Bayesian Inference from Non-Ignorable Network Sampling Designs

Simón Lunagómez[1], Edoardo Airoldi[1,2]
[1]Harvard University Department of Statistics, [2]Harvard FAS Center for Systems Biology

Introduction

Consider individuals interacting in a social network and a response that can be measured on each individual. We are interested in making inferences on a population quantity Q that is a function of both the response and the social interactions. In this work, we rely on the notion of ignorability for coarse data proposed by Heijtan and Rubin. This notion is the key element for assessing when to include the functional form of the sampling mechanism in the likelihood. A design that has to be included in the likelihood in order to obtain legitimate inferences is called non-ignorable. The objective of this work is to propose a general framework for performing Bayesian inference when the sampling mechanism on the social network is non-ignorable.

On Ignorability

There are situations where computing the likelihood depends not only on who is included in the sample, but also on how the sample was collected:

- The probability distribution of the sampling design may depend on features corresponding the portion of the graph that was not sampled.
- Two realizations of the sample mechanism can lead to the same observed subgraph, but the probabilities associated to the samples are different.

The concept of coarse data is a generalization to the notion of missing data that encompasses situations such as rounding and censoring, where the data are partially observed. Denote by W and Z the full observed and the coarse data, respectively. Think of the coarse data as $Z = Z(W, V)$, where V denotes a random function. We denote its conditional distribution $h(\cdot \mid W)$. Let w, z and v denote the realizations of W, Z and V, respectively. Given these components, Heijtan and Rubin [5] define:

$$r(z; w, v) = \begin{cases} 1 & \text{if } z = Z(w, v); \\ 0 & \text{if } z \neq Z(w, v). \end{cases}$$

It follows that, the uncertainty regarding the coarsening is described by:

$$\kappa(z \mid w) = \int_{\mathcal{V}} r(z; w, v) h(v \mid w) dv.$$

Therefore, the correct likelihood is given by:

$$L_C(\theta; z) \propto \int_{\mathcal{W}} p(w \mid \theta) \kappa(z \mid w) dw.$$

This likelihood can be approximated by:

$$L_G(\theta; z) \propto \int_{\{w: Z(w)=z\}} p(w \mid \theta) dw.$$

- The main theorem in Heijtan and Rubin [5] gives the conditions under which both likelihoods lead to equivalent inferences.
- We translated the framework of Heijtan and Rubin [5] to the context of social networks and proved that sampling mechanisms such as Respondent-Driven Sampling (RDS) [4] are not ignorable.

General Model & Assumptions

We assume a data generative model of the form:

$$p(Y, C, \mathcal{G}, \alpha, \gamma) = p(\alpha) p(\mathcal{G} \mid \alpha) p(C \mid \mathcal{G}) p(\gamma) p(Y \mid \mathcal{G}, \gamma).$$

This means that we see the social network $\mathcal{G}$ as a realization of a random graph model with parameter vector α. The distribution of the coarsening mechanism C is conditional on a given realization of $\mathcal{G}$; it is not necessary to know how the graph was generated to sample C. We assume that $\mathcal{G}$ induces a dependence structure on the response vector Y. To fully specify such distribution, additional parameters (denoted by γ) are needed, that explains the term $p(Y \mid \mathcal{G}, \gamma)$.
We included the factor $p(C \mid \mathcal{G})$ to deal with ignorability issues. Still, we need to model the partially observed data: that is the purpose of adding the factor $p(\mathcal{G} \mid \alpha) p(\alpha)$. We assume that the dependence structure modeled by $p(Y \mid \mathcal{G}, \gamma)$ is a Markov Random Field (MRF).

Inference

- Let $\mathcal{G}_{INC}$ and $\mathcal{G}_{EXC}$ denote, respectively, the observed and partially observed portions of $\mathcal{G}$;
- Y_{INC} and Y_{EXC} are defined in an analogous manner for the response vector.
- $\mathcal{G}_{INC}$ and Y_{INC} have to be augmented to compute $p(Y \mid \mathcal{G}, \gamma)$.
- There is uncertainty regarding how many nodes to augment.
- The observed data contains little information about how to augment the graph.

Let $\theta_k = (Y_{EXC}, \mathcal{G}_{EXC}, \gamma, \alpha)$. Apply Bayesian Model Averaging (BMA, for short. See [7] and Section 7.4 of [8]).

$$p(Q \mid Y_{INC}, \mathcal{G}_{INC}, C) = \sum_k p(\mathcal{M}_k) \int_{\Theta_k} p_k(Q \mid \theta_k) p(\theta_k \mid Y_{INC}, \mathcal{G}_{INC}, C) d\theta_k$$

The way $p(\mathcal{M}_w)$ is constructed goes as follows:

- We obtain D samples from $p(\mathcal{G}, \alpha) = p(\alpha) p(\mathcal{G} \mid \alpha)$.
- For each sample a realization of $(C \mid \mathcal{G})$ is obtained.
- Therefore, we have MC versions of $(\mathcal{G}, \mathcal{G}_{INC})$.
- From them the number of nodes and edges to be augmented is computed.

To compute $p_k(Q \mid \theta_k)$:

- For each iteration of the MCMC, construct $Y_{MC} = (Y_{INC}, Y_{EXC})$.
- Let Q be the sample mean of Y_{MC}.

To compute $p(\theta_k \mid Y_{INC}, \mathcal{G}_{INC}, C)$:

- We implemented Metropolis-Hastings with a mixture of kernels.

Model Specification

For $\mathcal{G}$ an Erdös-Rényi model was assumed [3], with a single probability of inclusion $\alpha \in (0, 1)$. A Beta(ω_1, ω_2) was used as prior for α. Our specification for $p(C \mid \mathcal{G})$ is an RDS [4] with m coupons per wave and sample size k. For this paper we will assume that y_i is binary, and

$$\Pr\{y_i = 1 \mid Y_{-i}, \mathcal{G}, \gamma\} = \Phi\left(\psi + \sum_{\{k | A(i,k)=1\}} \zeta y_k\right), \qquad (1)$$

where A is the adjacency matrix for $\mathcal{G}$ and $\gamma = (\psi, \zeta)$. In other words $p(Y \mid \mathcal{G}, \gamma)$ is specified as a MRF based on a probit model. We used a scaled Beta as prior for ζ, *i.e.*

$$p(\zeta \mid \eta_1, \eta_2, \delta) = \frac{1}{B(\eta_1, \eta_2)} \frac{\zeta^{\eta_1 - 1} (\delta - \zeta)^{\eta_2 - 1}}{\delta^{\eta_1 + \eta_2 - 1}} \times \mathbb{I}_{(0,\delta)}(\zeta), \qquad (2)$$

and a similar distribution as prior for ψ:

$$p(\psi \mid \nu_1, \nu_2, \xi) = \frac{1}{B(\nu_1, \nu_2)} \frac{(\psi + \xi)^{\nu_1 - 1} (-\psi)^{\nu_2 - 1}}{\xi^{\nu_1 + \nu_2 - 1}} \times \mathbb{I}_{(-\xi,0)}(\psi). \qquad (3)$$

MCMC Sampling

The proposals we used to update each parameter $(Y_{EXC}, \mathcal{G}_{EXC}, \psi, \zeta)$ are the following:

- For ζ we use a mixture kernel, where one component is a random walk reflecting at 0 and δ, the second component is the prior (Equation 2) and the third is a uniform distribution over $(0, \delta)$.
- For ψ we implement a mixture kernel, where one component is a random walk reflecting at $-\xi$ and 0, the second component is the prior (Equation 3) and the third is a uniform distribution over $(-\xi, 0)$.
- For the extra-sample edges in $\mathcal{G}_{EXC}$ (connect a sampled node to a non-sampled one) we take each augmented node k and count its number of neighbors h_k according to the current version of $\mathcal{G}$, then we delete the corresponding edges and then obtain a random sample of size h_k over the nodes for which $C = 1$. New edges are added which connect those h_k nodes to k.
- For the intra-sample nodes in $\mathcal{G}_{EXC}$ (connect two sampled nodes) we first delete all the current intra-sample edges and then add a new set of edges at random while respecting the restriction that no edge will exist between nodes included in the same wave and preserving the current density of the graph.
- For Y_{EXC} we use a mixture kernel. In one of them we impute using independent Bernoullis with mean equal to the average of Y_{INC}. For the other we sample Y_{EXC} from the conditional distribution $p(Y_{EXC} \mid Y_{INC}, \mathcal{G}, \zeta, \psi)$.

To give an idea of the challenges associated with the computation of the updates, we outline the procedure we used to compute the Metropolis ratio for ψ. Let $q(\cdot \mid \cdot)$ denote the proposal distribution; it follows that the Metropolis ratio is of the form:

$$\frac{p(\psi^{(t+1)}) p(Y \mid \mathcal{G}, \zeta, \psi^{(t+1)})}{p(\psi^{(t)}) p(Y \mid \mathcal{G}, \zeta, \psi^{(t)})} \times \frac{q(\psi^{(t)} \mid \psi^{(t+1)})}{q(\psi^{(t+1)} \mid \psi^{(t)})}.$$

The factor involving the proposal distribution and the prior for ψ is easy to handle. Let us focus on the factor regarding the MRF.

$$\frac{p(Y \mid \mathcal{G}, \zeta, \psi^{(t+1)})}{p(Y \mid \mathcal{G}, \zeta, \psi^{(t)})} = \frac{p(Y \mid \mathcal{G}, \zeta, \psi^{(t+1)}) / p(0 \mid \mathcal{G}, \zeta, \psi^{(t+1)})}{p(Y \mid \mathcal{G}, \zeta, \psi^{(t)}) / p(0 \mid \mathcal{G}, \zeta, \psi^{(t)})} \times \Lambda.$$

The quotients can be computed by using an argument similar to Brook's Lemma [1]:

$$\frac{p(Y \mid \mathcal{G}, \zeta, \psi)}{p(0 \mid \mathcal{G}, \zeta, \psi)} = \frac{p(y_1, y_2, \ldots y_{N_{MC}} \mid \mathcal{G}, \zeta, \psi)}{p(0, y_2, \ldots y_{N_{MC}} \mid \mathcal{G}, \zeta, \psi)} \times \cdots \frac{p(0, 0, \ldots 0, y_{N_{MC}} \mid \mathcal{G}, \zeta, \psi)}{p(0, 0, \ldots 0 \mid \mathcal{G}, \zeta, \psi)}$$

where N_{MC} is the length of Y_{MC}. Each of the terms on the right side is computed applying the identity:

$$\frac{p(y_1, y_2, \ldots y_k, \ldots y_{N_{MC}} \mid \mathcal{G}, \zeta, \psi)}{p(y_1, y_2, \ldots y_k^\star, \ldots y_{N_{MC}} \mid \mathcal{G}, \zeta, \psi)} = \frac{p(y_k \mid Y_{-k}, \mathcal{G}, \zeta, \psi)}{p(y_k^\star \mid Y_{-k}, \mathcal{G}, \zeta, \psi)}.$$

Under certain assumptions regarding $\mathcal{G}$, Λ is a function of quantities previously computed and $p(y_1 = 0 \mid \zeta, \psi)$, which can be computed analytically by using the methodology proposed by Kaiser and Cressie [6].

Data

We applied our methodology to the data derived from the study discussed in [2]. This was a large RDS study implemented in a single location, namely the community of Campinas in the state of Sao Paulo, Brazil. Since RDS was used, non-ignorability is an issue. The aim of the study was to infer the prevalence of HIV among gay men in Campinas, Brazil. The study comprised 658 men who have sex with men. The inclusion criteria used for this study, were:

- born male;
- had anal or oral sex with another man or transvestite in the past six months;
- 14 years of age or older;
- reside in the Metropolitan area of Campinas.

RDS was implemented using 16 seeds and 3 coupons per subject ($m = 3$).

Results

Simulation Study

We conducted a simulation study in order to gain better understanding on the performance of our method:

- For the graph topology we used a Small World (SW) model on a circle. The degree on the initial state (the lattice, before re-wiring) was set as 8. Four probabilities for re-wiring were considered: 0.15, 0.35, 0.75, and 0.95.
- To understand the impact of the strength of the dependence among the responses, we varied the parameters of the MRF to represent *low* dependence ($\psi = -0.82$, $\zeta = 0.01$). and *high* dependence ($\psi = -1.1$, $\zeta = 0.15$).
- The size of the underlying network was set as 100, the sample size was fixed in 35. RDS was run using a single seed and 3 coupons in all cases. The specification of the MRF that we used in combination with the random graph model imply $Q = 0.2$ in all scenarios.
- For each scenario we simulated 100 datasets. The average bias for each regime is shown in the table.

Estimator	Dependence	0.15	0.35	0.75	0.95
Bayes	high	0.087	0.045	0.021	0.017
VH	high	0.024	0.026	0.024	0.025
Bayes	low	0.094	0.051	0.023	0.019
VH	low	0.021	0.022	0.021	0.021

Real Data

- We applied our method to the data discussed in de Mello, Pinho, *et al* [2]. The table summarizes the results they obtained:

	Naive		Volz-Heckathorn	
$\hat{Q}$	0.0789	$(0.0577, 0.1001)$	0.0711	$(0.0466, 0.0955)$

- Since no prior information for the social network $\mathcal{G}$ is available, a sensitivity analysis was conducted. For the Erdös-Rényi model [3], the density and the size of the graph were allowed to vary.

N	Density	Mean	SD	0.025	0.05	0.95	0.97
1316	0.1	0.113	0.007	0.096	0.101	0.123	0.125
1316	0.05	0.114	0.008	0.097	0.100	0.128	0.131
1316	0.01	0.109	0.017	0.076	0.077	0.136	0.140
1316	0.005	0.119	0.008	0.104	0.106	0.134	0.136
1316	0.001	0.116	0.009	0.099	0.101	0.133	0.136
1316	$\frac{1}{N}$	0.114	0.006	0.101	0.103	0.125	0.129
2632	0.1	0.133	0.007	0.119	0.122	0.144	0.147
2632	0.05	0.135	0.008	0.118	0.122	0.150	0.151
2632	0.01	0.138	0.014	0.108	0.110	0.158	0.161
2632	0.005	0.146	0.011	0.122	0.123	0.166	0.169
2632	0.001	0.155	0.009	0.138	0.140	0.169	0.173
2632	$\frac{1}{N}$	0.129	0.015	0.096	0.101	0.152	0.156

- Summaries of the posterior for Q. These include: mean, standard deviation, and quantiles.
- Results differ from not model-based approaches; we claim that this is because we are correcting for biases due to the non-ignorability of the design.

Discussion

- Our methodology is the first one that includes a probability model that deals with all relevant sources of uncertainty.
- We are able to deal with non-ignorable designs.
- The approach proposed is highly modular: We allow for different distributional specifications for each of the model's components.

References & Acknowledgments

[1] D. Brook, "On the Distinction Between the Conditional Probability and the Joint Probability Approaches in the Specification of Nearest-Neighbour Systems," *Biometrika*, vol. 51, pp. 481–483, 1964.
[2] M. de Mello, A.A. Pinho, M. Chinaglia, W. Tun, A. Barbosa Junior, and M.C.F.J. Ilario, "Assessment of risk factors for HIV infection among men who have sex with men in the metropolitan area of Campinas city, Brazil, using respondent-driven sampling," *Washington DC: Population Council*, 2008.
[3] P. Erdos, and A. Renyi, "The Evolution of Random Graphs," *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, vol. 5, pp. 17–61, 1960.
[4] D.D. Heckathorn, "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations," *Social Problems*, vol. 44, pp. 174–199, 1997.
[5] D.F. Heitjan, and D. Rubin, "Ignorability and Coarse Data," *Annals of Statistics*, vol. 19, pp. 2244–2253, 1991.
[6] M.S. Kaiser, and N. Cressie, "The Construction of Multivariate Distributions from Markov Random Fields," *Journal of Multivariate Analysis*, vol. 73, pp. 199–220, 2000.
[7] A. Raftery, D. Madigan, and C. Volinsky, "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with discussion)," *Oxford University Press, Bayesian Statistics, Ed. by Berger, J.O. and Bernardo, J.M. and Dawid, A.P. and Lindley, D.V. and Smith, A.F.M.*, p. 323–349, 1996.
[8] C. Robert, "The Bayesian Choice, Second Edition," *Springer-Verlag*, 2001.