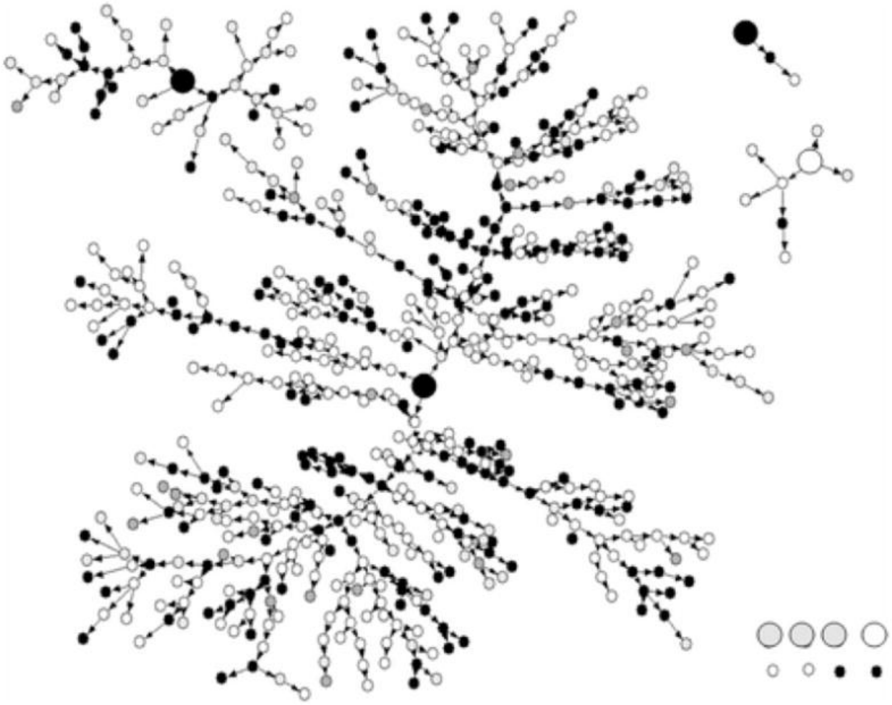

Figure 1: Recruitment networks of an RDS study

The idea of responses being correlated is especially relevant in healthcare domain. Consider the case where researchers need to obtain an estimate of HIV prevalence rate in IDU population. Since injection drug users often share needles and other injection equipment with social network members, and HIV easily transmits via such sharing in the presence of infectious agent, thus the friends (i.e. neighbors in the social network) of an infected IDU are more likely to be infected than friends of a non-infected IDU.

State-of-the-art estimator RDS II [2,3] addresses the problem of non-uniform recruitment probability. And under the assumption of 1) single initial seed, 2) single recruitment, 3) initial seed recruited according to the node degree, the RDS sampling can be fitted into the framework of importance sampling, and the marginal distribution of each sample is the stationary distribution of a random walk on the social network. However the above three assumptions are rarely met in practice: 1) 10 initial seeds is a more common setting, 2) single recruitment often yields a premature termination of the sampling procedure, thus multiple recruitment is almost always adopted, 3) lacking the knowledge of the degree distribution of the hidden population, initial seeds are recruited according to their reachability rather than node degree.

Even when the above idealized sampling requirements are satisfied, Goel *et al.* [4] showed the variance of the estimator is so large, that its performance is greatly attenuated. In view of this, we aim to construct an estimator with smaller mean squared error than RDS II. We seek to exploit the correlation of responses, note this piece of information is not utilized in computing RDS II, but we deem it relevant and informative.

The rest of the paper is organized as follows: Section 3 formulates the problem; Section 4 describes the algorithm for generating simulated ground truth data and constructing our estimator; Section 5 elaborates experimental results to compare the performance of RDS II and our estimator.

3 Problem Formulation

3.1 Notations

Denote by $V = \{X_1, ..., X_N\}$ the vertex set and by $E = \{< X_i, X_j >: i \sim j\}$ the edge set of a social network. Assume the structure of the network follows the configuration model with specified degree distribution which is known. We then depicted a certain trait of nodes (e.g. HIV status) in the social network by letting X_i takes values in $\{+1, -1\}$ for any $1 \leq i \leq n$. Specially, $X_i = +1$ as the i-th people has the trait, while $X_i = -1$ as he/she doesn't.

Of interest is the prevalence rate

$$p = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} \frac{X_i + 1}{2}\right]$$

based on a sub-graph $\mathcal{D}$ which was provided by a respondent-driven sampling procedure on the network

$$V(\mathcal{D}) \subset V, \quad E(\mathcal{D}) \subset E$$

And $\mathcal{D}$ can be represented as a tree or a forest according to the number of seeds in respondent-driven sampling.

3.2 Ising Model

It's natural to characterize both the prevalence rate and the correlated level by the Ising model [5]. In its simplest form, the Ising Model consists of a lattice of binary variables $X_i \in \{-1, +1\}$ that are locally connected horizontally and vertically with pairwise potentials. There can also be an external field applied to the variables that biases them toward a particular state. The total energy of a simple Ising model we consider here is defined as

$$H(X) = -\alpha \sum_{i \in V} X_i - \beta \sum_{<i,j> \in E} X_i X_j$$

where the first sum is over all nodes and the second over all edges of the lattice. α, β are the level of prevalence and the level of pairwise correlation. The probability distribution over states of the lattice is

$$\begin{aligned} p(x) &\propto \exp\left(-H(x)\right) \\ &= \exp\left(\alpha \sum_{i \in V} X_i + \beta \sum_{<i,j> \in E} X_i X_j\right) \end{aligned}$$

As we can see in the form of $p(x)$, the Ising model can be simply extended to a general network of binary variables. And, noting that Ising model defines a two-parameter exponential family with parameter (α, β) and minimal sufficient statistics $(\sum_{i \in V} X_i, \sum_{<i,j> \in E} X_i X_j)$. Reparametrization of exponential family gives

$$(2p-1)N = \mathbb{E}\left[\sum_{i \in V} X_i\right] = \frac{\partial \psi(\alpha, \beta)}{\partial \alpha}$$

$$\rho = \mathbb{E}\left[\sum_{<i,j> \in E} X_i X_j\right] = \frac{\partial \psi(\alpha, \beta)}{\partial \beta}$$

where

$$\psi(\alpha, \beta) = \log\left(\sum_{X \in \{-1,+1\}^N} \exp\left(\alpha \sum_{i \in V} X_i + \beta \sum_{<i,j> \in E} X_i X_j\right)\right)$$

is the cumulative generating function. Although the closed form of ψ is available, it's computationally prohibitive in practice.

4 Methods

4.1 Gibbs Sampling

Gibbs Sampling can be used to draw from the Ising Model a sample $X \in \{-1, +1\}^N$ as the ground truth. Using the local Markov property of Ising Model, one can get conditional probability distributions which play the key role in Gibbs Sampling.

$$\text{logit} P\left(X_i = +1 | X_{V \backslash i}\right) = \exp\left(2\alpha + 2\beta \sum_{<i,j> \in E} X_j\right)$$

With conditional probability distributions, we can carry out Gibbs Sampling.

```
Randomly select $X \in \{-1, +1\}^N$
for $iteration = 1 \ldots m$ do
    for $i \in V$ do
        $X_i \sim 2 \times \text{Ber}\left(P\left(X_i = +1 | X_{V \backslash i}\right)\right) - 1$
    end for
end for
```

When m is sufficiently large, X will "forget" its initial distribution and eventually converges to the equilibrium distribution $p(x)$.

4.2 Respondent-driven Sampling Simulation

Starting from 10 initial seeds drawn randomly from V, the "Snowballing" process of existing sample members recuriting future sample members continues until the desired sample size $n = 500$ is reached. At most 3 neighbors are recruited in the study, which is resonant of the practice in many RDS studies. All information about the nodes and edges in the observed sample is denoted by $\mathcal{D}_{obs}$.

4.3 Approximate Bayesian Computation

As shown in *Section 3.2*, p is a function of $\theta = (\alpha, \beta)$. We would like to estimate $\theta = (\alpha, \beta)$ first. It's an unconventional parameter inference problem since the likelihood function $l(\theta; \mathcal{D}_{obs})$ is unvailable. Approximate Bayesian Computation (ABC) [6] is an algorithm developed for these likelihood-free scenarios. Here is MCMC-based version of ABC algorithm.

```
Initialize $\theta^{(0)}$ from a prior $\pi(\theta)$
for $k = 1 \ldots K$ do
    generate $\theta'$ from Markov kernel $q(\cdot|\theta^{(i-1)})$
    generate $\mathcal{D}'$ by Gibbs sampling in the Ising model with parameter $\theta'$ and respondent-driven sampling
    if thenD' is "similar" to $\mathcal{D}_{obs}$ and Unif$(0,1) < \frac{\pi(\theta')q(\theta^{(i-1)}|\theta')}{\pi(\theta^{i-1})q(\theta'|\theta^{i-1})}$
        let $\theta^{(i)} = \theta', \mathcal{D}^{(i)} = \mathcal{D}'$
    else
        let $\theta^{(i)} = \theta^{(i-1)}, \mathcal{D}^{(i)} = \mathcal{D}^{(i-1)}$
    end if
end for
```

We say $\mathcal{D}$ is "similar" to $\mathcal{D}_{obs}$ if $\rho(\mathcal{D}, \mathcal{D}_{obs}) < \epsilon$ where ρ is a distance metric. The ABC-MCMC algorithm results in a sample of $\theta^{(0)}, ..., \theta^{(K)}$ from approximate posterior distribution

$$\theta_k \sim p\left(\theta | \rho(\mathcal{D}, \mathcal{D}_{obs}) < \epsilon\right) \approx p\left(\theta | \mathcal{D} = \mathcal{D}_{obs}\right)$$

Such a sample from posterior distribution yields an estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$. Applying another standard MCMC approach yields the corresponding $\hat{p}$.

5 Experiments

First we compare the performance of our estimator with that of RDS on simulated data. We generate simulated social networks (with 10,000 nodes) of a hidden population, and we consider two types of network structures, Erdos-Renyi random graph and small world network. After generating the social network, we label each node as 1 or 0 to indicate whether or not an individual has a specific feature, by employing the Ising model with parameters (α, β). The values of (α, β) are found via trial and error, until we obtain a reasonable (p, ρ).

Then we conduct experiments on Project 90 dataset [7], obtained by a study that began in 1987 to examine the influence of network structure on the propagation of infectious disease. We want to test the performance of our estimator when the assumed network structure in the model is not in alignment with the true network structure.

Since the MCMC sampling for posterior estimation is computationally intensive and time-consuming, we run the program on the cluster. For each dataset, 200 trials of experiments are conducted to calculate the mean, variance and mean squared error of estimators. The results are summarized in Table 1 and Table 2.

Table 1: Compare RDS and probabilistic-model-based estimator (PM) on simulated data

p	Graph	Mean(RDS)	Mean(PM)	Var(RDS)	Var(PM)	MSE(RDS)	MSE(PM)
0.0429	SW	0.0423	0.0375	0.00010	0.00007	0.00011	0.00010
0.0418	SW	0.0427	0.0369	0.00012	0.00006	0.00012	0.00008
0.0167	SW	0.0174	0.0163	0.00005	0.00003	0.00005	0.00003
0.0067	SW	0.0062	0.0071	0.00002	0.00003	0.00002	0.00001
0.0749	ER	0.0744	0.0556	0.00043	0.00012	0.00043	0.00049
0.0304	ER	0.0290	0.0155	0.00014	0.00003	0.00014	0.00015
0.0296	ER	0.0206	0.0289	0.00015	0.00003	0.00015	0.00011

Table 2: Compare RDS and PM on Project 90 dataset

p	Mean(RDS)	Mean(PM)	Var(RDS)	Var(PM)	MSE(RDS)	MSE(PM)
0.4336	0.4157	0.4043	0.0022	0.0001	0.0025	0.0005
0.0562	0.0642	0.1619	0.0007	0.0023	0.0008	0.0161
0.0162	0.0199	0.0674	0.0001	0.0006	0.0001	0.0026
0.0867	0.1091	0.0850	0.0029	0.0006	0.0034	0.0007
0.0683	0.0874	0.1840	0.0011	0.0015	0.0014	0.0169
0.0088	0.0096	0.0210	0.0001	0.0001	0.0001	0.0001
0.0237	0.0298	0.0654	0.0003	0.0002	0.0003	0.0016
0.0307	0.0355	0.0407	0.0003	0.0001	0.0003	0.0001
0.0646	0.0605	0.0719	0.0005	0.0004	0.0005	0.0004
0.0442	0.0453	0.0812	0.0003	0.0003	0.0003	0.0012
0.1736	0.1817	0.3326	0.0013	0.0011	0.0013	0.0339
0.0131	0.0177	0.0271	0.0002	0.0001	0.0002	0.0003

We can see that RDS II suffers from large variance while our estimator suffers large bias. In terms of mean squared error, for the small world networks, our estimator consistently outperforms than RDS II. But for the random graph and real network, our estimator beats RDS II only half of the times.

6 Conclusion

In this project, we investigated the inference problem related to respondent-driven sampling. We provided a new perspective by resorting to probabilistic models, with the intuition that incorporation of the relationship between sampled individuals might improve the performance of the estimator. To generate simulated networks, Ising model is employed. However, experiments show that our estimator suffers from large bias, it almost always underestimates the prevalence rate. One future direction is to investigate on the sources of the large bias for our method, and how the underlying network structure affects the performance of the estimator.

References

[1] Heckathorn DD (1997) Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Probl* 44:174-199.

[2] Volz E, Heckathorn DD (2008) Probability based estimation theory for respondent-driven sampling. *J Off Stat* 24:79-97.

[3] Goel S, Salganik MJ (2010) Assessing Respondent-Driven Sampling. *Proc Natl Acad Sci* 107:6743-6747.

[4] K. Binder (2001) Ising model. *Hazewinkel, Michiel, Encyclopedia of Mathematics*, ISBN 978-1-55608-010-4

[5] Salganik MJ (2006) Variance Estimation, Design Effects and Sample Size Calculations for respondent-driven sampling. *J Urb Health* 83:i98i112.

[6] Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci* 100: 15324-15328.

[7] Klovdahl AS (1994) Social networks and infectious disease: The Colorado Springs study. *Soc Sci Med* 38: 79-88.