FUNDAÇÃO GETULIO VARGAS
SCHOOL OF APPLIED MATHEMATICS

LUCAS MACHADO MOSCHEN

PREVALENCE ESTIMATION AND BINARY REGRESSION METHODS FOR RESPONDENT-DRIVEN SAMPLING WITH OUTCOME UNCERTAINTY

Rio de Janeiro
2021

LUCAS MACHADO MOSCHEN

PREVALENCE ESTIMATION AND BINARY REGRESSION METHODS FOR RESPONDENT-DRIVEN SAMPLING WITH OUTCOME UNCERTAINTY

Bachelor dissertation presented to the School of Applied Mathematics (FGV/EMAp) to obtain the Bachelor's degree in Applied Mathematics.

Area of Study: Bayesian statistics.

Advisor: Luiz Max Carvalho

Rio de Janeiro

2021

Ficha catalográfica elaborada pela BMHS/FGV

Moschen, Lucas Machado
Prevalence estimation and binary regression methods for respondent-driven sampling with outcome uncertainty/ Lucas Machado Moschen. – 2021.
110f.

Bachelor Dissertation (Undergraduation) – School of Applied Mathematics.

Advisor: Luiz Max Carvalho.
Includes bibliography.

1. Bayesian statistics. 2. Respondent-driven Sampling. 2. Sensitivity and specificity. I. Carvalho, Luiz Max. II. School of Applied Mathematics. III. Prevalence estimation and binary regression methods for respondent-driven sampling with outcome uncertainty

I dedicate this dissertation to my grandfather Adenir, who always encouraged me to study, even when he did not understand it.

Acknowledgements

TODO

"If your experiment needs a statistician, you need a better experiment."
Ernest Rutherford

Abstract

Hard-to-reach populations are difficult to access for researchers or refuse to enrol in public health surveys, making enumeration and sampling challenges. Respondent-driven sampling (RDS) is a chain-referral technique used to recruit individuals from hard-to-reach populations. The survey encourages the participants to recruit their peers, giving incentives to each recruitment and for participation. Since there is no enumeration of the subjects, RDS is a non-probabilistic sampling strategy. Moreover, the graphical structure of RDS suffers from missing data, and several assumptions about the recruitment process are necessary. de After having the sampled individuals, understanding their characteristics is a focus in epidemiology, given that these are usually high-risk populations to some diseases. Therefore, estimating the disease prevalence, the proportion of infected individuals, and the dependence among other observed variables is a critical step for public decision making. Diagnostic tests for disease identification are subject to misclassification, and incorporating their accuracy corrects biases in the prevalence estimation problem. This work proposes the use of regression techniques for prevalence estimation in respondent-driven samples. We use conditionally autoregressive models to represent correlation among the individuals induced by recruitment.

In modern statistics, understanding situations with unknown information and quantifying them plays a significant role. We use Bayesian inference for uncertainty quantification for our models. In the Bayesian paradigm, probability distributions for quantities of interest represent the belief about them. We discuss different prior specification approaches for the parameters and examine uncertainty about the graph structure using a graphical model of RDS. To perform sampling from the parameter distribution, we used the Hamiltonian Monte Carlo sampler. Diagnostics of this method helped to improve our model programming. Verification of the model through simulation and external datasets showed robust results, and we propose model extensions for the limitations of this work. amostras dirigidas por respondentes

Keywords: respondent-driven sampling, regression analysis, Bayesiande inference, prevalence estimation, misclassification, sensitivity, specificity

Resumo

Populações de difícil acesso são difíceis para pesquisadores se aproximarem ou se recusam a se inscrever em pesquisas de saúde pública, tornando a enumeração e amostragem desafios. O Respondent-driven sampling (RDS) é uma técnica de referência em cadeia usada para recrutar indivíduos de populações difíceis de alcançar. A pesquisa incentiva os participantes a recrutarem seus pares, dando incentivos a cada recrutamento e à participação. Como não há enumeração dos sujeitos, o RDS é uma estratégia de amostragem não probabilística. Além disso, a estrutura gráfica do RDS sofre com a falta de dados e várias suposições sobre o processo de recrutamento são necessárias.

Após a obtenção dos indivíduos amostrados, o entendimento de suas características é foco da epidemiologia, visto que se trata de populações geralmente de alto risco para algumas doenças. Portanto, estimar a prevalência da doença, a proporção de indivíduos infectados e a dependência a outras variáveis observadas é uma etapa crítica para a tomada de decisão pública. Os testes de diagnóstico para identificação de doenças estão sujeitos a erros de classificação e incorporar suas acurácias corrige vieses no problema de estimação de prevalência. Este trabalho propõe o uso de técnicas de regressão para estimativa de prevalência em Respondent-driven sampling. Usamos modelos condicionalmente autorregressivos para representar a correlação entre os indivíduos induzida pelo recrutamento.

Na estatística moderna, entender situações com informações desconhecidas e quantificá-las desempenha um papel significativo. Usamos inferência bayesiana para quantificação de incerteza para nossos modelos. No paradigma bayesiano, as distribuições de probabilidade para quantidades de interesse representam a crença sobre elas. Discutimos diferentes abordagens de especificação anterior para os parâmetros e examinamos a incerteza sobre a estrutura do grafo usando um modelo gráfico de RDS. Para realizar a amostragem da distribuição dos parâmetros, usamos o amostrador Hamiltonian Monte Carlo. Os diagnósticos deste método ajudaram a melhorar a programação do nosso modelo. A verificação do modelo por meio de simulação e conjuntos de dados externos mostrou resultados robustos, e propomos extensões do modelo para as limitações deste trabalho.

Palavras-chave: respondent-driven sampling, análise de regressão, inferência bayesiana, estimação de prevalência, classificação errada, sensibilidade, especificidade

List of Figures

List of Tables

List of abbreviations and acronyms

RDS	Respondent-driven sampling
TP	True positive
TN	True negative
FP	False positive
FN	False negative
PWID	People who inject drugs
MSM	Men who have sex with men
FSW	Female sex workers
MCMC	Markov chain Monte Carlo
RDS-SH	Salganik and Heckathorn estimator for prevalence in Respondent-driven sampling
RDS-VH	Volz and Heckathorn estimator for prevalence in Respondent-driven sampling
RDS-SS	Successive sampling estimator for prevalence in Respondent-driven sampling
CAR	Conditionally autoregressive models
SAR	Simultaneous Autoregressive models
GLM	Generalized Linear models
IAR	Intrinsically Autoregressive models
HMC	Hamiltonian Monte Carlo
BFMI	Bayesian fraction of missing information
ESS	Effective sample size
CDC	Centers for Disease Control and Prevention
HIV	Human immunodeficiency virus

List of symbols

$\in$	Belongs to
$\Sigma_{i=1}^{n} x_i$	Sum of the variables $x_1, x_2, \ldots, x_n$
$\Pr(A)$	Probability of an event A
M^T	Transpose of matrix M
$\mathbb{1}$	Indicator function
$\text{tril}(M)$	Lower triangle matrix of M
$\mathbb{R}_{>0}$	Set of positive real numbers
$\det(M)$	Determinant of M
$\sim$	Is distributed as
iid	Independent and identically distributed
Φ	Normal cumulative distribution
exp	Exponential
$\int$	Integral
$\mathbb{E}(X)$	Expected value of random variable X
$\text{Var}(X)$	Variance of random variable X
A^*	Hermitian matrix of A

Contents

1 Introduction

Hidden or hard-to-reach populations have two key features: no sampling frame exists, and the individuals have privacy concerns about participating in surveys. The former occurs because the population's size and boundaries are unknown. Furthermore, the subjects suffer from stigmatization or engage in illegal behaviour (HECKATHORN, 1997), which complicates learning about them. Moreover, if the frequency of the condition of interest is low, there are high logistic costs involved. Some examples are heavy drug users, sex workers, homeless people, and men who have sex with men.

Methods to reach these populations, such as, for example, snowball sampling (GOODMAN, 1961), key important sampling (DEAUX; CALLAGHAN, 1985), and targeted sampling (WATTERS; BIERNACKI, 1989) have been developed. Heckathorn (1997) introduced the Respondent-Driven Sampling (RDS) to fill some gaps from other strategies he depicted in his work. In his proposed approach, the researchers select a handful of individuals from the target population and give them coupons to recruit their peers. The individuals receive a reward for being recruited and for recruiting, which creates a dual incentive system. After Heckathorn (1997), several papers studied this topic more deeply.

After sampling from the target population, a questionnaire or a disease test is conducted. When considering a disease, an essential quantity for epidemiologists is the proportion of infected people, called the prevalence. The estimation of this quantity can be crucial for public decision making. However, the diagnoses are subject to measurement error, and considering their accuracy is a vital step (REITSMA et al., 2005). One common way to do this is jointly measuring *sensitivity* and *specificity*. Sensitivity measures the ability of the test to detect the condition, whereas specificity refers to the test's capacity to verify its absence. These quantities are often negatively correlated, and a higher sensitivity of a screening test reduces the specificity and vice versa.

Furthermore, other variables accessed in the survey can be possible risk factors for the disease and need better understanding within the target population. From correlational studies, one can analyze causal dependencies. The literature of regression analysis in RDS samples is not established yet (AVERY, 2020, p. 15) and conceptual questions are yet to be addressed. We develop a hierarchical model to represent the structure of the process.

Because of our lack of knowledge about the complex interactions in Nature, it is necessary to model the uncertainty of the process under study. The models we develop to represent these interactions are approximations subjected to error. It leads to the famous quote: "all models are wrong, but some are useful." In finite sample experiments, the model parameters should reflect our state-of-art knowledge about them. The Bayesian Statistics

paradigm handles uncertainty quantification defining the unknown quantities as random variables, and new data update the beliefs about these quantities.

This work proposes to study the survey method RDS, a chain-referral method to interview hard-to-reach populations when necessary to estimate the prevalence of some binary condition from this population. The modelling strategy follows a hierarchical model with regressor variables. The modelling also accounts for sensibility and sensitivity given the imperfection of the detection tests. We apply these methods using the Hamiltonian Monte Carlo method in Simulated and real data.

The dissertation is organized as follows: Chapter 2 describes the theoretical foundations. Chapter 3 discusses each block of the model and validades its properties in simulated data. Model extensions are also discussed in the end of the chapter. Chapter 4 describes two datasets and utilize the model on them. At last, Chapter 5 concludes the main aspects of this work.

2 Theoretical background

In this chapter, we shall describe the theoretical background taken under consideration for the developed models and analysis, including the prevalence estimation problem (Section 2.1), Respondent-driven sampling (Section 2.2), Bayesian statistics (Section 2.4), and computational methods (Section 2.5) used in our research.

2.1 Prevalence estimation problem

The study of how health-related conditions are distributed among populations is known as *Epidemiology* (ROTHMAN; GREENLAND; LASH, et al., 2008, p. 32), which aims to derive valid estimates for potential causes from diseases that affect people. It is a fundamental research area in policy formulation, implementation of prevention programs, and development of laws. In order to accomplish these goals, epidemiologists use *measures of disease frequency*, including *incidence* and *prevalence*. The former is related to the proportion of new cases of a disease given a period of time, while the latter is the proportion of individuals exposed at time t and it is the object of study of this section. An interesting point is the following:

> Diseases with high incidence rates may have low prevalence if they are rapidly fatal or quickly cured. Conversely, diseases with very low incidence rates may have substantial prevalence if they are nonfatal but incurable. (ROTHMAN; GREENLAND; LASH, et al., 2008, p. 46).

As a result, prevalence is determined by both incidence and the duration of disease. Noordzij et al. (2010, p. c18) highlights that prevalence reveals the burden of a disease in respect to its effects on society, such as monetary costs, quality of live, and morbidity. They also comment that when measured periodically, the evolution of prevalence can identify potential causes of the infection and prevention and care methods. We remark that when it is impossible to test every individual at the same time, we may assume that all individuals remain exposed to the disease at time of the last tested individual.

Consider a population of interest and a known condition, such as, for instance, a disease or a binary behavior, such as smoking status. A diagnostic test is done in the individuals to measure the presence or the absence of this condition, such as serological tests. Mathematically, we denote $\theta \in (0, 1)$ the prevalence of the condition, which is the parameter of interest. Let I be a index set for the individuals. We also denote Y_i^{true} the

indicator function of the presence of the condition in the ith individual, that is,

$$Y_i^{\text{true}} = \begin{cases} 1, & \text{if individual } i \text{ has the condition.} \\ 0, & \text{otherwise.} \end{cases}$$

Assume for simplicity that all tests are performed at time t. Assume that Y_i indicates the result of the test, then

$$Y_i = \begin{cases} 1, & \text{if test was positive in individual } i. \\ 0, & \text{otherwise.} \end{cases}$$

Since it is not usually feasible to test every individual in the population, it is necessary to randomly select individuals from the population. On that point, other sampling approaches may be better options, such as stratified random sampling, systematic sampling, and two-stage cluster sampling (DANIEL, 2011, p. 125). From that experiment, we get a sample $y = \{y_1, ..., y_n\}$. Based on that outcomes the Maximum Likelihood Estimator is the following expression

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i, \tag{2.1}$$

which is an estimator for the *apparent prevalence*, that is, the probability of a positive outcome.

However, this estimator assumes that the diagnostic test used is perfect, which is often incorrect. It is also not interesting when the samples are not randomly selected (See Section 2.2). From that point, it is crucial to regard the evaluation of the diagnostic procedure by some measurement. Šimundić (2009, p. 2) presents several options with different aspects, such as the *likelihood ratio*, *sensitivity and specificity*, and *the area under the ROC curve*. In this work, we consider the sensitivity and specificity of the test.

A perfect test would discriminate every sick individual from the non-sick ones. Given that there is no such thing, we suppose having a *gold standard test* that is the best available test (VERSI, 1992) to diagnose a particular disease. Its result is a proxy for the real Y_i^{true} and

> in the context of infectious diseases, a gold standard can be a very precise molecular test that detects the presence of the pathogen's genetic material, polymerase chain reaction (PCR) for instance. (BASTOS; CARVALHO; GOMES, 2021, p. 125).

From the gold standard, we can evaluate a second test, typically faster or cheaper. The possible results upon comparing these tests are presented in table 1. The definitions for each initials in the table are the following:

a) true positive (TP): when both tests agree that the individual has the disease;

b) true negative (TN): when both tests agree that the individual does not have the disease;

c) false positive (FP): when the test under evaluation has a positive diagnose, despite the gold standard being negative;

d) false negative (FN): when the test under evaluation has a negative diagnose, despite the gold standard being positive.

Chart 1 – Two-by-two table that compares the result from the gold standard to the test under evaluation.

	$Y = 0$	$Y = 1$
$Y^{\text{true}} = 0$	TN	FP
$Y^{\text{true}} = 1$	FN	TP

Source: Prepared by the author (2021) and based on Bastos, Carvalho, and Gomes (2021, p. 126).

Remark 2.1.1. When a gold standard test is not available, which is called *no gold standard situations* (RUTJES et al., 2007, p. 1), other methods should be considered, such as the construction of reference standard by giving the patients either different or the same tests and combining the results somehow. Rutjes et al. (2007) does a literature review on the topic.

For now, we drop the index i in the random variables Y_i and Y_i^{true}. Let $p = \Pr(Y = 1)$ be the probability of a positive test. We call p the *apparent prevalence* since it is what the researchers observe. Equation (2.1) is an estimator for it. We also have that $\Pr(Y^{\text{true}} = 1) = \theta$. Notice that p depends on the used test, while θ does not. In prevalence estimates, we will only have $\theta = p$ if the test is perfect or the test is the gold standard itself. Define the following:

Definition 2.1.1 (Sensitivity)**.** Probability of a positive test correctly identified. In mathematical terms, conditioned on $Y^{\text{true}} = 1$, the *sensitivity* γ_s is the probability of $Y = 1$:

$$\gamma_s = \Pr(Y = 1 | Y^{\text{true}} = 1). \tag{2.2}$$

Definition 2.1.2 (Specificity)**.** Probability of a negative test correctly identified. In mathematical terms, conditioned on $Y^{\text{true}} = 0$, the *specificity* γ_e is the probability of $Y = 0$:

$$\gamma_e = \Pr(Y = 0 | Y^{\text{true}} = 0). \tag{2.3}$$

Theorem 2.1.1 (Relation between prevalence and apparent prevalence)**.** *These quantities are related by the following equation:*

$$p = \gamma_s \theta + (1 - \gamma_e)(1 - \theta). \tag{2.4}$$

Proof. This is a direct application of the definition of conditional probability and the countable additivity axiom of Probability:

$$\begin{aligned}
p &= \Pr(Y=1) = \Pr(Y=1, Y^{\text{true}}=1) + \Pr(Y=1, Y^{\text{true}}=0) \\
&= \Pr(Y=1|Y^{\text{true}}=1)\Pr(Y^{\text{true}}=1) + \Pr(Y=1|Y^{\text{true}}=0)\Pr(Y^{\text{true}}=0) \\
&= \Pr(Y=1|Y^{\text{true}}=1)\Pr(Y^{\text{true}}=1) \\
&+ (1-\Pr(Y=0|Y^{\text{true}}=0))(1-\Pr(Y^{\text{true}}=1)) \\
&= \gamma_s\theta + (1-\gamma_e)(1-\theta).
\end{aligned}$$

□

The intuition behind this equation is pretty simple: the proportion of positive test counts the correct identified exposed individuals and the incorrect identified not exposed. Equation (2.4) also reveals that if $\gamma_s = \gamma_e = 1$, we have the trivial case $p = \theta$. Moreover, if $\gamma_s = \gamma_e = 0.5$, we have that $p = 0.5$ and there is no information about θ.

A frequentist approach assumes that θ is fixed and unknown. Its inference is based on the point estimate for the apparent prevalence $\hat{p}$ given in Equation (2.1), along with a Confidence Interval, such as the Wald Confidence interval built with a normal approximation. In order to provide a point estimate for $\hat{\theta}$, Rogan and Gladen (1978, p. 73) propose

$$\hat{\theta}^{RG} = \frac{\hat{p} - (1-\gamma_e)}{\gamma_s + \gamma_e - 1}. \tag{2.5}$$

Suppose a disease with prevalence $\theta = 0.01$. In this case, we would have that $p \approx 1 - \gamma_e$ by equation (2.4). Given the randomness, it is possible to have $\hat{p} < 1 - \gamma_e$, which would define a useless estimative for θ. Besides that, Confidence Intervals for that expression does not include uncertainty about γ_e and γ_s. On the other side, a Bayesian approach let θ be a random variable, allowing the researcher to incorporate their uncertainty on the prior distribution, which is explained in Section 2.4. It also allows to include uncertainty in sensitivity and specificity of the test. According to Branscum, Gardner, and Johnson (2005):

> Diagnostic-test evaluation is particularly suited to the Bayesian framework because prior scientific information about the sensitivities and specificities of the tests and prior information about the prevalences of the sampled populations can be incorporated. (BRANSCUM; GARDNER; JOHNSON, 2005, p. 1).

Therefore, this work focuses on the Bayesian paradigm.

2.1.1 Correlation between sensitivity and specificity

A general method for a diagnostic or screening test is to construct a continuous scale measuring some related quantity to the disease and to define a cut-off number, such that values higher than the threshold indicate the presence of the illness. Suppose the

cut-off is high, almost the maximum value of the scale. Therefore, the majority of the population will be tested negative. There will be a lot of false-negative individuals but a few false-positive ones, which implies that sensitivity is low and specificity is high. If the threshold is smaller, the opposite effect happens. Nonetheless, sensitivity and specificity are negatively correlated. Parikh et al. (2008, p. 46) gives a more practical example about an intraocular pressure test and the impact of different threshold specifications. This correlation can be noticed in meta-analysis studies as presented by Guo, Riebler, and Rue (2017, p. 1).

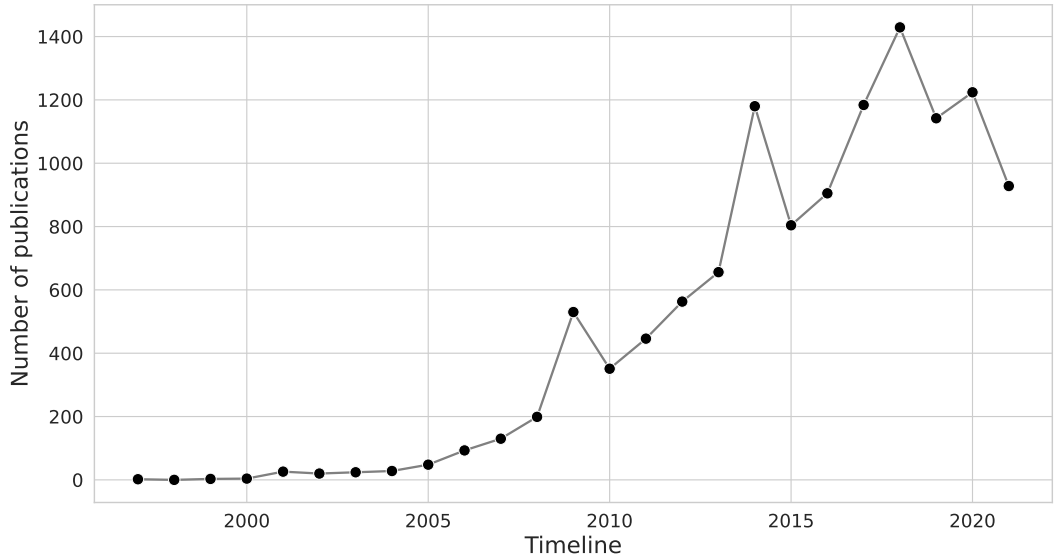
2.2 Respondent-driven sampling

RDS is a procedure developed by Heckathorn (1997) to survey *hidden* or *hard-to-reach populations*, whose main characteristic is the absence of a sampling frame, i.e., it is not possible to enumerate its individuals since size and boundaries are unknown. The second characteristic of these populations is the confidentiality concerns, given that membership is stigmatized or illegal. With that aspect, traditional sampling methods which produce probability samples are infeasible. To overcome this, Snowball Sampling (GOODMAN, 1961) is the most common method, and it relies on the respondents to nominate more subjects within the population as a snowball. Examples of studied groups include people who inject drugs (PWID), men who have sex with men (MSM), and female sex workers (FSW) (GILE; BEAUDRY, et al., 2018, p. 66).

Heckathorn (1997)'s proposal was to specialize this method without the need of nominating peers. In this approach, the researchers select some individuals, called *seeds* from the target population, and give them a fixed amount of *recruitment coupons* to recruit their acquaintances. Each recipient of the coupons reclaims it in the study site, is interviewed, and receives more coupons to continue the recruitment. This process occurs until it reaches some stopping criteria, such as the sample size achieving some desired number. The sampling is without replacement, so the participants cannot be recruited more than once. Moreover, the respondents inform how many subjects from the population they know. Other less usual methods include Key Important Sampling (DEAUX; CALLAGHAN, 1985) and Targeted Sampling (WATTERS; BIERNACKI, 1989), both are convenience sampling methods.

According to Gile, Beaudry, et al. (2018, p. 66), there are two main advantages of RDS over other snowball samplings. First, the fixed number of recruitment coupons enforces the network gets deeper and distant from the seeds, which reduces the dependence of the final sample from the initial chosen by researchers. Second, since the recruited subjects do not have to name their peers, confidentiality is maintained until the recruitment is completed. Other problems cited by Heckathorn (1997, p. 175) include biases towards individuals who are more cooperative, biases by masking when the participants do not

Figure 1 – Publications by year with the term "Respondent driven sampling" from 1997 to 2021.



Source: htpps://app.dimensions.ai. Exported on October 31, 2021.

name friends for the next wave to protect them, and individuals with more links may be oversampled. RDS offers a solution with a *dual incentive system*, explained in Subsection 2.2.1.

Since the creation of the method by Heckathorn, several papers have been published, as Figure 1 presents. The figure was produced searching publications with the term "Respondent-driven sampling." These works generally aim to give basis to public health policies. Good examples in Brazil are Damacena et al. (2019), Mota (2012), and Bastos, Bastos, et al. (2018). Damacena et al. (2019) apply the RDS method to carry out a biological and behavioral surveillance study in FSW populations from twelve cities in Brazil. Mota (2012) proposes the RDS method in MSM populations from ten cities in Brazil. Bastos, Bastos, et al. (2018) study several sexually transmitted infections among transgender women from twelve Brazilian cities.

2.2.1 Details about the sampling procedure

The RDS method was expanded by Heckathorn (2002). It detailed two aspects: introducing a way to correct *homophily* biases that is the tendency for individuals to connect to others similar to them, and *personal network size* or *degree* that is the number of connections of an individual within the target population. It also presented a bootstrapping procedure to quantify uncertainty about inferences. Salganik and Heckathorn (2004) slightly modified the RDS procedure and introduced proof that under some regularity conditions, RDS estimators were asymptotically unbiased. World Health Organization (2013) is a reference to know how to execute an RDS survey. According to it:

> Seeds are non-randomly selected members of the survey population who initiate the RDS recruitment process. From each seed, a recruitment chain is expected to grow. Seeds play an extremely important role in conducting an RDS survey. (WORLD HEALTH ORGANIZATION, 2013, p. 70).

No rule was established on the number of seeds to start the sampling. It typically varies from 2 to 32, with the mean being 10 (WORLD HEALTH ORGANIZATION, 2013, p. 70). The number cannot be small since unsuccessful recruitments are common. A diverse choice among the target population may accelerate the convergence to equilibrium. It also allows the access to isolate and subpopulations. After this selection, three coupons are distributed to each participant. The coupons must have information about survey site location, an unique identification code, telephone number, and opening hours. Gile, Beaudry, et al. (2018, p. 67) highlights that "this number is chosen to strike a balance between the inferential desire [...] and the practical necessity of guarding against early termination of the sample trees."

Subjects receive a reward for being interviewed and recruiting their peers within the target population, which establishes a dual incentive system. The *primary incentive* is the *individual-sanction-based control*, so there is a reward for participating in the survey. The second one is the *group-mediated social control* that influences the participants to induce others to comply to get the remuneration for the recruitment. When social approval is relevant for the members, recruitment can be more efficient and cheaper. It happens because material incentives are converted into peer-based symbolic since there is social influence involved. In conclusion, consenting to be recruited provide material and symbolic motivation to both recruiter and recruited.

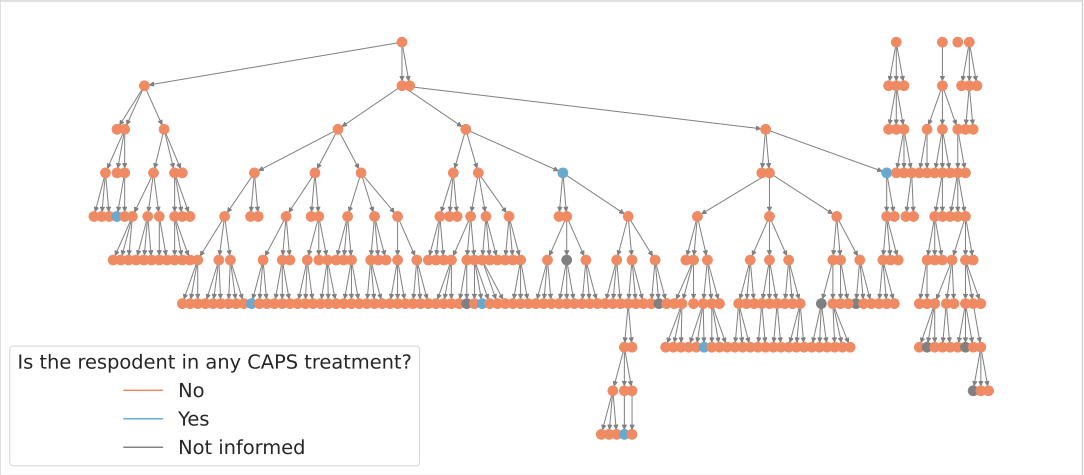
For an illustrative example, Figure 2 presents a recruitment structure based on a respondent-driven sample among 303 heavy drug users from Curitiba collected between July 28, 2008, and October 18, 2009 (SALGANIK; FAZITO, et al., 2011, Web Appendix). Five seeds were chosen within the population, the fifth being a month after the other four since the fourth seed was unsuccessful. Each participant received three coupons and the mean number of recruited individuals per recruiter was around 0.98.

2.2.2 Assumptions and statistical properties

RDS is a successful recruitment method for accessing hard-to-reach populations since the respondents recruit most of the participants. On the other hand, this characteristic also makes it hard to derive statistical properties without making strong assumptions of the recruitment process. Some hypotheses are related to specific models, which are presented in Section 2.2.3:

a) sampling is not uniformly random among the individuals since some have more

Figure 2 – RDS structure among heavy drug users in Curitiba.



Source: Data extracted from (SALGANIK; FAZITO, et al., 2011) and figure prepared by the author (2021). The respondents were asked whether they are in any "Centro de Atenção Psicossocial (CAPS)" (Psychosocial Care Center) treatment program for drug use.

connections than others, which gives them a higher probability of being recruited. Those with more contacts should reduce the weighting in the inferences, but this also relies on another assumption: self-reported degree should be accurately measured (GILE; HANDCOCK, 2010, p. 297);

b) recruitment is without replacement, given that respondents are not allowed to participate more than once. It compromises inferences since the probability of inclusion in the survey also depends on the number of individuals participating until the recruitment time (GILE; HANDCOCK, 2010, p. 299). To derive an RDS estimator, Volz and Heckathorn (2008, p. 81) require a small sampling fraction to compensate for breaking this assumption;

c) *homophily* is the tendency of individuals to connect within the same group. For instance, men tend to recruit more men than women. If the process has zero homophily, it indicates that individuals do not regard the group to recruit. On the other hand, if homophily is one, all the connections are intragroup (HECKATHORN, 2002, p. 20). Heckathorn (2002, p. 21) proved that under certain conditions (see Subsection 2.2.3), the respondent-driven sample is unbiased with respect to homophily if it is equal for each group;

d) the connections generated by the RDS process item b) violate the independence between the samples through *clustering*, i.e., people are more likely to connect to those similar (AVERY, 2020, p. 14);

e) RDS produces a branching structure that makes it impossible to observe links between two people who don't recruit each other (GILE; HANDCOCK, 2015, p. 17). It constitutes a missing data problem, according to Crawford (2016, p.

190);

f) in apparent contradiction to item b), to the distribution achieve its convergence and remove the biases induced by the initial sample, enough waves of recruitments are necessary (HECKATHORN, 1997, p. 186);

g) Goel and Salganik (2009, p. 2225) defines *bottleneck* as the probability of cross-group recruitment. It happens when the recruitment chain remains inside an identified subgroup of individuals. In that situation, "studies should be conducted separately within each tier." (GILE; BEAUDRY, et al., 2018, p. 75). As an expository example, Toledo et al. (2011, p. S139) observes strong geographical heterogeneity among a population of heavy drug users in Rio de Janeiro.

To check these assumptions and to analyze aspects of RDS, Gile, Johnston, and Salganik (2015) developed a tool of diagnostics, including plots, such as the bottleneck plot, and suggested questions for the questionnaire, such as asking the respondents the speed of coupon distribution.

2.2.3 Models for the RDS Process

Since its inception, several authors have tried to better understand and model the RDS process because of its non-probabilistic nature. Each modelling approach aims to approximate even more the network structure to yield more reliable inferences. In this section, we present some of them. Let $G = (V, E)$ be an undirected graph representing the hidden population, such that $|V| = N$, and $A \in \{0, 1\}^{N \times N}$ its adjacency matrix, where $A_{ij} = 1$ if there is a connection between individuals i and j, and $A_{ij} = 0$ otherwise. We denote $|V|$ to mean the number of nodes, and $|E|$ the number of edges in the graph G. The choice of an undirected model for the hidden population is very common, but not mandatory. The degree of a person is, therefore, $d_i = \sum_{j=1}^{N} A_{ij}$.

Besides the following models, there are two additional and relevant works to cite. Goel and Salganik (2009) described RDS as a Markov chain Monte Carlo (MCMC) to analyse the structure created by the recruitment links. They deeply discussed the problems that bottlenecks can cause. McLaughlin (2021) develops a Bayesian model for the recruitment process considering preferential selection based on covariates.

2.2.3.1 First-order Markov process

The first model was proposed by Heckathorn (1997). He argues RDS recruitment has the characteristic that "any subject's recruits are a function of his or her type, such as his or her ethnicity; and not of previous events, such as who recruited the recruiter" (HECKATHORN, 1997, p. 182). Consequently, recruitment is modelled as a first-order

Markov chain in the space of states generated by the categorical variables, such as ethnicity or gender. The evidence for the above statement is based on chi-square analysis. By these hypotheses, the paper uses the following well-known results from Markov chain theory:

Theorem 2.2.1 (Convergence to equilibrium)**.** *Let $\{Z_n\}_{n\in\mathbb{N}}$ be the recruitment process. Given that the state space is finite, if the Markov chain is irreducible and aperiodic, then it converges to the stationary distribution and is independent of the initial sample (HECKATHORN, 1997, p. 183).*

Proof. A proof is outlined in (LEVIN; PERES, 2017, p. 52-53). □

Theorem 2.2.2 (Geometric rate of convergence)**.** *The convergence of the Markov chain generated by RDS recruitment converges to the stationary distribution at a geometric rate (HECKATHORN, 1997, p. 186).*

Proof. The same proof given in (LEVIN; PERES, 2017, p. 52-53), demonstrates geometric convergence. □

Moreover, he establishes conditions for unbiased samples:

Theorem 2.2.3 (Unbiased samples)**.** *A respondent-driven sample produces an unbiased sample if all groups have same homophily, that is, the probability of selecting a member within the same group for any group is the same (HECKATHORN, 1997, p. 192).*

Proof. Heckathorn (1997, p. 191 - 192) presents a proof for this fact. □

Heckathorn (2002, p.22) extended this model under the hypothesis that relationships between the individuals are reciprocal. The Random Walk model simplifies this concept by proposing that each recruitment in the social network G occurs between adjacent nodes with uniform probability and that the process begins with a unique seed. With the assumption that the graph has only one connected component and that the researchers chose the seed with probability proportional to its degree, Salganik and Heckathorn (2004, p. 209-218) derives sampling probabilities. A proof of asymptotic convergence to the stationary distribution

$$\pi_j^* = \frac{d_j}{\sum_{i=1}^{N} d_i} \tag{2.6}$$

is provided (SALGANIK; HECKATHORN, 2004, p. 234-235). The authors ponder limitations regarding the validity of these assumptions in real applications and assert that "Empirically checking the reasonableness of the assumptions and further research related to the robustness of the estimation procedure are both problems worthy of further study." (SALGANIK; HECKATHORN, 2004, p. 230).

2.2.3.2 Successive sampling (SS)

The problem with the Random Walk approach with replacement is the assumption of a small sample fraction. It induces biases in prevalence estimates since population size can be small, implying that convergence will not occur or the sample fraction will be high. Both cases break the assumption. To adjust for finite population effects, Gile (2011) suggests a successive sampling approach. Along with the sampling, the recruitment probability is proportional to the size of the remaining not recruited population.

The procedure starts sampling an individual i with probability proportional to degree d_i. After, it selects another individual with probability proportional without replacement, given by expression (2.7) (GILE, 2011, p. 136).

$$\Pr(G_j = g_j \mid G_1 = g_1, \ldots, G_{j-1} = g_{j-1}) = \begin{cases} \dfrac{d_{g_j}}{2|E| - \sum_{i=1}^{j-1} d_{g_i}}, & g_j \notin \{g_1, \ldots, g_{j-1}\} \\ 0, & \text{otherwise,} \end{cases} \tag{2.7}$$

such that $G_i = g_i$ is the event of the selection of individual g_i in the step i. To estimate probabilities, this model assumes that the degree distribution and the population size N are known (GILE, 2011, Table 2, p. 144), the latter not being necessary to the Random Walk with replacement model.

2.2.3.3 Graphical Structure model

Crawford (2016) presented a model to probabilistically reconstruct the subgraph whose nodes are the respondents and edges are their connections. He considered the information brought by the waiting times between recruitments and the remaining coupons with the recruiters to define a probability distribution on the space of subgraphs.

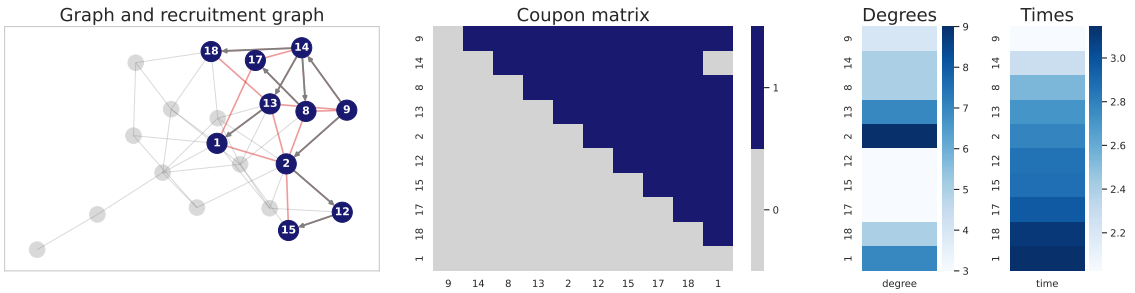
Definition 2.2.1 (Recruitment graph)**.** The *recruitment graph* $G_R = (V_R, E_R)$ represents the recruited individuals and the recruitment edges. Therefore $i \in V_R$ if individual $i \in V$ was recruited, and $(i, j) \in E_R$ if individual $\{i, j\} \in E$ and individual i recruited individual j. Notice that G_R is a *forest*, that is, a collection of trees. (CRAWFORD, 2016, p. 193).

Denote $n = |V_R|$. Given that each individual can be sampled only once, it is not possible to observe the *recruitment-induced subgraph*, that is

Definition 2.2.2 (Recruitment-induced subgraph)**.** It is the induced subgraph $G_S = (V_S, E_S)$ generated by V_R, that is, $V_S = V_R$ and $\{i, j\} \in E_S$ if $i, j \in V_R$ and $\{i, j\} \in E$. (CRAWFORD, 2016, p. 192).

Denote $\boldsymbol{t} = (t_1, \ldots, t_n)$ the vector of recruitment times of the individuals such that $t_1 < \cdots < t_n$, and $\boldsymbol{d} = (d_1, \ldots, d_n)$ the degrees of the individuals in the same order. Then we define

Figure 3 – Example of simulated data from Crawford model.



Source: Prepared by the author (2021). In the network graphic, the blue nodes were sampled, and the red edges are the non-observed links between sampled.

Definition 2.2.3 (Coupon matrix)**.** The *coupon matrix* $C \in \{0,1\}^{n\times n}$ is defined by $C_{ij} = 1$ if the ith subject has at least one coupon just before the jth recruitment event. The row order is the same of $\boldsymbol{t}$. (CRAWFORD, 2016, p. 193).

From the RDS process, the observed data is $\boldsymbol{Z} = (G_R, \boldsymbol{d}, \boldsymbol{t}, C)$. Figure 3 illustrates how the data is presented in a simple situation. The algorithm to generate simulated data was self-made based on Crawford (2016)'s assumptions.

Definition 2.2.4 (Compatibility)**.** Let $\hat{G}_S = (\hat{V}_S, \hat{E}_S)$ be an estimate for G_S. The subgraph $\hat{G}_S$ is *compatible* with data $\boldsymbol{Z}$ if

a) $v \in V_R$ if and only if $v \in \hat{V}_S$;

b) $\forall (i,j) \in E_R, \{i,j\} \in \hat{E}_S$;

c) $\forall v \in V_R, \sum_{u \in V_R/\{v\}} \mathbb{1}\{\{u,v\} \in \hat{E}_S\} \leq d_v$. (CRAWFORD, 2016, p. 197).

We denote $\mathcal{C}(\boldsymbol{Z})$ the set of all compatible subgraphs for $\boldsymbol{Z}$.

After the recruitment time t_i, individual i is a recruiter until their coupons or non-recruited neighbors are exhausted. A node is *susceptible* if it has a link to a recruiter. An edge is susceptible if it connects a recruited and a susceptible node. After j being recruited, every $\{i,j\} \in E$ with $i \in V_R$ is no longer a susceptible edge. Moreover, Crawford (2016, p. 194) assumes that each recruitment time has exponential distribution with parameter λ and it is independent of the recruiter characteristics, neighbors, and all other waiting times. This assumption may fail when homophily is strong. Some interesting propositions follows from this construction (CRAWFORD, 2016, p. 195), but here we focus on G_S.

Let $\tilde{A} \in \{0,1\}^{n\times n}$ be the adjacency matrix of a compatible estimated subgraph, that is, $[\tilde{A}]_{ij} = 1$ if and only if $\{i,j\} \in \hat{G}_S$. Then

$$[AC]_{ij} = \sum_k [A]_{ik}[C]_{kj} = \sum_k \mathbb{1}(\{i,k\} \in \hat{G}_S \text{ and } k \text{ can recruit in } t_j),$$

that is the number of recruiters connected to i just before the j^{th} recruitment, when $j \leq i$. Let u_i be the number of edges linking the sampled node i with others not sampled. Then,

$$[C^T u]_i = \sum_k [C]_{ki} u_k = \sum_k \mathbb{1}(k \text{ can recruit at } t_i) \cdot \#\text{susceptible edges of } k$$

Proposition 2.2.1. *The likelihood of the recruitment times* $w = (0, t_2 - t_1, ..., t_n - t_{n-1})$ *is*

$$L(w|G_S, \lambda) = \left(\prod_{k \text{ isn't seed}} \lambda s_k \right) \exp(-\lambda \boldsymbol{s}^T w), \tag{2.8}$$

where

$$\boldsymbol{s} = \text{tril}(\tilde{A}C)^T 1 + C^T u$$

indicates the number of susceptible edges just before each recruitment. (CRAWFORD, 2016, p. 197).

Proof. A proof of this proposition is given in the online Appendix of (CRAWFORD, 2016). □

Setting $T(\tilde{A}) = -\lambda \boldsymbol{s}$ and $B(\tilde{A}) = \sum_{k \text{ isn't seed}} \log(\lambda \boldsymbol{s}_k)$, the likelihood from above can be normalized to obtain the probability

$$P(\tilde{A}|w) \propto \exp\left[T(\tilde{A})^T w + B(\tilde{A})\right]$$

which can be interpreted as an Exponential Random Graph Model (ERGM) (CRAWFORD, 2016, p. 198). Finally, from a Bayesian perspective (see Section 2.4), one can define prior distributions for G_S and λ to obtain,

$$p(G_S, \lambda|G_R, C, d, t) \propto L(w|G_S, \lambda)\pi(G_S, \lambda), \tag{2.9}$$

where $\pi(G_S, \lambda)$ is a prior density. An application of this model was the estimation of the hidden population size with the additional assumption that the graph G has Erdős-Rényi distribution (CRAWFORD; WU; HEIMER, 2018).

2.2.4 Prevalence estimators

In this subsection, we outline five very common RDS proportion estimators presented in the literature based on the modelling from Subsection 2.2.3. They are apparent prevalence estimators and can be used for prevalence estimate through equation (2.5) in a frequentist approach. Then:

a) *naive estimator*: it is the sample proportion

$$\hat{\theta}_{\text{naive}} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

as in equation (2.4);

b) *Salganik-Heckathorn RDS estimator* (RDS-SH): Considering the Random Walk approximation, Salganik and Heckathorn (2004) built this estimation regarding the sampling probabilities. Let $N_T = \sum_{i \neq j} A_{ij} y_i (1 - y_j)$ be the number of connections between individuals with and without the disease, $\bar{d}_1 = \frac{\sum_{i=1}^{N} \sum_{j \neq i} A_{ij} y_i}{\sum_{i=1}^{N} y_i}$ the mean degree of ill individuals, $\bar{d}_0$ the mean degree of not ill individuals with a similar formula, and $N_1 = N \cdot \theta$. Salganik and Heckathorn (2004, p. 218) derives that

$$\theta = \frac{\bar{d}_0 c_{01}}{\bar{d}_0 c_{01} + \bar{d}_1 c_{10}},$$

where

$$c_{01} = \frac{N_T}{(N - N_1)\bar{d}_0} \text{ and } c_{10} = \frac{N_T}{N_1 \bar{d}_1}.$$

Therefore, the prevalence estimator is

$$\hat{\theta}_{\text{SH}} = \frac{\widehat{d}_0 \hat{c}_{01}}{\widehat{d}_0 \hat{c}_{01} + \widehat{d}_1 \hat{c}_{10}}, \tag{2.10}$$

such that $\widehat{d}_0, \widehat{d}_1, \hat{c}_{01}$, and $\hat{c}_{10}$ are estimated for the corresponding quantities.

c) *Volz-Heckathorn RDS estimator* (RDS-VH): With similar assumptions to the previous one, Volz and Heckathorn (2008, p. 85) shows that the inclusion probability of individual i in the sample is $\pi_i \propto d_i$ and the corresponding proportion estimator is

$$\hat{\theta}_{\text{VH}} = \frac{\sum_{i=1}^{n} y_i d_i^{-1}}{\sum_{i=1}^{n} d_i^{-1}}. \tag{2.11}$$

The assumptions for $\hat{\theta}_{\text{VH}}$ were highlighted in Subsubsection 2.2.3.1, and are summarized in (Table 1 GILE; BEAUDRY, et al., 2018, p. 71).

d) *Successive sampling estimator:* (RDS-SS) Under the successive sampling approximation for RDS, Gile (2011, p. 137-138) derives an estimate considering the without replacement assumption. It is of the form

$$\hat{\theta}_{\text{SS}} = \frac{\sum_{i=1}^{n} y_i w_i}{\sum_{i=1}^{n} w_i}, \tag{2.12}$$

where w_i is calculated algorithmically, taking into account the finite population effect. If the sampling fraction is small, this estimator is similar to RDS-VH estimator. Otherwise, when it grows, RDS-VH is biased. The limitation of RDS-SS estimator is that N is assumed to be known, which is rarely the case. Gile (2011, p. 140) did a sensitivity analysis on population size estimate and found out that "if the hidden population size is unknown, but sufficiently large, the estimator is not sensitive to the working population size." (GILE; HANDCOCK, 2015, p. 16)

e) *RDS-B estimator*: (BASTOS; BASTOS, et al., 2018) proposes a pseudo-posterior approach to estimate prevalence. Let

$$Y_i \sim \text{Bernoulli}(\theta_i) \text{ with } \text{logit}(\theta_i) = \alpha,$$

where logit is explained in Section 2.3.1. Defines $\delta_i \propto n \cdot d_i^{-1}$ such that $\sum_{i=1}^{n} \delta_i = n$, based on the weights suggested by Volz and Heckathorn (2008). The pseudo-likelihood is written as follows:

$$L(\alpha \mid Y = y) = \prod_{i=1}^{n} \Pr(Y_i = y_i \mid \alpha)^{\delta_i}.$$

From a Bayesian perspective (see Section 2.4), inferences are based on the posterior distribution and Bastos, Bastos, et al. (2018, p. S18) used weakly informative priors for α. In this case, a pseudo-posterior is used. This estimator has the advantage of allowing prior information as convenient, but it suffers from the same limitations as VH and SH estimators, since the weights are derived from a Random Walk approximation.

Ott et al. (2019) and Fellows (2019) extended these estimators. The former presented a similar estimator to SH estimador, yet more robust. The latter introduced homophily into the model. Besides these estimators, Avery (2020) suggested binary logistic regression methods and other extensions through Generalized Linear Models (see Section 2.3.1).

2.2.5 Regression methods

According to Gile, Beaudry, et al. (2018, p. 86), "RDS suffers from two particular challenges for multivariate modeling: unknown sampling weights and unknown dependence structure." These two problems led to different approaches in the literature. Avery (2020, p. 13-15) has a good review on the topic. Spiller (2009) suggests to model dependence as mixed effects. Bastos, Pinho, et al. (2012) perform a binary regression to prevalence estimation through a hierarchical model where correlation structure was modelled as a Conditionally autoregressive (CAR) model (see Section 2.3.2). Yauck et al. (2021) includes homophily in a similar model, but with a Simultaneous Autoregressive (SAR) model (BANERJEE; CARLIN; GELFAND, 2003, p. 98) for correlation (see Section 3.5).

2.2.6 Bootstrap methods for uncertainty quantification

Quantifying the uncertainty about unknown quantities is an objective of modern statistics. Since most of the work on RDS was based on frequentist inference, bootstrapping is the most used technique. (SALGANIK, 2006) introduced a procedure assuming the Markov chain model on the categories built by Heckathorn. For each bootstrap dataset, the idea is:

a) select a starting node randomly from the whole sample;

b) let x_i be the category of the sampled individual. From the set of nodes recruited by some recruiter within the category x_i, select a node randomly;

c) the procedure ends when the sample size is n and calculates the point estimate.

After estimating the prevalence for each bootstrap dataset, the distribution of the point estimates derives confidence intervals and variance estimators. However, they tend to underestimate the uncertainty as pointed out by Goel and Salganik (2010, p. 6746). Gile, Beaudry, et al. (2018, p. 80) point out some of the limitations of this procedure that were addressed in the literature.

More recently, Baraff, McCormick, and Raftery (2016) developed the tree bootstrap. Different from Salganik (2006), the bootstrap samples are trees similar to the original one, rather than a unique chain of contacts. The design is:

a) select the first wave of seeds with replacement;

b) for each recruiter, sample with replacement from their recruits;

c) the process ends when the number of waves reaches the same as in the original tree.

After this process, it estimates variance and confidence intervals. Green, McCormick, and Raftery (2020) proved this estimator is consistent in the m-trees context.

2.3 Modelling strategies

In this section, we briefly describe two modelling strategies used throughout the dissertation: Generalized Linear Models (GLM) and CAR models.

2.3.1 Generalized linear models

Let $\boldsymbol{y} \in \mathbb{R}^n$ be a realization of a random variable $Y : \Omega \to \mathbb{R}^n$ associated with a phenomena such that each component Y_i is independent of the others. Set $\mu = \mathbb{E}[Y]$. The classical linear model assumes that $Y_i \overset{iid}{\sim} \text{Normal}(\mu_i, \sigma^2)$ and $\mu_i = \boldsymbol{X}_i\beta$, such that $\beta \in \mathbb{R}^k$ is an unknown parameter vector and $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ is the data, where $\boldsymbol{X}_{ij}$ is the measure of the j-th covariate in the i-th individual. Non-constant variance (heteroscedasticity) for each Y_i is a possible variation for this model.

GLM extend the above model. In order to understand this extension, we follow McCullagh and Nelder (2019, p. 27) setting

$$\eta = \boldsymbol{X}\beta \quad \text{and} \quad \eta_i = g(\mu_i), i = 1, \ldots, n,$$

such that $g(\cdot)$ is a monotonic differentiable function and is named *link function*. Therefore, "the link function relates the linear predictor η to the expected value μ "(MCCULLAGH;

NELDER, 2019, p. 31). Notice that in the classical linear model, g is the identity function, but it can be generalized. Another possible generalization is the distribution of Y, which may be any from the Exponential Family of distributions (ROBERT, 2007, p. 115).

When Y_i has Bernoulli distribution with probability of success $\mu_i \in (0,1)$, the link function must have its image over the open interval $(0,1)$ and domain in the real line. The classical are the following:

a) *logit*: $\eta = \log(\mu/(1-\mu))$ that represents the log odds of $Y_i = 1$;

b) *probit*: $\eta = \Phi^{-1}(\mu)$ where the $\Phi(\cdot)$ is the Normal cumulative distribution function;

c) *complementary log-log*: $\eta = \log(-\log(1-\mu))$.

This work focus on Logistic regression, which is the most common inferential procedure for binary response.

2.3.2 Conditionally autoregressive models

The *Conditionally Autoregressive* models have their first appearance in Besag (1974) with the objective of modelling spatial interactions among a finite number of random variables representing different regions. The joint probability specification is given by (BANERJEE; CARLIN; GELFAND, 2003, Section 3.3.1)

$$\omega_i \mid \omega_j, j \neq i \sim \text{Normal}\left(\rho \sum_j b_{ij}\omega_j / b_{i+}, \tau^{-1}/b_{i+}\right), i = 1, \ldots, n,$$

where $b_{i+} = \sum_{j=1}^{n} b_{ij}$. By Brook's Lemma (BROOK, 1964)

$$p(\omega_1, \ldots, \omega_n) \propto \exp\left\{-\frac{\tau}{2}\omega^T(D_b - \rho B)\omega\right\}, \tag{2.13}$$

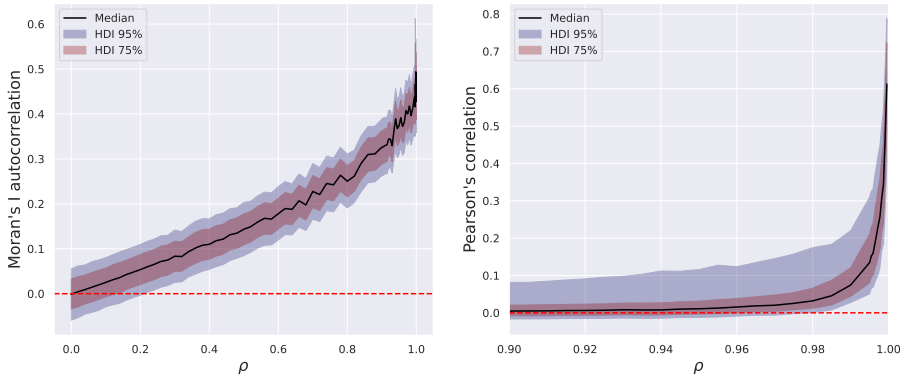
where $[D_b]_{ij} = b_{i+}$ and B is a symmetric *proximity matrix*, which connects the individuals. B_{ij} can measure the distance between i and j or indicate if they are connected. Relation (2.13) defines a normal distribution for $\omega_1, \ldots, \omega_n$ with mean zero and covariance matrix $[\tau(D_b - \rho B)]^{-1}$. The parameter τ is the spatially variation precision, while ρ controls spatial dependence. When $\rho = 1$ the model is called *Intrinsically Autoregressive* (IAR) and when $\rho = 0$, the regions are independent.

For the variables $\omega_1, \ldots, \omega_n$ have a proper prior distribution, the matrix $D_b - \rho B$ must be non-singular. This condition is met if $\rho \in (\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the smaller and higher eigenvalues of $D_b^{-1/2} B D_b^{-1/2}$, respectively (BANERJEE; CARLIN; GELFAND, 2003, p. 94). By simple calculations, one can prove that $\lambda_{\min}^{-1} < 0 < \lambda_{\max}^{-1}$, then this interval is not empty.

For many applications, CAR reproduces the strong spatial correlation between neighbors only when ρ is close to the limits. Moreover, its interpretation is not so clear.

To verify this fact, we generate a random matrix $\tilde{B}$ with binary entrances and adjust $B = 0.5(\tilde{B} + \tilde{B}^T)$ yielding a symmetric matrix. We fix $\tau = 1$ and for each ρ we generate 10000 datasets of 500 individuals from CAR model. Moran's I spatial autocorrelation and the distribution of the Pearson's correlation of each pair were calculated. Figure 4 presents the HDI for the distribution of the Pearson's correlations among the individuals and the Moran's I autocorrelation for different values of ρ. Notice the non-linearity of Pearson's graphic with only high values of ρ generating large correlations. Table 1 shows that when n increases, with $\rho = 0.95$, the Pearson's correlation decreases fast and ρ has to be even higher for observing any higher value.

Figure 4 – Moran's I spatial autocorrelation and Pearson's correlation statistics for different values of ρ



Source: Prepared by the author (2021). The blue regions indicates the HDI 95%, while the red on is the HDI 75%. The black line denotes the median.

Table 1 – 75% Interval for Pearson's correlations among individuals for different values of n

n	Interval 75%	
	Quartile 12.5%	Quartile 87.5%
10	0.48	1.0
50	0.05	0.32
100	0.03	0.17
500	-0.002	0.032
1000	-0.007	0.02

Source: Prepared by the author (2021).

2.4 Bayesian statistics

We can represent our beliefs and information about unknown quantities through probabilities. There are two more common interpretations: frequentist and Bayesian. While the frequentists define probability as the limit of a frequency in a large number of trials, the Bayesians represent an individual's degree of belief in a statement that is updated

given new information. This philosophy allows assigning probabilities to any event, even if a random process is not defined (STATISTICAT, 2016).

In 1761, Reverend Thomas Bayes wrote for the first time the Bayes' formula relating the probability of a parameter after observing the data with the evidence (written through a likelihood function) and previous information about the parameter. Pierre Simon Laplace rediscovered this formula in 1773 (ROBERT, 2007), and this theory became more common in the 19th century. After some criticisms, a modern treatment considering Kolmogorov's axiomatization of the theory of probabilities started after Jeffreys in 1939.

Therefore, Bayesian inference is the process of inductive learning using Bayes' rule, where inductive means that characteristics of a population are learned from a subset of it. We generally express numerical characteristics of the population as a parameter θ which is indirectly observed through numerical descriptions y of the population. Both are uncertain until the observation of a sample, when its information can decrease our uncertainty about the population characteristics (HOFF, 2009, p. 1-2).

The set of all possible outcomes y forms the *sample space* $\mathcal{Y}$, while the set of all possible parameters forms the *parameter space* Θ. Bayesian inference is composed by the following:

a) *prior distribution:* a probability distribution defined over Θ that quantifies our beliefs about θ before observing the data;

b) *sampling model:* a probability distribution of the data generation process that express our belief that $y \in \mathcal{Y}$ is the outcome when $\theta \in \Theta$ is true. When it is seen as function of the parameter, it is called *likelihood function*;

c) *loss function:* only in a decision theory framework, it measures the error of a estimative $\delta \in \Theta$ in comparison to θ;

d) *posterior distribution:* once we get the data y, it represents our updated beliefs about the parameter. All the inferences are based on this probability distribution.

Bayes' theorem establishes that when the sampling model is absolutely continuous with respect to some measure ν with conditional density $f_{Y|\theta}(y \mid \theta)$, and the prior distribution is a well defined probability measure μ_θ, the posterior distribution $\mu_{\theta|Y}(\cdot \mid y)$ is absolutely continuous with respect to μ_θ almost surely and its Radon-Nikodym derivative is (SCHERVISH, 2012, p. 16)

$$\frac{d\mu_{\theta|Y}}{d\mu_\theta}(\theta|y) = \frac{f_{Y|\theta}(y \mid \theta)}{\int_\Theta f_{Y|\theta}(y|t)d\mu_\theta(t)}. \tag{2.14}$$

When the prior distribution is absolutely continuous with respect to the Lebesgue measure, equation (2.14) resumes to

$$p(y|\theta) = \frac{f_{Y|\theta}(y \mid \theta)\pi(\theta)}{\int_\Theta f_{Y|\theta}(y \mid \theta)\pi(t)\,dt}. \tag{2.15}$$

Another important concept used throughout the text is the following

Definition 2.4.1. Let $\mathcal{F}$ be a family of probability definitions parametrized by $\theta \in \Theta$. $\mathcal{F}$ is *conjugate* for a likelihood $f_{Y|\theta}(y \mid \theta)$ when for every $\pi \in \mathcal{F}$, the posterior $p(y \mid \theta) \in \mathcal{F}$. The prior is called *conjugate prior* for the likelihood $p(y \mid \theta)$, and prior and posterior are *conjugate distributions*.

2.5 Computational methods

Over the text, we use a state-of-art implementation of *Hamiltonian Monte Carlo* (HMC) to sample from the posterior distribution of the parameters. HMC is an advanced technique, and it is especially effective for hierarchical models. In this section, we illustrate the method and its diagnosis. We also discuss the *Metropolis-within-Gibbs* technique used to sample graphs from Crawford (2016)'s model.

Markov chain Monte Carlo is a technique for exploring the space of a target probability distribution using dependent samples built by a Markov chain. Through Markov chain's convergence theory, the method guarantees that, under some regularity conditions, the chain will eventually search all the space. Warm-up iterations are necessary before the sampling to achieve the stationary distribution. Based on the samples, we can derive Monte Carlo estimators for quantities of interest.

HMC improves this idea by using the differential geometry of the target probability distribution with auxiliary parameters called *momenta*. The first appearance in the literature occurred at the end of the 1980s (BETANCOURT, 2017, p. 3). Nowadays, this method is a very popular in computational statistics. The following is based on Betancourt (2017) and Betancourt (2016).

Instead of randomly moving in the parameter space with uninformed jumps, the technique uses the direction from the vector field given by the gradients to trace out a trajectory through the *typical set* - the high mass region, which has a significant contribution to the expectations. However, if it only used the gradient, the trajectory would pull towards the mode of the distribution, so we need to impose more geometric constraints. As in a physical system, the method endows momentum to the system to turn it conservative.

First, we extend the parameter space $\Omega \subseteq \mathbb{R}^D$ introducing auxiliary momentum parameters $p \in \Omega^C$ of dimension D. We also extend the probability distribution $\pi(q)$ in Ω to a joint distribution called *canonical distribution* through

$$f_{\Omega,\Omega^c}(q,p) = f_{\Omega^c|\Omega}(p \mid q)\pi(q),$$

where $f_{\Omega^c|\Omega}(p \mid q)$ is the *cotangent disintegration*. By marginalizing q in $f_{\Omega,\Omega^c}(q,p)$, we recover $\pi(q)$. Particularly, we use the *Hamiltonian* function $H(q,p)$ to define

$$\pi(q,p) = e^{-H(q,p)}, \tag{2.16}$$

with $H(q,p)$ indicating the energy at point q. By relation (2.16),

$$H(q,p) = -\log f_{\Omega^c|\Omega}(p \mid q) - \log \pi(q) = K(p,q) + V(q),$$

where K is the *kinetic energy* and V is the *potential energy*. The kinetic energy is subject to implementation. The *Hamiltonian equations* generate the vector field thorough the following expression

$$\begin{aligned}\frac{dq}{dt} &= \frac{\partial H}{\partial p} = \frac{\partial K}{\partial p} \\ \frac{dp}{dt} &= -\frac{\partial H}{\partial q} = -\frac{\partial K}{\partial q} - \frac{dV}{dq}.\end{aligned}$$

The Hamiltonian defines a family of maps $\phi_t : (q,p) \to (q,p), \forall t \in \mathbb{R}$ and $\phi_t \circ \phi_s = \phi_{s+t}$, which is called *Hamiltonian flow*. From a starting point q we can lift to $\Omega \times \Omega^C$ sampling $p \sim f_{\Omega^C|\Omega}(\cdot \mid q)$, and explore the space through the Hamiltonian flow $\phi_t(q,p)$. After the exploration, project to Ω to obtain a new value of q. It remains to choose the cotangent disintegration, the integration time t, and numerical approximations for the Hamiltonian equations.

"The Hamiltonian flow preserves the Hamiltonian" (METACADEMY, 2021), which implies that it remains in the same level set $H^{-1}(E)$ of the starting point (q,p). The Markov transition jumps from a level set to another, explore that level set and then project back to Ω. Therefore, there are two main phases of the technique: exploration across level sets, and along each level set. Betancourt (2016, p. 6) decomposes the joint distribution into the distribution over level sets and the marginal energy distribution $\pi_H = \pi_{H^{-1}(E)}\pi_E$.

After the sampling, diagnosing the inferences is an important step that help to show if they are valid. Throughout the text, we use them to help the understanding of problems and limitations, and the construction of possible solutions. Below there are the main available diagnostics:

a) $\hat{R}$ greater that 1.01: this is a diagnostic borrowed from the MCMC theory. We run M parallel chains and compare between and within chain estimates thorough variance estimators. If the starting point is negligible, we expect each chain be similar, yielding $\hat{R} \approx 1$. Otherwise, this value will be higher. The threshold is subject to discussion, but the value 1.01 is suggest by Vehtari et al. (2019, p. 4).

b) *divergences*: energy level sets, which contain regions of high curvature, are challenging to the finite time discretization of the integrator. These regions compromise the accuracy of the result, causing divergences and speeding up the trajectories towards infinite energies. Divergences make inferences invalid. Decreasing the step size is a first way to understand the problem;

c) transitions that hit the *maximum tree depth*: while divergences point out invalid estimates, the maximum tree depth indicate inefficient searches. A reparametrization of the model and increasing its value are possible solutions for this warning.

d) low E-BFMI value: As stated by Betancourt (2016, p. 6), we expect the distribution of the energies induced by the position q be similar to the marginal distribution π_E. To verify this, it is used the the *Bayesian fraction of missing information* given by the expression

$$BFMI = \frac{\mathbb{E}_\pi[\mathrm{Var}_{\pi_{E|q}}(E \mid q)]}{\mathrm{Var}_{\pi_E}[E]},$$

which measures the insufficient energy variation in the momentum resampling. When this value is close to 0, the average variance of the energy induced by q is much smaller than the marginal, which represents a very slow exploration through level sets. In other words, the Markov chain did not efficiently walk through the target distribution. Reparametrization of the model is a good option that we use here.

e) low *effective sample size* (ESS): the ESS quantifies the necessary independent samples from the target distribution to give similar estimator obtained by the MCMC algorithm. It is divided in bulk-ESS for measuring the efficiency in location estimates and tail-ESS for tail estimates, such as 5% and 95% quantiles.

f) *trace plot* and other visual diagnostics: the traceplot indicates the space explored for each parameter. It can diagnose if some chain got stuck in a region or if there are two different chains. Other visual diagnostics such as posterior samples scatter plots can show the quality of the estimates.

One limitation of HMC is the need for a continuous probability space for the parameters to define the gradient correctly. Because of that, it can not sample the graphs from Crawford (2016) model, and other method should be used. The two commonly used procedures for MCMC algorithms besides HMC are *Metropolis-Hastings* and *Gibbs sampler*. We give a short discussion about these methods.

Metropolis-Hastings builds a Markov chain starting from an arbitrary initial point $X^{(0)}$ and a probability distribution with density $q(y \mid x)$ called *proposal distribution*. The transition kernel is designed to have the target distribution π as the stationary distribution. For each iteration t, we draw X from $q(\cdot \mid X^{(t-1)})$ and accept the new value with probability $\alpha(X \mid X^{(t-1)})$, where

$$\alpha(X \mid X^{(t-1)}) = \min\left(1, \frac{\pi(X)q(X^{(t-1)} \mid X)}{\pi(X^{(t-1)})q(X \mid X^{(t-1)})}\right). \tag{2.17}$$

Imposing some condition on q, such as $\pi(x), \pi(y) > 0 \implies q(x \mid y) > 0$, the chain is irreducible (ROBERT; CASELLA; CASELLA, 2004, p. 274) and we can assume convergence of Monte Carlo estimates to quantities of interest.

On the other hand, Gibbs sampler builds a Markov chain using the full conditional probabilities and it is more popular in high-dimensional distributions. One particular formulation that we use in this work is the *Systematic scan Gibbs sampler*. Suppose $X = (X_1, X_2)$ and let $X^{(0)} = (X_1^{(0)}, X_2^{(0)})$ be an arbitrary initial point. For each iteration t, we sample $X_1^{(t)} \sim \pi_{X_1|X_2}(\cdot \mid X_2^{(t-1)})$ and $X_2^{(t)} \sim \pi_{X_2|X_1}(\cdot \mid X_1^{(t)})$. The kernel of this distribution admits π as stationary distribution. With some regularity condition, we can show that the sampling generates a irreducible and recurrent Markov chain such shat Monte Carlo estimates converge to the integrals we are interested in.

When the full conditionals are not available, Metropolis-Hastings can be used inside the Gibbs sampling cycle generating a sub-chain of size T. The scheme generated is called *Metropolis-within-Gibbs*. Gamerman and Lopes (2006, p. 213) states that it is usual to use $T = 1$ since large values are unnecessary.

3 Methodology for prevalence estimation

Fisher (1922, p. 311) stated that the objective of statistics is to reduce the data since its volume is impossible to comprehend by researchers. In that sense, few parameters should represent the whole phenomenon catching the most relevant information. Years later, J. Neyman studied the theory of modelling which can be divided in three aspects (LEHMANN, 2012, p. 161):

a) models of complex phenomena are created by putting together simple building elements that the researcher is familiar with and can handle;

b) there are two types of models: the *explanatory models*, which will be focused on this work, and the *interpolatory formulae*.

c) An explanatory theory necessitates a thorough understanding of the scientific context of the problem. In this regard, we investigated questions involving Respondent-driven sampling and prevalence estimation as introduced in Chapter 2.

In this chapter, we develop models that enclose these ideas building each block separately. For a Bayesian modelling, we assume that each parameter of the model has a probability distribution that incorporates the researcher's uncertainty about it. For each individual, we observe k covariates that are possible risk factors represented by the vector $\boldsymbol{x}_i \in \mathbb{R}^k$ of the i^{th} individual. We denote θ_i the probability of the i-th individual have been exposed to the disease that depends on the prevalence θ and $\boldsymbol{x}_i$. We also consider the dependence of sampling from RDS as a spatial random effect. The probability of positive test in the i^{th} individual is denoted by p_i.

Another important feature of the model is that sensitivity and specificity have the same distribution for all individuals and it only depends on the test used. This is an assumption that must be analysed for each particular case. For instance, COVID-19 Sofia test has different sensitivity and specificity for symptomatic and asymptomatic individuals (Table 1 MITCHELL et al., 2021, p. 3).

From above, we develop three different models: the first considers perfect tests, that is, $\gamma_s = \gamma_e = 1$ and no spatial random effect; the second considers imperfect tests, regarding γ_s and γ_e, but ignoring the RDS structure; and the third one has imperfect tests and RDS structure. Some considerations are made to improve the model's limitations.

The implementation of the following models were in the statistical computation platform Stan (CARPENTER et al., 2017) within Python Interface PyStan (RIDDELL; HARTIKAINEN; CARTER, 2021) which uses an implementation for HMC algorithm. All the codes are provided in Appendix C. For plotting the diagnosis and the distributions,

ArviZ (KUMAR et al., 2019) and Matplotlib (HUNTER, 2007) Python packages were used. For handling networks, we use NetworkX (HAGBERG; SWART; S CHULT, 2008). All experiments were run on a Linux PC with Intel(R) Core(TM) i5-7200U 2.5GHz processor (4 cores) and 8 GB of memory.

3.1 Perfect tests

The first model supposes the samples are independent and the test is perfect, which means that $\theta_i = p_i$ for all i. Therefore it only considers the risk factors $\boldsymbol{x}_i$.

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Bernoulli}(\theta_i), \\ g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T \beta, \end{aligned} \tag{3.1}$$

where $g(\cdot)$ is the logit function. The parameter $\beta \in \mathbb{R}^k$ is the risk effects. For Bayesian inference, priors on β and θ must be included. We suppose they are independent before the data and use $\beta \sim \text{Normal}(\mu_\beta, \Sigma_\beta)$ and $\theta \sim \text{Beta}(a^p, b^p)$, where the vector $\mu_\beta \in \mathbb{R}^k$, the symmetric positive-definite matrix $\Sigma_\beta \in \mathbb{R}^{k\times k}$, and the positive real values $a^p, b^p \in \mathbb{R}_{>0}$ are fixed hyperparameters. Inferences about β and θ are based on the posterior distribution. Keeping the notation of Section 2.3.1, we denote the covariate matrix by $\boldsymbol{X}$.

Remark 3.1.1 (Interpretation of prevalence). According to the model formulation, if the risk factors are zero, i.e. $\boldsymbol{x}_i = 0$, the probability of the i-th individual having been exposed is the prevalence θ, which means that in a population with no risk effects, the probability of a person having the disease is exactly the proportion in this population.

3.1.1 Identifiability

Identifiability is a key concept in statistical modelling. It formalizes the basic idea of identifying a model parameter from data. Roughly, a non-identifiable model cannot distinguish values of the parameter after observing data through the sampling model. A formal definition regards the likelihood function (XIE; CARLIN, 2006, p. 3459):

Definition 3.1.1. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be the family of probability distributions for $\mathcal{Y}$. This model is *identifiable* if for any $\theta', \theta'' \in \Theta$,

$$\forall y \in \mathcal{Y}, P_{\theta'}(Y = y) = P_{\theta''}(Y = y) \implies \theta' = \theta''.$$

The family distribution from model (3.1) is the logistic regression parametrized by (θ, β) and conditioned on observing the regressor $\boldsymbol{X}$, with $\mathcal{Y} = \{0, 1\}^n$. Defining $\beta_0 = g(\theta)$, we may rewrite it as

$$Y_i \mid \tilde{\beta}, \tilde{\boldsymbol{x}}_i \sim \text{Bernoulli}(g^{-1}(\tilde{\boldsymbol{x}}_i^T \tilde{\beta})),$$

such that $\tilde{\beta}$ concatenate β_0 and β, and $\tilde{\boldsymbol{x}}_i$ concatenate 1 and $\boldsymbol{x}_i$. Küchenhoff (1995, p. 7) gives a formal proof for the identifiability of this representation.

In the Bayesian paradigm, inferences are based on the posterior distribution. Therefore, identifiability should consider the prior distribution. Lindley (1972, p. 46) argued that proper priors are sufficient to handle identifiability problems in the Bayesian perspective, which means that a well-defined posterior probability distribution is enough for parameter identification. A formal definition for *Bayesian identifiability* is the following: if $p(\theta \mid \beta, y, \boldsymbol{X}) = p(\theta \mid \beta)$, the data y is uninformative for θ when β is known. The definition is analogous if β and θ change places. However, Gelfand and Sahu (1999, p. 248) proved that this definition is equivalent to likelihood identifiability.

Despite the identifiability of the model, it may be hard to sample from the posterior distribution depending on the value of $\boldsymbol{x}$. As an example, consider the following experiment:

(i) generate 500 covariates $X_i \sim \text{Normal}(15, 1)$;

(ii) let $\beta = 0.1$, $\theta = 0.1$, and $\theta_i = g^{-1}(g(\theta) + X_i\beta)$ for $1 \leq i \leq 500$;

(iii) for each i, sample $Y_i \sim \text{Bernoulli}(\theta_i)$;

(iv) let $a^p = 1$, $b^p = 1$, $\mu_\beta = 0$, and $\Sigma_\beta = 1$ the hyperparameters for the prior distributions (weakly informative);

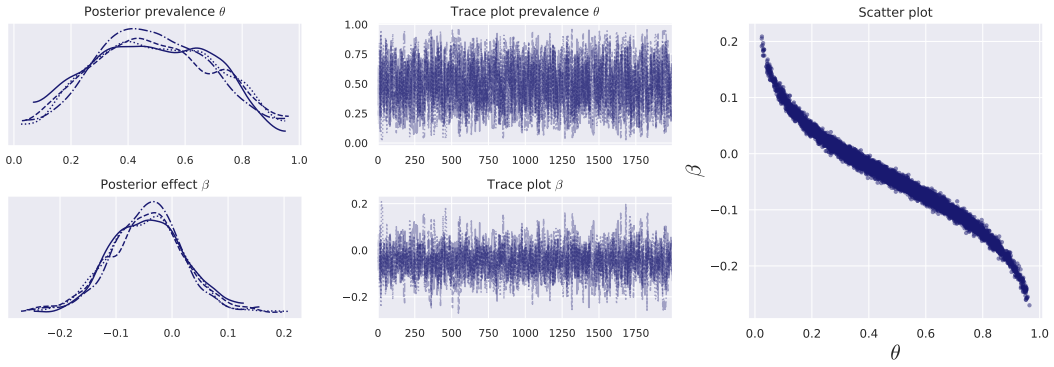
(v) make 1000 warm-up and 1000 sampling iterations using Stan given the data $(Y_1, X_1), \ldots, (Y_n, X_n)$.

(vi) make 2000 warm-up and 2000 sampling iterations using Stan given the data $(Y_1, X_1), \ldots, (Y_n, X_n)$.

The HMC sampler took around 8.39s. Figure 5 presents the results through the posterior distribution, the trace plot, and the strong posterior correlation between θ and β. To address this problem, subtracting the mean $\bar{x}$ is a default procedure (OGLE; BARBER, 2020, p. 5). After centering the data around the mean, the HMC sampler took around 1.39s, and the improved results are shown in Figure 6.

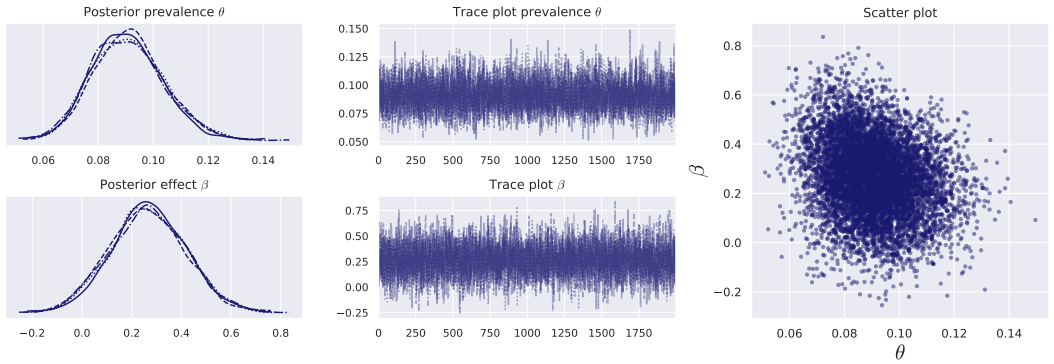
We observe that the interpretation of prevalence from Remark 3.1.1 changes from centred and uncentered since the meaning of $\boldsymbol{x}_i = 0$ is different. Along with this discussion, it is usual to divide the centred variable by its standard deviation, to put all predictors on a common scale. Discussions about the problems caused by standardizing are outside of the scope of this work. Gelman (2008) suggests to divide de continuous variables by 2 times de standard deviation to allow "the coefficients to be interpreted in the same way as binary deviation." (GELMAN, 2008, p. 2867) Binary inputs are not standardized since their coefficients are easily interpretable.

Figure 5 – Posterior distribution, trace plot, and posterior samples of parameters θ and β from model (3.1) with uncentered covariate.



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs.

Figure 6 – Posterior distribution, trace plot, and posterior samples of parameters θ and β from model (3.1) with centralized covariate.



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs.

Other identifiability problems arising from the input variables are collinearity and *separation* (GELMAN; JAKULIN, et al., 2008, p. 1360-1361). The latter occurs if a linear combination of a subset of the predictors gives a perfect prediction for the binary outcome. For instance, when a linear combination of the predictors is greater than a threshold if and only if $y = 1$.

3.1.2 Simulated data

To present a sanity check about the functionality of model (3.1) and to validate the properties of the estimation procedure, we simulate fake data from the model and make inferences about the result. We follow the experiment from Section 3.1.1. Table 2 summarizes the experiment settings.

We primarily look at the settings from experiment 1. With a non-informative prior for θ (Beta$(1/2, 1/2)$) and a weakly informative for β (zero mean and covariance matrix four times the identity matrix), Figure 7 shows the posterior distributions for the

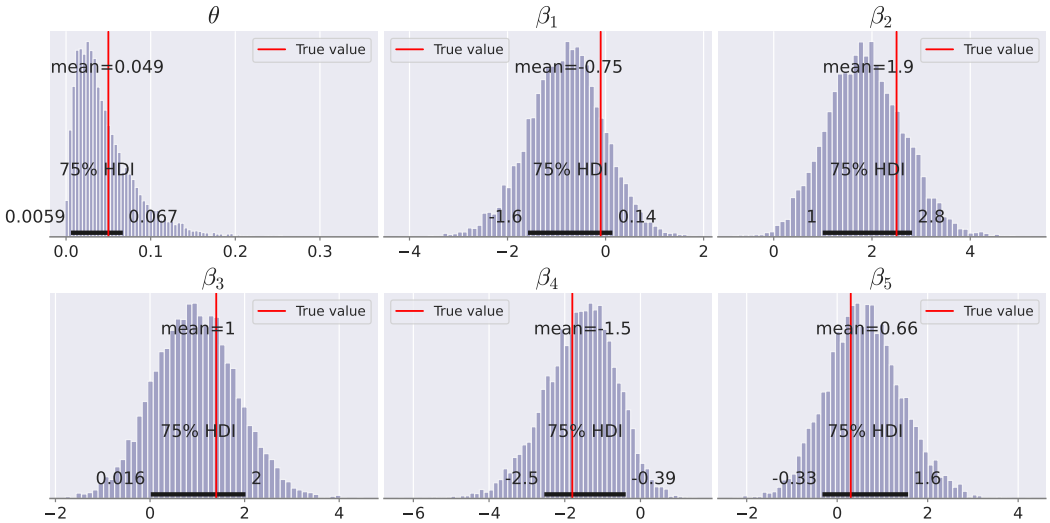
Table 2 – Experiment settings for the simulation of model (3.1).

Experiment	n	k_c (normal)	k_c (cauchy)	k_b	β	θ
1	100	3	0	2	[-0.1, 2.5, 1.4, -1.8, 0.3]	0.05
2	100	3	0	2	[-0.1, 2.5, 1.4, -1.8, 0.3]	0.9
3	100	2	2	1	[-0.1, 2.5, 1.4, -1.8, 0.3]	0.1
4	5000	40	5	5	F distribution	0.1

Source: Prepared by the author (2021). We denote n for number of samples, k_c for the number of continuous variables, and k_b for binary variables. Between paranthesis, *normal* means that the variables were generated from a Multivariate Normal with prespecified parameters, and *cauchy* from a Cauchy distribution. F distribution is Normal($\mu = 0, \sigma = 2$) with probability 0.3, and 0 otherwise.

parameters. The prevalence estimate is good despite the non-informative prior. When the distance between the prior and the true value is large, the inferences seem to be biased. However, this makes sense regarding the model. For instance, for β_2, before observing the data, we put 0.7 mass probability for values less than 0.1. The data decreased it to 0.125. This highlights the importance of a well-defined prior distribution. The values for bulk-ESS were greater than 3000 for all parameters, while Tails ESS were greater than 2200 with 1000 warmup and 1000 sampling iterations, and 4 chains. For all parameters $\hat{R} = 1$. Trace plots and scatter plots were also good and we omit here since they do not bring new information for the discussion.

Figure 7 – Posterior distribution for parameters of model (3.1) with experiment 1 settings.

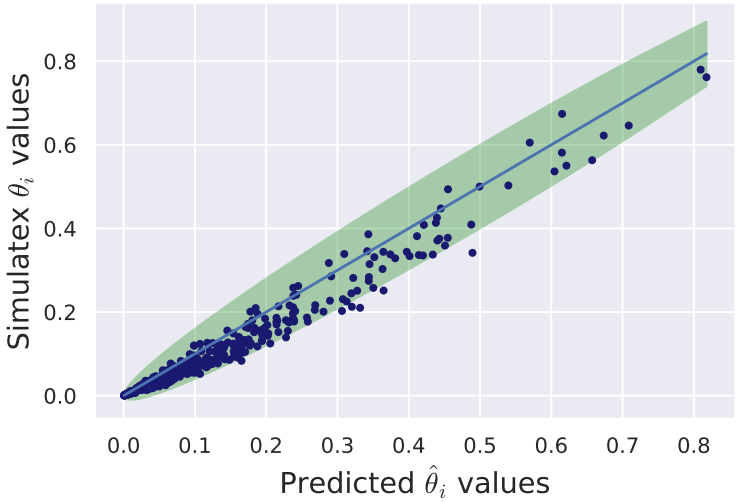


Source: Prepared by the author (2021) and based on Stan and ArviZ outputs. The red line represents the true value inputed for the simulation.

Figure 8 compares the predicted and simulated probabilities of having the disease θ_i. Although we are performing Bayesian inference, frequentist properties can be accessed through simulation. After 1000 simulations varying the input data Y, the 75% credible interval included the true parameters in 75.8%, 78.8%, 76.4%, 77.5%, 67.3%, and 72.2% of the times, respectively for $\theta, \beta_1, \ldots, \beta_5$. Each simulation had 100 samples and weakly

informative priors for β and θ.

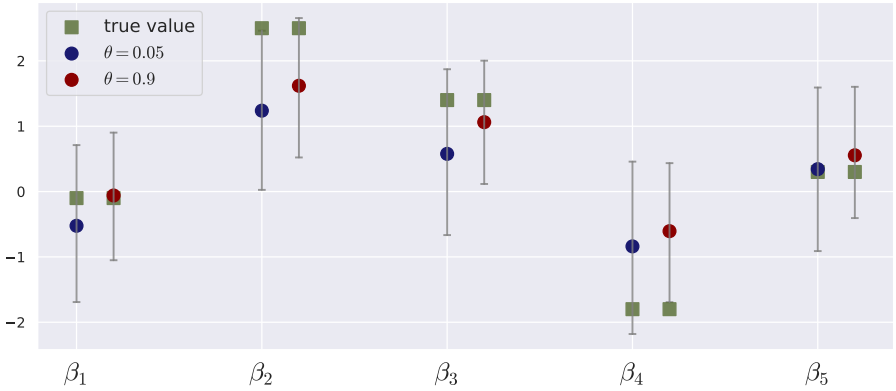
Figure 8 – Comparing predicted and simulated probabilities of having the disease from model (3.1) with experiment 1 settings.



Source: Prepared by the author (2021) and based on Stan output. The green area is delimited by the curves generated by $2\sqrt{\theta_i(1-\theta_i)/n}$, where $n = 500$ is the number of points. This area is a proxy for ± 2 standard-error bounds.

Experiment 2 is used to see if these properties repeat when the prevalence is higher. The same regressors were used for the comparison, but the input data Y were generated with different prevaleces. When prevalence was 0.9, the estimates were a little higher for all coefficients as Figure 9 presents. This is related to the fact that the posterior mean of θ underestimated the true value for this experiment. After increasing the number of sampled individuals, the estimates were closer, as expected.

Figure 9 – Comparing posterior mean and 95% credibility intervals for β in model (3.1) with the same regressors $\boldsymbol{X}$ but different prevalences.

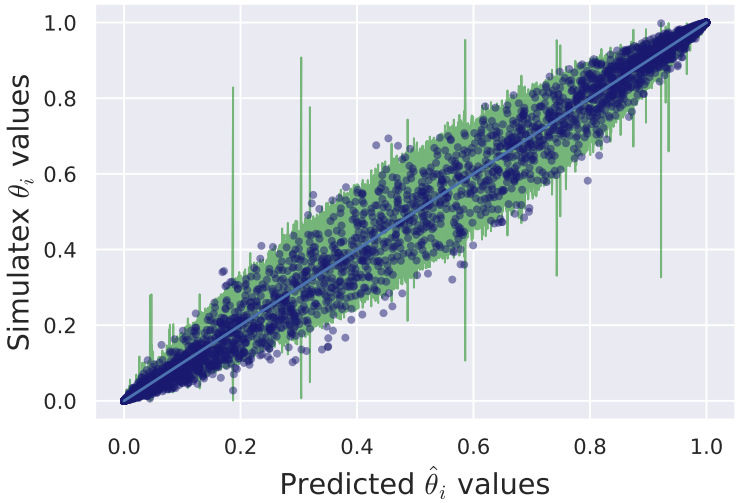


Source: Prepared by the author (2021) and based on Stan output.

The third experiment aims to analyse what happens if some covariates have a heavier tail. No big difference was noticed despite the existence of some individuals very

different from the others. At last, the fourth experiment increases the dimensionality to observe the number of effective samples. Each chain took around three minutes, instead of the 3s needed for the previous experiments. From the 51 parameters, 48 had the true values in the 95% HDI credible interval. The bulk-ESS was greater than 4500 for 95% of the parameters. Figure 10 presents how the predicted probabilities for each individual behaves in this case.

Figure 10 – Comparing predicted and simulated probabilities of having the disease from model (3.1) with high dimension for experiment 4 settings.



Source: Prepared by the author (2021) and based on Stan output. The green area indicated the 95% credible interval for each predicted θ_i.

3.2 Sensitivity and specificity

In this section, we describe a model for estimating the sensitivity and specificity of a diagnostic test. This model is relevant to analyze and experiment with different prior specification approaches. Suppose having a gold standard test and another test, for instance, a simpler, faster, or less invasive one, which we want to estimate the accuracy by the sensitivity and specificity. In this scenario, true positive (negative) individuals are those who tested positive (negative) by the gold standard. Therefore, in a population with n_{γ_s} true positives and n_{γ_e} true negatives, we denote

$$\begin{aligned} y_{\text{pos}} \mid \gamma_s &\sim \text{Binomial}(n_{\gamma_s}, \gamma_s), \\ y_{\text{neg}} \mid \gamma_e &\sim \text{Binomial}(n_{\gamma_e}, \gamma_e), \end{aligned} \tag{3.2}$$

such that y_{neg} are negative tests on known negative subjects and y_{pos} are positive tests on known positive. Chart 2 presents the Two-by-two formulation from Chart 1 with the model objects.

In Bayesian analysis, we have to define a prior distribution with density π for the parameters (γ_e, γ_s). For this, we consider three different approaches:

Chart 2 – Two-by-two table with the model specification.

	$Y = 0$	$Y = 1$	Total
$Y^{\text{true}} = 0$	y_{neg}	$n_{\gamma_e} - y_{\text{neg}}$	n_{γ_e}
$Y^{\text{true}} = 1$	$n_{\gamma_s} - y_{\text{pos}}$	y_{pos}	n_{γ_s}
Total	$n_{\gamma_s} + y_{\text{neg}} - y_{\text{pos}}$	$n_{\gamma_e} + y_{\text{pos}} - y_{\text{neg}}$	$n_{\gamma_s} + n_{\gamma_e}$

Source: Prepared by author (2021).

a) prior distributions are specified independently for each parameter and each one has a beta distribution, i.e,

$$\pi(\gamma_e, \gamma_s) = \pi(\gamma_e)\pi(\gamma_s) \propto \gamma_s^{a_s}(1-\gamma_s)^{b_s}\gamma_e^{a_e}(1-\gamma_s)^{b_e},$$

for a_s, b_s, a_e, and b_e being pre-determined positive real hyperparameters;

b) bivariate normal distribution in the log odds space, i.e,

$$(\text{logit}(\gamma_e), \text{logit}(\gamma_s)) \sim \text{Normal}(\mu_\gamma, \Sigma_\gamma),$$

such that the vector $\mu_\gamma \in \mathbb{R}^2$ and the covariance matrix $\Sigma_\gamma \in \mathbb{R}^{2\times 2}$ are pre-determined hyperparameters;

c) a bivariate beta distribution described in Appendix A with parameters α_1, α_2, α_3, $\alpha_4 \in \mathbb{R}_{>0}$.

If more studies about the same diagnostic test are available, a *hierarchical partial pooling* approach can be adopted for prior specification, as explained by Gelman and Carpenter (2020, p. 1272-1274) and by Guo, Riebler, and Rue (2017, p. 2-3).

3.2.1 Independent beta distribution priors

If the knowledge of the specificity affects the range of most possible values of the sensitivity, or vice-versa, there is antecedent information about the correlation between the parameters. When this is not the case, a possible independent prior formulation is the usage of Beta distribution since it is bounded in the interval $[0, 1]$ and it is reasonably flexible in its shape. Another good reason for this choice is that the beta distribution forms a conjugate family with the likelihood binomial distribution (see Definition 2.4.1), which is more tractable numerically. Therefore we have the following prior specification

$$\gamma_s \sim \text{Beta}(a_s, b_s),$$
$$\gamma_e \sim \text{Beta}(a_e, b_e),$$

which leads to the following posterior distribution from the likelihood (3.2):

$$\gamma_s \mid y_{\text{pos}} \sim \text{Beta}(a_s + y_{\text{pos}}, b_s + n_{\gamma_s} - y_{\text{pos}}),$$
$$\gamma_e \mid y_{\text{neg}} \sim \text{Beta}(a_e + y_{\text{neg}}, b_e + n_{\gamma_e} - y_{\text{neg}}).$$

Notice that this particular likelihood function does not add any correlation to the parameters since it treats each one separately. The interpretation of the beta distribution parameter is in terms of the number of successes for the first parameter and failures for the second parameter. With respect to Section 2.1.1, since the likelihood from this model does not add any correlation to the posterior distribution, the prior distribution has to give this information to it, when necessary.

3.2.2 Bivariate normal distribution in the log odds space

This approach was designed by Chu and Cole (2006) to jointly analyse sensitivity and specificity from a set o studies. In their work, the prior specification allows the incorporation of regressors. We consider it without the regressors, which simplifies to

$$\begin{pmatrix}\text{logit}(\gamma_s)\\ \text{logit}(\gamma_e)\end{pmatrix} \sim \text{Normal}(\mu_\gamma, \Sigma_\gamma), \text{ with } \Sigma_\gamma = \begin{pmatrix}\sigma^2_{\gamma_s} & \rho\sigma_{\gamma_s}\sigma_{\gamma_e}\\ \rho\sigma_{\gamma_s}\sigma_{\gamma_e} & \sigma^2_{\gamma_e}\end{pmatrix},$$

such that $\sigma_{\gamma_s} > 0$ and $\sigma_{\gamma_e} > 0$ are the standard deviations from log odds of sensitivity and specificity, respectively, and ρ is the correlation between the parameters in the log odds space. The possible problem with this prior approach is that the moments of logit normal distribution are not in closed form and there is no available formula to compute $\mathbb{E}[\gamma_s]$ from the parameters of the normal distribution (WILL KURT, 2021).

3.2.3 A bivariate beta prior

A common practice is to define the beta distribution as a prior distribution over $[0, 1]$. When more dimensions are necessary, the Dirichlet distribution is a possible generalization with the restriction that the parameters live in the simplex of lower dimension, i.e., if $\boldsymbol{x} \in [0, 1]^d$ has Dirichlet distribution, there is the restriction $\sum_{i=1}^d \boldsymbol{x}_i = 1$. Because of that, Olkin and Trikalinos (2015) build a bivariate beta distribution with positive probability in $(0, 1)^2$, with marginals having beta distribution and correlation over the interval $(-1, 1)$. Appendix A presents a detailed derivation. The prior specification is as follows:

$$\begin{aligned}(U_1, \ldots, U_4) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_4),\\ \gamma_e &= U_1 + U_2,\\ \gamma_s &= U_1 + U_3.\end{aligned}$$

Prior distributions can be placed on the hyperparameters α_i. In this work, we employ

$$\alpha_i \sim \text{Gamma}(a^i, b^i), \quad a^i, b^i > 0, \quad \text{for } i = 1, \ldots, 4.$$

Suppose the researcher has prior information about specificity and specificity, such as their mean and correlation.

To specify the prior hyperparameters using prior information, Section A.4 discusses the results when the researcher prespecifies $m_s = \mathbb{E}[\gamma_s], m_e = \mathbb{E}[\gamma_e], v_s = \text{Var}(\gamma_s), v_e =$

$\text{Var}(\gamma_e)$, and $\rho = \text{Cor}(\gamma_s, \gamma_e)$. Since system (A.15) usually has no solution, an optimization problem is solved with m_s and m_e fixed, and the other parameters being an approximation of the researcher's input values. For more details, see Appendix A.

a) having α_i fixed: we search for $\alpha_i = \hat{\alpha}_i > 0$ thorough the values of m_s, m_e, ρ, v_s, and v_e. An optimization problem is searching for $\hat{\alpha}_i$ which gives moments $\text{Var}(\gamma_s), \text{Var}(\gamma_e)$, and $\text{Cor}(\gamma_s, \gamma_e)$ as close as possible to the input values, and $m_s = \mathbb{E}[\gamma_s], m_e = \mathbb{E}[\gamma_e]$. A variation of this method would include m_s and m_e in the optimization problem and it is suggested when believes about m_s and m_e are less strong.

b) having α_i as a hierarchical parameter: we first estimate $\hat{\boldsymbol{\alpha}}$ the same way as described above and set $\mathbb{E}[\alpha_i] = a^i/b^i = \hat{\alpha}_i \implies a^i = b^i\hat{\alpha}_i$. The parameter $b_i = \hat{\alpha}_i/\text{Var}(\alpha_i)$ is a inversely proportional quantity to the spread of parameter α_i. The interesting thing about this approach is that it allows the prior to move more freely, specially when the input values are far from the estimated ones.

Remark 3.2.1. When α is a random variable, the adapt delta parameter had to be increased to 0.9, since some divergences were found.

3.2.4 Comparing the prior specifications with simulated data

Now we are going to compare the three prior specification methods. For each of the following three situations, we are going to simulate 1000 datasets from the binomial likelihood with $n_{\gamma_s}, n_{\gamma_e} \sim \text{Poisson}(50)$, $\gamma_s \sim \text{Beta}(100, 0.15/0.85 \cdot 100)$ to ensure $\mathbb{E}[\gamma_s] = 0.85$, and $\gamma_e \sim \text{Beta}(100, 0.2/0.8 \cdot 100)$ to ensure $\mathbb{E}[\gamma_e] = 0.8$. The three situations are:

a) only vague information is available;

b) strong beliefs about the means and no information about correlation;

c) strong beliefs about the means and the correlation.

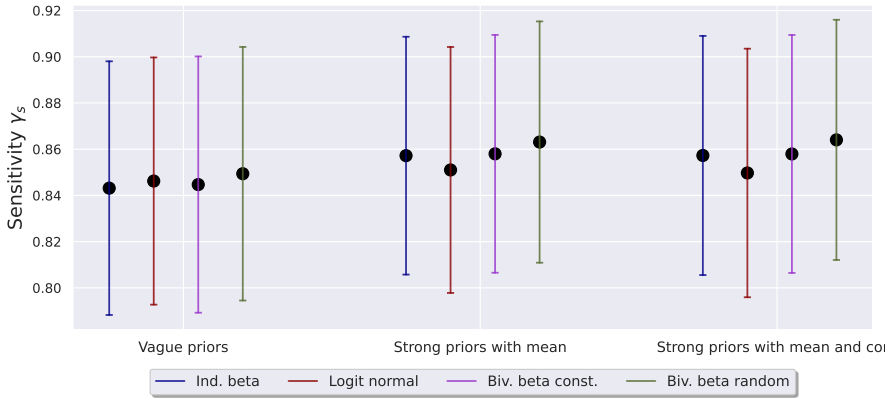
For each situation and each dataset, it was drawn 2000 samples from the posterior distribution and the HDI 75% interval and posterior mean were calculated. The coverage column counts the percentage of the times that the true values lied in the interval, while the MSE column calculates the mean squared error of the posterior mean with respect to the true value. We notice that the fourth prior approach had a little number of effective samples when compared to the other methods, and there is no big different among the approaches. The logit normal prior is worst when strong information is given. This may be related to the difficulty to convert information from the probability space to the log odds space. The estimation error decreased when information about the means and correlation is given. Figure 11 shows that the credible intervals change very little for each different approach and even for each quantity of information, which tells that the data is driving the posterior.

Table 3 – Comparing prior specification approaches in three different situations.

Situation	Prior approach	Coverage		MSE 10^{-3}	
		Sens	**Spec**	**Sens**	**Spec**
item a)	Independent betas	73.8%	76.1%	2.531	2.843
	Logit normal	74.1%	74.5%	2.405	2.811
	Biv. beta constant α	75.6%	75.5%	2.388	2.625
	Biv. beta random α	74.9%	74.4%	2.264	2.546
item b)	Independent betas	74.1%	73.6%	2.009	2.363
	Logit normal	69.9%	71.2%	2.300	2.797
	Biv. beta constant α	75.2%	75%	1.952	2.316
	Biv. beta random α	74.7%	74.8%	2.167	2.454
item c)	Independent betas	74.3%	74.2%	2.007	2.365
	Logit normal	68.4%	71.5%	2.303	2.804
	Biv. beta constant α	74.3%	74.9%	1.989	2.364
	Biv. beta random α	74.5%	75.5%	2.229	2.504

Source: Prepared by the author (2021). Biv. means bivariate and Hits is the percentage of times that the estimated HDI 75% included the true value.

Figure 11 – The average posterior mean estimate and average HDI 75% intervals for each prior strategy and level of information for sensitivity.



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs.

By the above analysis, we choose the independent betas approach given it reduces the computational burden.

3.3 Imperfect tests

A slight modification of model (3.1) is to consider the imperfection of the test measured through specificity and sensitivity, remembering the relation of these quantities

to the apparent prevalence through equation (2.4). Hence, the model can be written as

$$\begin{aligned} Y_i \mid p_i &\sim \text{Bernoulli}(p_i) \\ p_i &= \gamma_s\theta_i + (1-\gamma_e)(1-\theta_i), \\ g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T\beta, \\ \beta &\sim \text{Normal}(\mu_\beta, \Sigma_\beta), \\ \theta &\sim \text{Beta}(a^p, b^p), \end{aligned} \tag{3.3}$$

with priors on (γ_e, γ_s) as studied in the previous section. It is important to highlight that we suppose prior the data that θ is independent of γ_e and γ_s, which is not necessarily true as pointed out by Leeflang et al. (2013), who concluded that specificity tends to be lower when prevalence is higher. Model (3.3) is an extension of the one presented by Gelman and Carpenter (2020) and studied by McInturff et al. (2004).

3.3.1 Identifiability

If the regressors $\boldsymbol{x}_i$ are not present in model (3.3), it is not identifiable with respect to its likelihood as pointed out by Gelman and Carpenter (2020, p. 1271). Intuitively, the problem happens because Y_i brings information about p_i which is subdivided in three parameters: θ, γ_s and γ_e. Regarding Definition 3.1.1 and dropping the index i, take $\theta = 0.1, \gamma_e = 0.9$ and $\gamma_s = 0.6$. Then,

$$p = 1 - \gamma_e + \frac{\gamma_s + \gamma_e - 1}{1 + e^{-g(\theta)}} = 0.15.$$

With $\gamma_e = 0.9$, $\gamma_s = 0.2$ and $\theta = 0.5$, the value of p is also 0.15, which implies that two different combinations of the parameters generate the same probability function for Y. As a consequence, the model is non-identifiable. Including the regressors, the calculations are harder. Suppose that $g(\theta)$ is increased by a real a. The effect of a on p_i is through $g^{-1}(g(\theta) + a + \boldsymbol{x}_i^T\beta)$, which depends on $\boldsymbol{x}_i$. Because of that, sensitivity and specificity cannot generally offset this difference, and identifiability cannot be proved or disproved.

Nevertheless, there are some tractable cases. For instance, if $\boldsymbol{x}_i = x_i$ is a binary variable, with the same reasoning, it can be shown that the model is non-identifiable. Moreover the problems concerning the covariates $\boldsymbol{X}$ appear here in the same manner. To avoid any identifiability problem, information should be added by the prior distribution, specially through γ_s and γ_s.

Below we present a practical situation where identifiability problems appear. We simulate data from the model with $\gamma_s = 0.8, \gamma_e = 0.85$ and $\theta = 0.1$. Moreover $\beta \in \mathbb{R}^5$ and $\boldsymbol{X} \in \mathbb{R}^{200\times 5}$ are chosen arbitrarily, the regressors being drawn from a normal distribution. For the estimation process, uniform prior for θ, γ_s and γ_e, and a normal prior with mean 0 and standard deviation 1 for each β_i. After 4000 iterations for warmup and 4000 for sampling, the results are summarized in Table 4 and Figure 12. Notice that the bulk

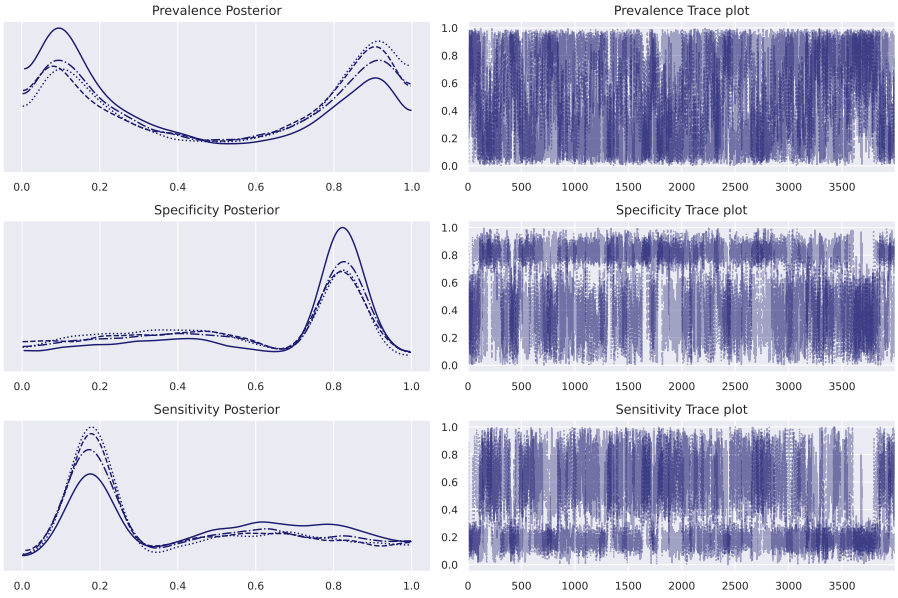
Table 4 – Results from HMC algorithm for the practical identifiability analysis in model (3.3).

	mean	**sd**	**mcse mean**	**mcse sd**	**ess bulk**	**ess tail**	$\hat{R}$
θ	0.500	0.340	0.022	0.016	291.0	3282.0	1.02
γ_e	0.585	0.277	0.019	0.015	231.0	2453.0	1.02
γ_s	0.415	0.279	0.020	0.014	241.0	2186.0	1.02

Source: Prepared by the author (2021) as a result of Stan diagnostics output. The meaning of the columns is: mean is the posterior mean; sd is the posterior standard deviation; mcse mean is the mean Markov Chain Standard Error; mcse sd is the standard deviation Markov Chain Standard error; ess bulk and ess tail are the Bulk and Tail effective sample sizes.

effective sample size is very small. The posterior mean are very bad estimates for the true values. The high density set is the union of two intervals for the prevalence, mich makes little sense in the real life.

Figure 12 – Posterior distribution and trace plot of Prevalence, Specificity and Sensitivity for model (3.3) with vague priors.



Source: Prepared by the author (2021) with output of Stan.

3.3.2 Simulated data

As an initial check for model (3.3), we use it to generate the data to verify if the estimation process is sufficiently reliable. The experiment is like the one explained in Section 3.1.1, but with sensitivity and specificity. We do not recommend vague priors on γ_s and γ_e because of the identifiability problem as mentioned above. We compare the estimates from model (3.1) in this context. Table 5 summarizes the experiments. We use a fixed $\beta = [-0.1, 2.5, 1.4, -1.8, 0.3]$ with two binary regressors and three continuous drawn from the normal distribution.

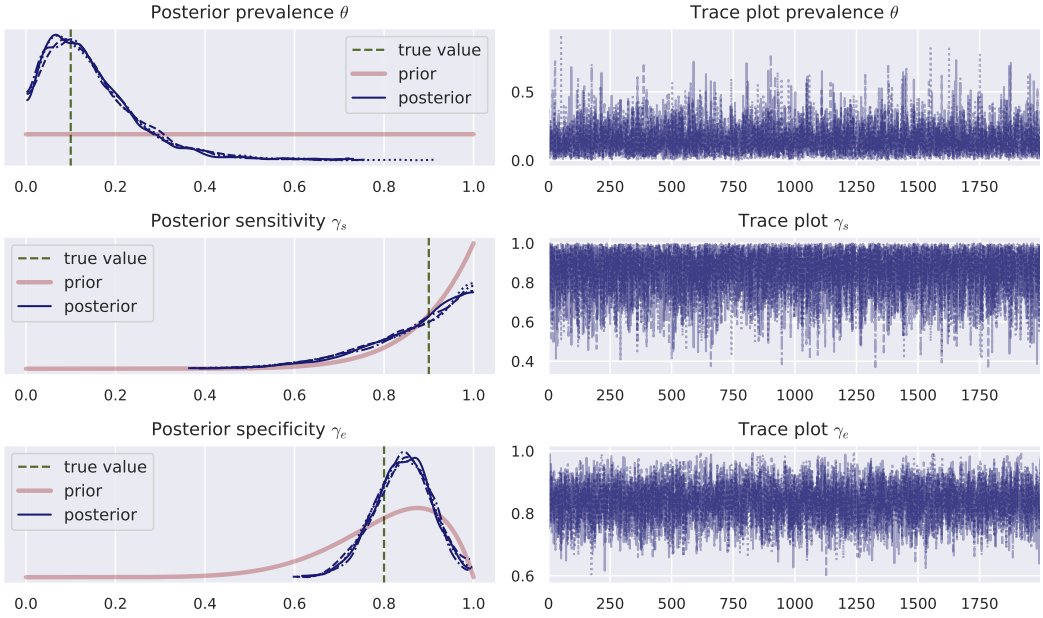
Table 5 – Experiment settings for the simulation of model (3.3).

Experiment	n	θ	γ_s	γ_e
1	100	0.1	0.9	0.8
2	100	0.02	0.85	0.85
3	2000	0.01	0.85	0.85
4	2000	0.1	0.6	0.95
5	2000	0.1	0.95	0.6

Source: Prepared by the author (2021). We denote n for number of samples.

For the first experiment, we placed vague priors on the prevalence and the effects and informative priors for the sensitivity and specificity. The algorithm took around 1.87s to perform 4000 iterations (2000 for a warm-up and 2000 for sampling). All the basic diagnostics from HMC were good. Figure 13 summarizes the posterior distribution. We also applied the first model in this dataset. For this model, the posterior mean of the prevalence was 0.148 (HDI 94% 0-0.326), while for the perfect test model, it was 0.238 (HDI 94% 0.096-0.375), a biased estimate. Gelman and Carpenter (2020, p. 1271) conclude in its application that "uncertainty in the population prevalence is in large part driven by uncertainty in the specificity." However, this effect is not so clear in this model. Figure 14 presents the resultant scatter plot of the posterior simulations. Here we observe that all parameters drive prevalence uncertainty.

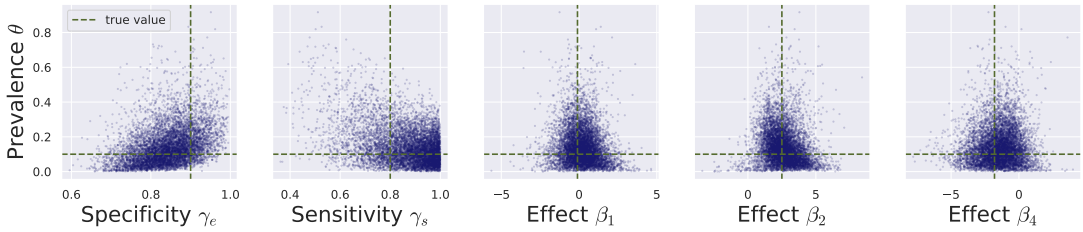
Figure 13 – Posterior distribution and trace plot for the first experiment of model (3.4)



Source: Prepared by the author (2021) from the Stan sampling result. The green line marks the true value for the simulation, while the red line represent the density of the prior distribution. Each blue line is a posterior distribution sampled from four different chains.

To verify frequentist properties, with the same specifications, we simulated 1000 datasets varying the test result Y and calculated the 75% credible interval. For each

Figure 14 – Scatter plot of the posterior simulations of prevalence, specificity, sensitivity and effects of model (3.4)

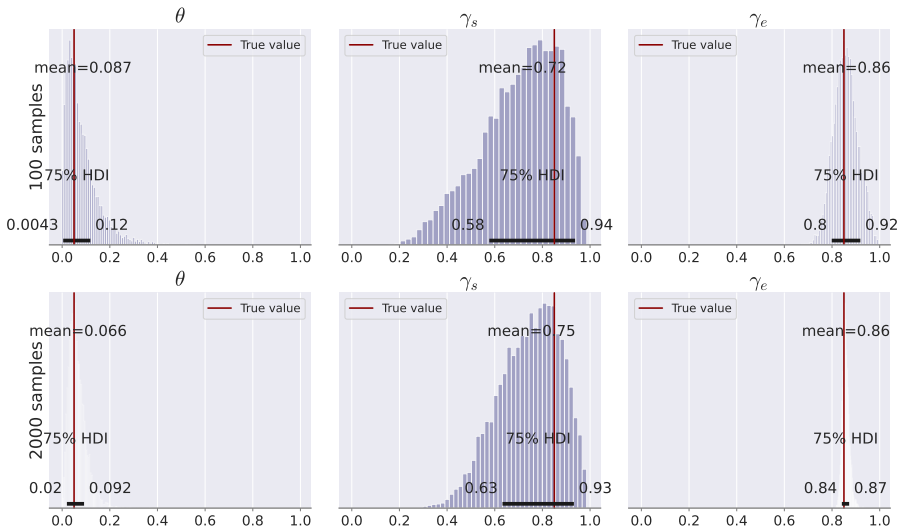


Source: Prepared by the author (2021) from the Stan sampling result.

experiment, we verified if the true parameter was included in the corresponding interval. This happened in 91.5%, 99.8%, 83%, 86.1%, 72.5%, 80.1%, 80.4%, and 70% of the times, respectively for $\theta, \gamma_s, \gamma_e, \beta_1, \ldots, \beta_5$.

The second experiment considers the case where the number of samples is not so high, but prevalece is low. It contrasts with the third experiment where many more samples are obtained. Figure 15 presents these differences. Notice that the uncertainty was decreased for all parameters, but specificity decreased the most. Even with a small quantity of data points, the model had a good performance. When the number of points increase, the credible intervals get narrower.

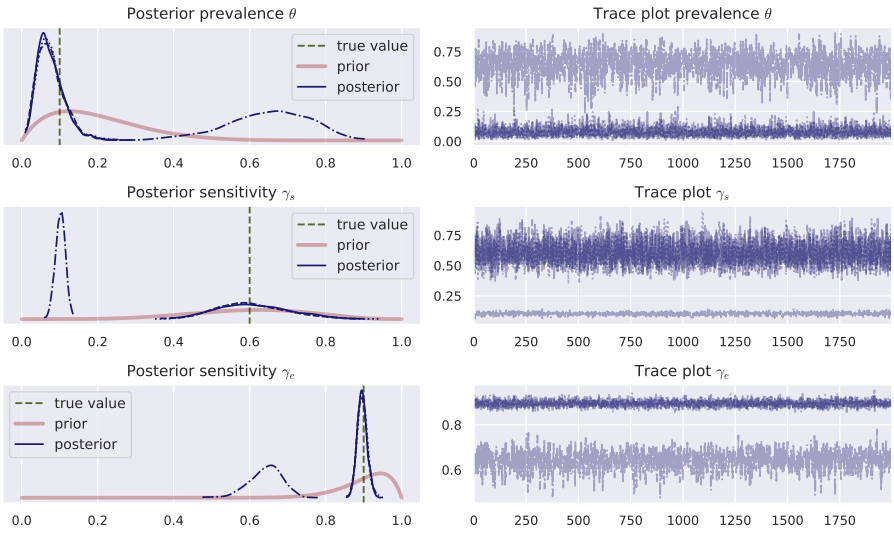
Figure 15 – Posterior distribution for θ, γ_s and γ_e from model (3.4) with experiments 2 and 3 settings



Source: Prepared by the author (2021) from the Stan sampling result. The red line represents the true value for each parameter.

The fourth and fifth experiments compare two opposite situations. The former sets a low sensitivity and a high specificity, while the latter has high sensitivity and low specificity. These examples are convenient for detecting the problem with identifiability in high dimensional data. Specifying the hyperparameters with $\alpha_p = 2$, $\beta_p = 8$, $\alpha_s = 6$,

Figure 16 – Posterior distribution and trace plot for the fourth experiment of model (3.4)



Source: Prepared by the author (2021) from the Stan sampling result. The green line marks the true value for the simulation, while the red line represent the density of the prior distribution. Each blue line is a posterior distribution sampled from four different chains.

$\beta_s = 4$, $\alpha_e = 18$, $\beta_e = 2$, with 2000 iterations for warmup the resulting posterior and trace is given by Figure 16. Notice that one of the chains was very far from the true value yielding a very high $\hat{R}$. Although the result seems awkward, it makes sense under the identifiability problem, since the chain produces a very similar probability of positive test. A stronger prior for γ_s can handle this effect, which means that the prior distribution depends on the size of the sample. A similar behavior happens for the fifth experiment.

3.4 Imperfect tests and respondent-driven sampling

After understanding the problem when not considering the specificity and the sensitivity of the diagnostic test for the estimation of θ, we focus on the sampling strategy studied in Section 2.2. One problem with RDS is that we cannot make probability statements without making assumptions about the sampling process. Since the participants recruit their peers, the sampled individuals depend on the recruiters and whom they recruited. In this section, we propose a model for the network dependence of RDS extending Bastos, Pinho, et al. (2012).

For now, the recruitment graph (see Definition 2.2.1) has no uncertainty incorporated, and we included it as a random effect on the model through a CAR model (see Section 2.3.2) in the Gaussian case. Besag (1974) introduced CAR for spatial effects, but they fit in this situation since, by adjacent sites, we understand recruitment. We remark that for RDS, we partially observe the corresponding map. If the entire map was available, we could interpret it as interaction or friendship depending on the population.

Following the notation of Section 2.2.3 and Section 2.3.2, we denote A for the adjacency matrix, where $[A]_{ij} = a_{ij} = 1$ if, and only if, i connects to j and 0 otherwise. We also denote $a_{i+} = \sum_j a_{ij}$. Besides the parameters from model 3.3, we use τ for the spatial precision parameter and ρ for controlling the dependence between neighbors. Hence, we specify the model as follows:

$$\begin{aligned}
Y_i \mid p_i &\sim \text{Bernoulli}(p_i) \\
p_i &= \gamma_s\theta_i + (1-\gamma_e)(1-\theta_i), \\
g(\theta_i) &= g(\theta) + \boldsymbol{x}_i^T\beta + \omega_i, \\
\beta &\sim \text{Normal}(\mu_\beta, \Sigma_\beta), \\
\omega_i \mid \omega_j, j \neq i &\sim \text{Normal}\left(\rho\sum_j a_{ij}\omega_j/a_{i+}, \tau^{-1}/a_{i+}\right), i = 1, \ldots, n, \\
\theta &\sim \text{Beta}(a^p, b^p) \\
\gamma_s &\sim \text{Beta}(a^s, b^s) \\
\gamma_e &\sim \text{Beta}(a^e, b^e) \\
\tau &\sim \text{Gamma}(a^\tau, b^\tau) \\
\rho &\sim \text{Unif}(0, \lambda_{\max}^{-1}).
\end{aligned} \tag{3.4}$$

We remind that $\omega \sim \text{Normal}\left(0, [\tau(D-\rho A)]^{-1}\right)$ as discussed in Section 2.3.2, such that $D_{ii} = a_{i+}$.

The prior specification of ρ is highly debated in the literature. Since we established that the lower bound is 0, we are saying that there is a positive correlation between the respondents, which is an usual assumption, but must be verified for each real application. As we noticed in Section 2.3.2, this correlation is strong only if ρ is close to $\lambda_{\max}^{-1}$, therefore the uniform distribution contrasts with this empirical knowledge. Banerjee, Carlin, and Gelfand (2003, p. 177) suggests the use of beta distribution for ρ with a large mean, "but this is controversial since there will typically be little true prior information available regarding the magnitude of α" (α is the parameter ρ in our notation. BANERJEE; CARLIN; GELFAND, 2003, p. 177). Lee (2011, p. 81) uses a discrete uniform distribution over $\{0, 0.05, 0.1, \ldots, 0.9, 0.95\}$.

The prior distribution on τ is also subject to discussion. Since it is a precision parameter, Simpson et al. (2017, p. 9, Theorem 1) prove that if the prior has finite mean, it *overfits*, which intuitively means that the prior puts not enough mass at the base model, in this case the model without spatial correlation. They calculate a penalized complexity prior for τ as the type-2 Gumbel distribution with density

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp(-\lambda\tau^{-1/2}), \tau > 0, \tag{3.5}$$

where $\lambda > 0$ determines the magnitude of the penalty. Simpson et al. (2017, p. 9) suggest to specify U and α so that $\Pr(1/\sqrt{\tau} > U) = \alpha \implies \lambda = -\log(\alpha)/U$. This distribution

can be rewritten in terms of the standard deviation $\sigma = 1/\sqrt{\tau}$ by the Change of Variables formula as follows

$$\pi(\sigma) = \frac{\lambda}{2}\sigma^3 \exp(-\lambda\sigma) \cdot 2\sigma^{-3} = \lambda \exp(-\lambda\sigma). \tag{3.6}$$

Lee (2013, p. 5) uses $\sigma^2 \sim \text{Unif}(0, M_\sigma)$ with $M_\sigma = 1000$ as default based on Gelman (2006) since "it is difficult to choose the hyperparameters so that it is non-informative for very small values of" (LEE, 2013, p. 4) referring to specification of a non-informative inverse-gamma distribution for σ^2. With that distribution, we have, by the Change of Variables formula,

$$\pi(\tau) = \frac{1}{M_\sigma}\tau^{-2}\mathbb{1}_{\{\tau > 1/M_\sigma\}}.$$

Comparison between parametrization of σ and τ showed that they are similar in sight of time of execution, energy and divergences, among others diagnostics. However, the mean estimate of σ is more controlled, while the median is very similar for both.

3.4.1 Identifiability

This model inherits all the problems with identifiability from the previous ones. Xie and Carlin (2006, p. 3470) discusses when two parameters depending on the individual are summed, one being a CAR component, while the other capturing heterogeneity among the regions. When $\rho = 1$, the prior on ω is improper and property of the posterior must be analysed in each case. Because of that, to identify the parameter θ, an additional constraint is necessary, such as

$$\sum_{i=1}^{n} \omega_i = 0. \tag{3.7}$$

Relation (3.7) is called in optimization as *hard constraint* since the solution must strictly satisfy it. In probability theory, it would mean to give a point mass distribution for the sum. In Stan, a common alternative is to put a *soft constraint* such as

$$\sum_{i=1}^{n} \omega_i \sim \text{Normal}(0, 0.0001/n), \tag{3.8}$$

that serves as a penalty term.

3.4.2 Stan implementation

Implementing this model was a long process, including several failures and some successes. This subsection summarizes how the errors happened, how we detected through diagnoses such as divergences and energy, and how we fixed them. For the following comparisons, the contact graphs are generated following an Erdős–Rényi model (ERDOS; RÉNYI, et al., 1960), which is the simplest way to draw a random network. Fixing a number of individuals N, a number of samples $n < N$ and a probability of contact P,

each node is connected by a link with probability P. The expected number of edges in this graph is $PN(N-1)/2$. To perform a RDS sampling, we follow Baraff, McCormick, and Raftery (2016, p. 14670):

a) 10 seeds are selected at random with probability proportional to their degree;

b) each participant receives three coupons and has probabilities of 1/3, 1/6, 1/6 and 1/3 to recruit, respectively, 0, 1, 2 or 3 individuals from their contacts;

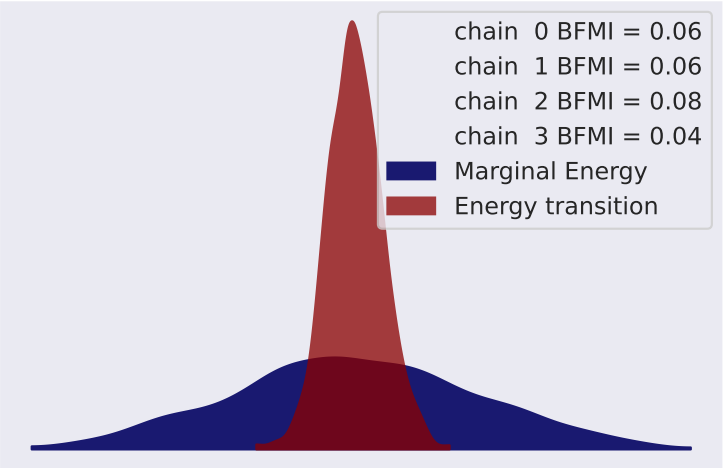
c) the last wave of recruitments happens when the number of samples is 150.

More details of the simulation process is given in Subsection 3.4.3.

Raw implementation

The first Stan implementation is exactly as presented in equation (3.4). Coding as we model is an advantage of Stan since it improves readability of the code. However this implementation suffers from some problems concerning the geometry of the parameter space. For this implementation we treat ρ as a fixed parameter. Performing 1000 iterations of the algorithm took around six minutes. The resulting inferences were not so bad despite the small number of iterations. Figure 17 presents a comparative plot between the marginal energy distribution and the distribution of the transition energy. Since they are very different, it means that the target distribution has heavy tails, which is challenging for the sampling. In this case, "the stochastic exploration between level sets will become so slow that after any finite number of transitions the exploration of the Markov chain will be incomplete." (BETANCOURT, 2017, p. 44).

Figure 17 – Energy plot raw Stan implementation of model (3.4)



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs. The blue curve is the marginal distribution of the energy, while the red is the distribution of the energy transition.

Vectorization of the variables

According to the Stan documentation (STAN DEVELOPMENT TEAM, 2021), the vectorized form of probability functions and operations such as matrix products are more efficient than loops and transformed parameters. Because of that, we apply this suggestion to the implementation. Notice that it does not solve the problem with the heavy tails of the target distribution, then the time to make 1000 iterations was around five minutes. Although the vectorization did not prove to be better for the model in particular, we keep on using it given the general recommendations of the documentation.

Non-centered parameterization

Parametrization of hierarchical models is a longstanding issue. Figure 18 shows how the posterior samples behavior in a hierarchical model. This effect is known as funnel since there is a high density region with low volume and a low density region with high volume (BETANCOURT; GIROLAMI, 2015, p. 1). The posterior correlation observed can be mitigated through a non-centered parameterization. Taking ρ fixed, we reparametrize ω as follows:

$$\begin{aligned} \omega_i^{\text{raw}} &\sim \text{Normal}(0, 1) \\ \omega &= \tau^{-1/2} V_\omega \cdot \omega^{\text{raw}}, \end{aligned}$$

where $V_\omega V_\omega^* = (D - \rho A)^{-1}$ is the Cholesky decomposition. We do the same thing for the effects β:

$$\begin{aligned} \beta_i^{\text{raw}} &\sim \text{Normal}(0, 1) \\ \beta &= \mu_\beta + V_\beta \cdot \beta^{\text{raw}}, \end{aligned}$$

where $V_\beta V_\beta^* = \Sigma_\beta$. With that little exchange, the time to run the four chains and 1000 iterations took around one minute, much better than previously. Figure 19 shows how this improves the results on the energy plot. However, this implementation is not efficient yet, since we are only taking a few iterations.

Efficient implementation

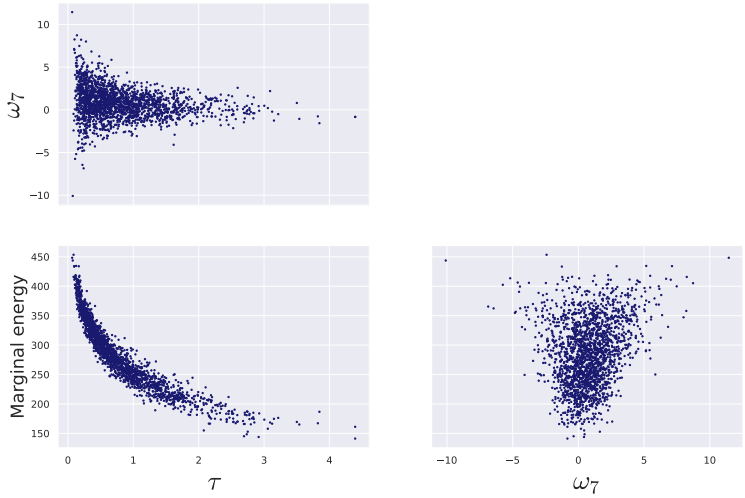
RDS connection matrix is essentially sparse since the number of edges is at most four times the number of individuals (one recruiter and three recruited). Because of that, we follow the efficient implementations of sparse CAR models from Donegan (2020) and Max Joseph (2021). The probability density of a multivariate normal distribution with zero mean is

$$f_\omega(\omega \mid \tau, \rho) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma^{-1})}} \exp\left\{-\frac{1}{2}\omega^T \Sigma^{-1} \omega\right\}, \tag{3.9}$$

where $\Sigma^{-1} = \tau(D - \rho A)$. Jin, Carlin, and Banerjee (2005, p. 955) proves the following relation

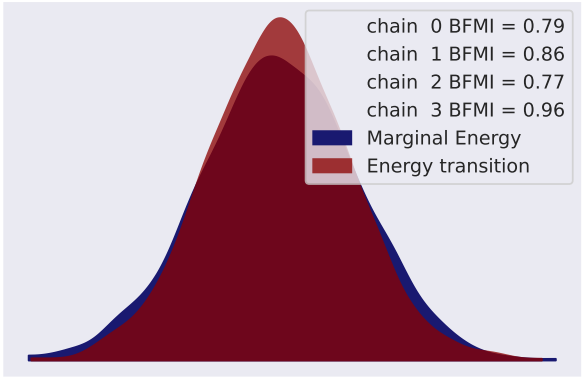
$$\det(\tau(D - \rho A)) = \tau^n \det(D - \rho A) \propto \tau^n \prod_{i=1}^{n} (1 - \rho\lambda_i),$$

Figure 18 – Scatter plot of posterior samples of few parameter from the inefficient Stan implementation of model (3.4) to visualize funnel effects



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs.

Figure 19 – Energy plot non-centered Stan implementation of model (3.4)



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs. The blue curve is the marginal distribution of the energy, while the red is the distribution of the energy transition.

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $D^{-1/2}AD^{-1/2}$. Notice that $\{\lambda_i\}_{i=1}^n$ are calculated before of the HMC algorithm. We also have that $\omega^T(D - \rho A)\omega = \omega^T D\omega - \rho\omega^T A\omega$. Since D is a diagonal matrix, $\omega^T D\omega = \sum_{i=1}^n D_i i\omega_i^2$ which can be easily calculated. Moreover, by the sparsity of A, $\omega^T A\omega$ can be performed with sparse matrix operations.

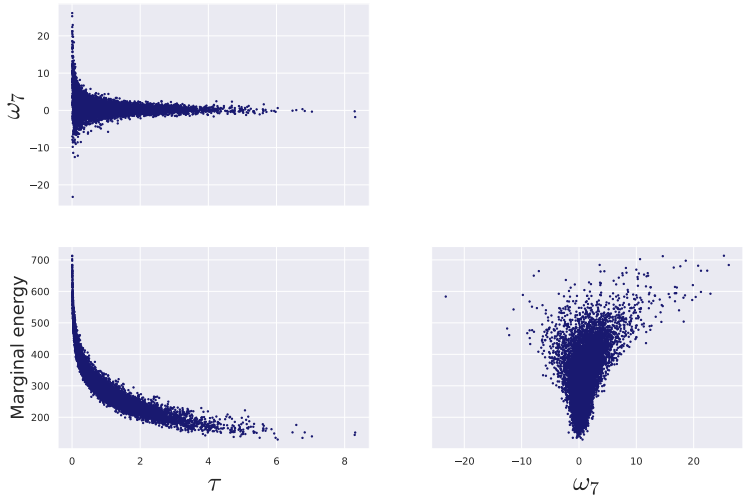
Given that Stan receives the log density and constants are not necessary, (3.9) can be simplified to

$$\log f_\omega(\omega \mid \tau, \rho) = \frac{n}{2}\log(\tau) + \frac{1}{2}\sum_{i=1}^{n}(1 - \rho\lambda_i) - \frac{\tau}{2}\omega^T(D - \rho A)\omega.$$

The time to perform 1000 iterations was around 8s, which is a pretty good reduction compared to the previous implementations. Despite that, the efficient implementation has the disadvantage of not allowing to rescale the multivariate normal distribution, given

that matrix V_ω does not have a sparse representation. Therefore, the energy plot is very similar to Figure 17. The scatter plot is also very similar to Figure 18, but since the we can draw more samples without worrying about time of execution, the funnel effects is more evident as presented in Figure 20.

Figure 20 – Scatter plot of posterior samples of few parameter from the efficient Stan implementation of model (3.4) to visualize funnel effects



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs.

Although the Cholesky decomposition of $D - \rho A$ is not sparse, we can rescale the parameter τ from the sparse representation of CAR model. Let

$$\omega_i^{\text{scaled}} \sim \text{Normal}(0, 1)$$
$$\omega = 1/\sqrt{\tau} \cdot \omega^{\text{scaled}}.$$

The efficient representation of the density of ω^{scaled} is

$$\log f_\omega(\omega^{\text{scaled}} \mid \rho) = \frac{1}{2}\sum_{i=1}^{n}(1 - \rho\lambda_i) - \frac{1}{2}(\omega^{\text{scaled}})^T(D - \rho A)\omega^{\text{scaled}},$$

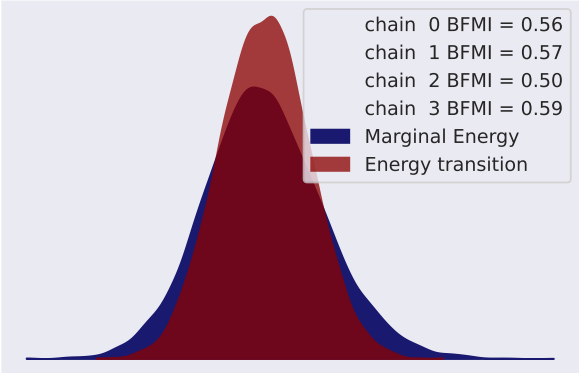
which is equivalent to $\log f_\omega(\omega \mid \tau = 1, \rho)$. With this simple modification, the energy plot improves as Figure 21 demonstrates. For comparison purposes, Figure 22 highlights how the funnel effect decreased after this change.

3.4.3 Simulated data

Here we present how the inferences change from different specifications of the parameters. First of all, we have to generate a graph to represent the hidden population. We use four different approaches:

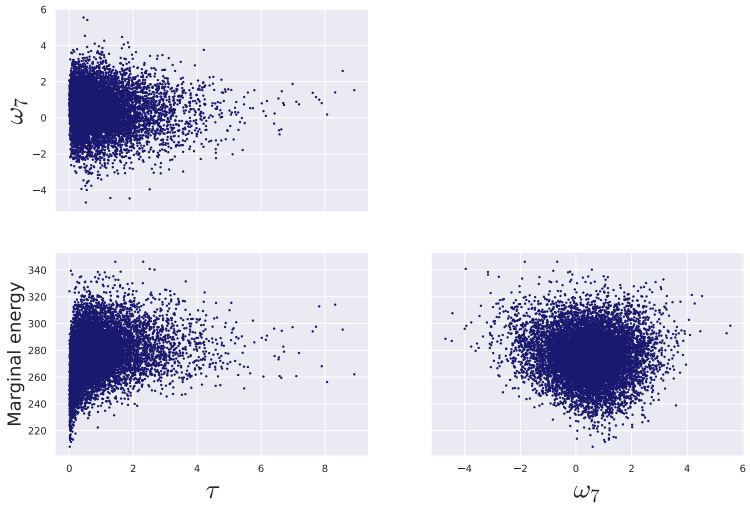
a) Erdős–Rényi (ER1) (ERDOS; RÉNYI, et al., 1960): let G be a graph with N nodes. For each pair of nodes, we connect them by a link with probability P.

Figure 21 – Energy plot efficient and scaled Stan implementation of model (3.4)



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs. The blue curve is the marginal distribution of the energy, while the red is the distribution of the energy transition.

Figure 22 – Scatter plot of posterior samples of few parameter from the efficient and scaled Stan implementation of model (3.4)



Source: Prepared by the author (2021) and based on Stan and ArviZ outputs.

The expected number of edges is $PN(N-1)/2$ and the average number of neighbors of a node is $P(N-1)$.

b) Erdős–Rényi with preferential attachment by a binary variable (ER2): from a graph G with N nodes, each node i receives a binary variable $Z_i \in \{0, 1\}$ with $\Pr(Z_i = 1) = 0.3$ creating two subpopulations. Two individuals within the same subpopulation are connected with probability P_1, while individuals with different values of Z connect with probability $P_2 < P_1$. Each subpopulation, conditioned on the binary variables, is a Erdős–Rényi graph.

c) Barabási-Albert (BA1) (BARABÁSI; ALBERT, 1999): let m_0 be a initial number of nodes. Each time, a node is added to the graph and connects to $m \leq m_0$ other nodes with probability P^i proportional to the degree of the nodes

without replacement. If k_i is the degree of node $i \in \{1, \ldots, m_0\}$, so $P^i \propto k_i$. The algorithm ends when the number of nodes is N. This property is called *preferential attachment* since nodes with higher degree are preferred for the new connections. The average degree is $2m - m/N - m^2/N$.

d) Barabási-Albert with preferential attachment by a binary variable (BA2): the only change in BA1 is that each node connects to other within the same population with probability P and to any node disregarding variable Z with probability $1 - P$.

With a graph G, we need to simulate a respondent-driven sampling. For that, we use two different approaches:

a) standard: this formulation was used by Baraff, McCormick, and Raftery (2016, p. 14670), as described in Section 3.4.2.

b) preferential: the variable Z influences the recruitment choice. For each coupon the individual possess, there is a probability R of choosing someone within the same subpopulation and $1 - R$ of choosing disregarding the variable Z.

After specifying the network and the sampling simulation, we obtain the matrices A and D. We set $\beta = [-0.1, 2.5, 1.4, -1.8, 0.3]$ with two binary regressors and three continuous drawn form the normal distribution, $\gamma_s = 0.9$, $\gamma_e = 0.85$, and $\theta = 0.1$. The second binary regressor is used for creating the subpopulation whenever necessary. Table 6 summarizes the experiments. All the experiments set $N = 10000$ and choose the other parameters to obtain an average number of neighbors of approximately 50. All experiments have 500 samples. The choice of $\rho = 0.9995$ reflects the observation of Table 1, since with $\rho = 0.95$ and $n = 500$, the Pearson correlation is too small. We can verify this empirically with a contingency table comparing the test result of the recruiter and the recruit.

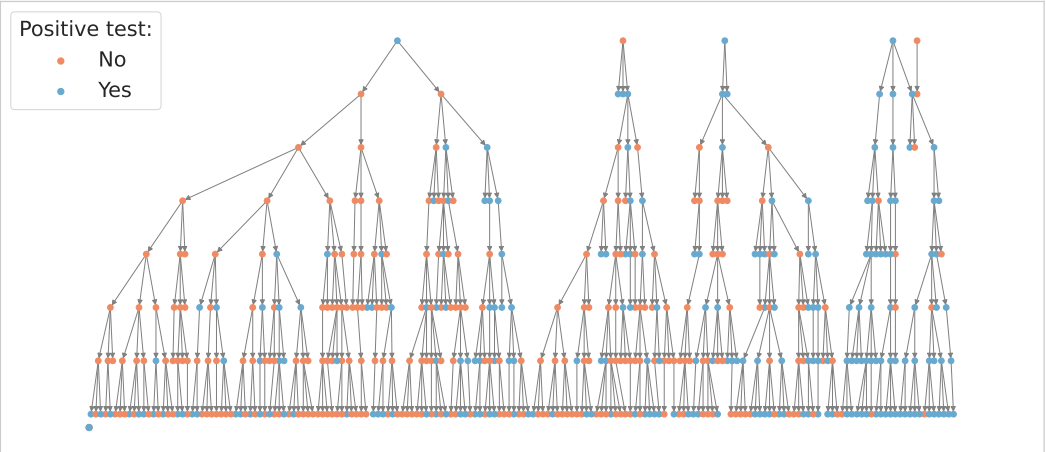
Table 6 – Experiment settings for the simulation of model (3.4).

Experiment	Graph	RDS	ρ	τ
1	(ER1, $P = 0.005$)	Standard	0.9995	1
2	(ER1, $P = 0.005$)	Standard	0.9995	10000
3	(ER1, $P = 0.005$)	Standard	0.5	10
4	(ER2, $P_1 = 0.008, P_2 = 0.002$)	Preferential($R = 0.5$)	0.9999	10
5	(BA1, $m = 25$)	Standard	0.9999	5
6	(BA2, $P = 0.5, m = 25$)	Preferential($R = 0.5$)	0.99999	5

Source: Prepared by the author (2021).

For the first experiment, Figure 23 shows the simulated RDS structure. We disregarded unsuccessful seeds, so, from the ten starting individuals, only five continued the recruitments. Running a Chi-square test with significant level 5% in the contingency table from Table 7 indicates the existence of a relation between the variables, as we wanted to simulate.

Figure 23 – Simulated RDS structure in an Erdős–Rényi graph.



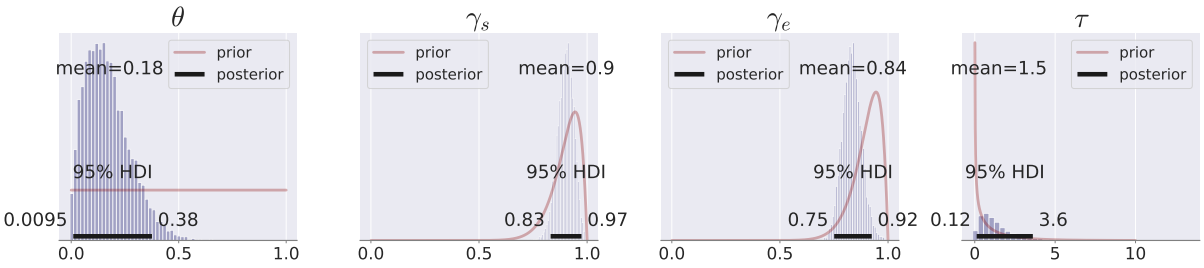
Source: Prepared by the author (2021) and based on NetworkX result. The colors indicate the result of the test, yes being positive and no being negative.

Table 7 – Contingency table of recruiter and recruited's test result from experiment 1

Recruiter's test result	Recruited's test result		Total by recruiter
	Negative	Positive	
Negative	230	103	333
Positive	83	71	154
Total by recruited	313	174	487

Source: Prepared by the author (2021).

Figure 24 – Posterior distribution of θ, γ_s, γ_e and τ for experiment 1 of model (3.4).

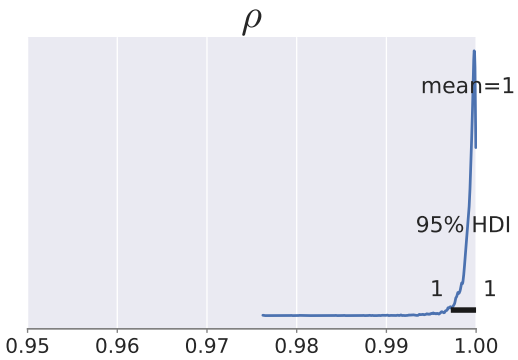


Source: Prepared by the author (2021) and based on Stan and ArviZ results.

Placing strong priors on γ_s and γ_e as usual, and weakly informative priors on the other parameters, we perform 4000 warmup and 2000 sampling iterations. The execution time was around nine minutes. Figure 24 shows the learning of the parameters from the data. An interesting thing is that γ_s moved the mode of the prior distribution closer to the true value. The bulk-ESS was less than a thousand for θ and ρ. Despite that, all the HMC diagnostics were good. Although the prior for ρ is a uniform distribution, the posterior is very tight (see Figure 25).

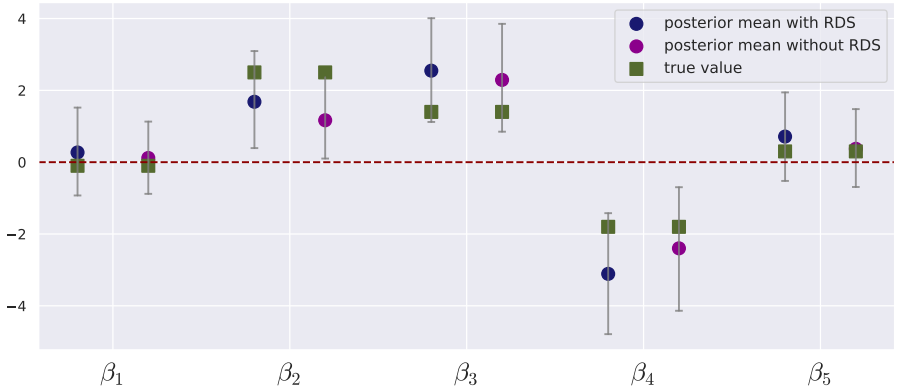
The credibility intervals for the parameters β were very large, as Figure 26 pictures.

Figure 25 – Posterior distribution of ρ for experiment 1 of model (3.4).



Source: Prepared by the author (2021) and based on Stan and ArviZ results.

Figure 26 – Comparing posterior mean and 94% credibility intervals for β in model (3.4) and model (3.3).



Source: Prepared by the author (2021) and based on Stan and ArviZ results.

Moreover, we performed a comparison between this model and the model without the RDS structure. As expected, almost all the intervals are broader when the model includes the RDS because of the uncertainty which ω adds. However, the estimates were not so different, implying that this model was not capable of proving to be better. The estimates of the prevalence were 0.105 (94% HDI of 0.0-0.38) for the model that includes RDS, while 0.13 (94% HDI of 0.008-0.27) for the other. This implies that the covariates influence the results much more on than the RDS itself regarding this model.

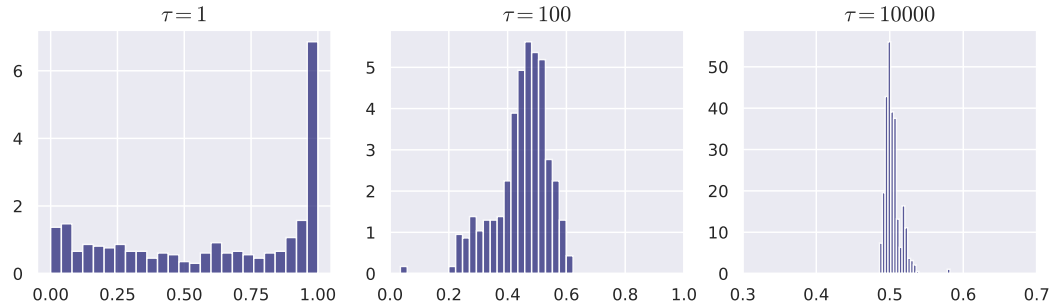
Strong priors on β decrease the posterior uncertainty, but they do not change the comparison between both models. For the second experiment, with $\tau = 10000$, time execution time was around one minute, much less than the previous one. In this experiment, the expected magnitude of ω is 100 times less than $\tau = 1$. Figure 27 presents the histogram of

$$\frac{1}{1+\exp(\omega_i)}, i = 1, \ldots, n.$$

When we include $\tau = 100$, we notice that the posterior of ρ, as in Figure 28, differs

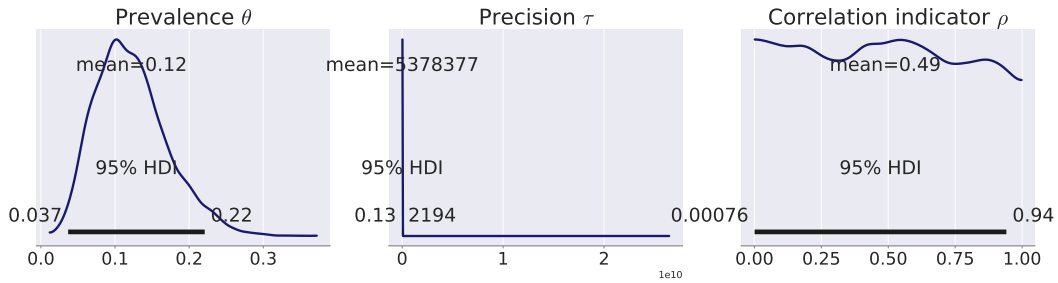
from Figure 25. It means that the correlation between recruiter and recruit in the data is not strong. Because of this, we have to set a higher ρ. We highlight that there is a non-trivial relation between ρ and τ. Figure 28 also answers what happens when ρ is small, as in the third experiment. Unfortunately, as higher as ρ gets, the harder is the geometry of the problem, then more execution time is required. In particular, with $\tau = 100$ and $\rho = 0.999995$, the chains had very problematic divergences. We also noticed that the beta prior for ρ with high mass around 1, such as specifying $\alpha_\rho = 900$ and $\beta_\rho = 1$, for instance, provokes inefficient searches, which is shown to us through the maximum tree depth. This should be investigated in future work.

Figure 27 – Comparing the transformed values of ω by the inverse logit for different magnitudes of τ



Source: Prepared by the author (2021). Note that the axis x of the third graphic was tightened because the density is very concentrated.

Figure 28 – Posterior distribution for θ, τ and ρ from model (3.4) when $\tau = 100$.

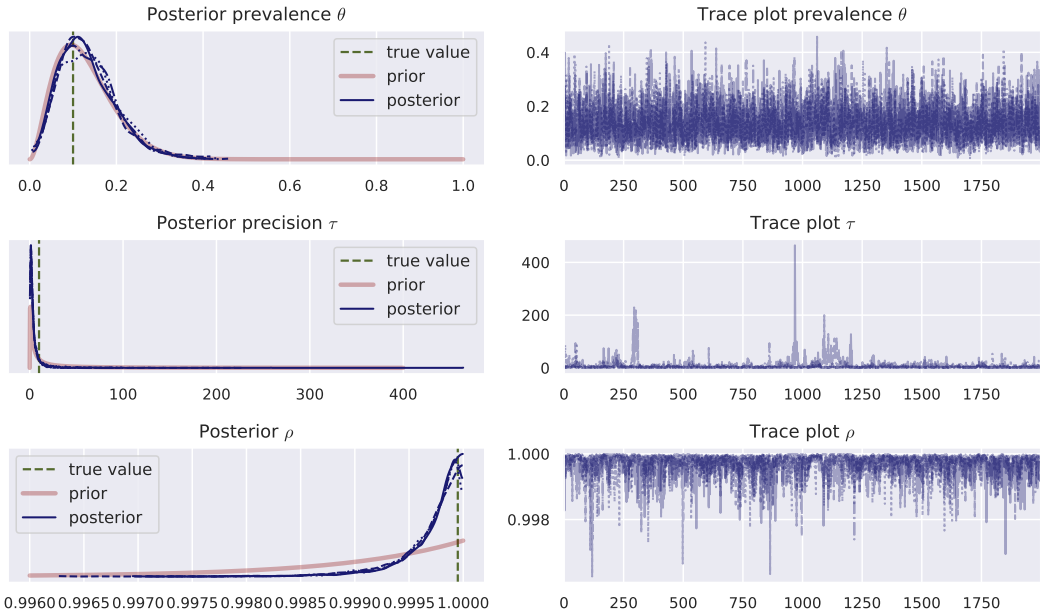


Source: Prepared by the author (2021) and based on Stan and ArviZ results.

Another important aspect is the prior distribution for τ, as discussed when we presented model (3.4). The gamma distribution gives low probability to the model without the spatial effect ω. With Gumbel prior, τ has more freedom to visit high values. Figure 29 shows how this impacts the inferences with $\lambda = \log(10)$, which is equivalent to setting $\Pr(1/\sqrt{\tau} > 1) = 0.1$. The prior distribution for θ was very strong leading to little learning from the data.

At last, we compare the last three experiments to verify if the inferences for θ are robust for different graph structures. We compare the posterior 94% HDI and the

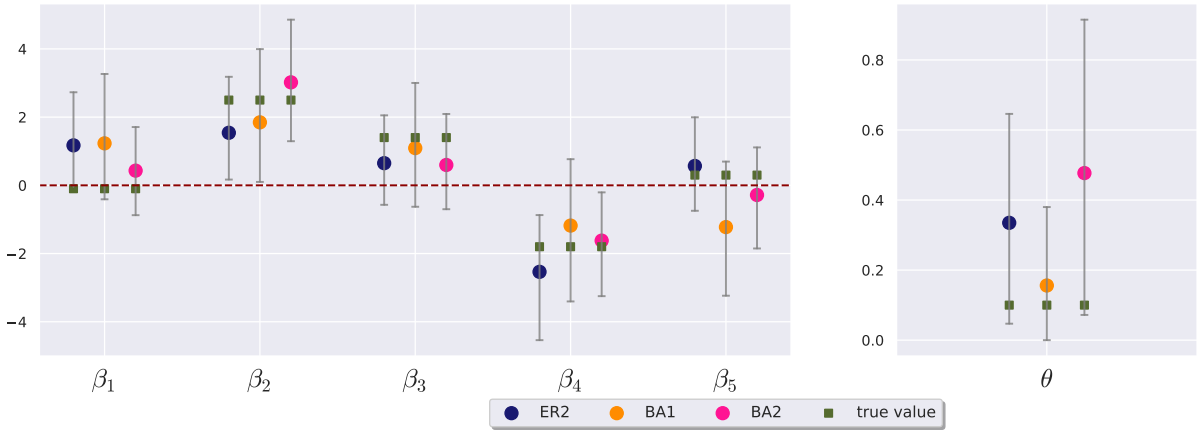
Figure 29 – Posterior distribution and trace plot for θ, τ and ρ from model (3.4)



Source: Prepared by the author (2021) and based on Stan and ArviZ results.

posterior mean for the parameter β for each different graph structure in Figure 30. Weakly informative priors on β and θ were posed. Notice that the credible intervals for β are very similar in the three structures, including β_2, the parameter related to the variable used for preferential recruitment in the models ER2 and BA2. The credible intervals are wide, as observed before. The credible intervals for θ presents something already noted in the literature (see (GILE; BEAUDRY, et al., 2018) for instance).

Figure 30 – Comparing posterior mean and 94% credibility intervals for β and θ for model (3.4) from three different graph structures.



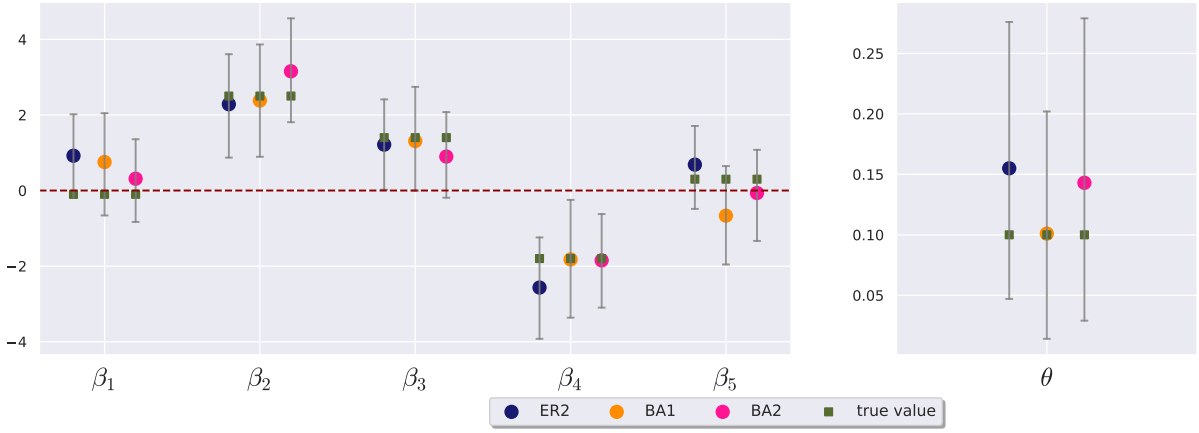
Source: Prepared by the author (2021) ans based on Stan result.

The estimates of the model not considering sampling weights and homophily generate biased estimates when RDS structure has these characteristics. This is a relevant limitation of model (3.4). However, the credible intervals included the true value, which is

an evidence of uncertainty quantification. Performing model (3.3) over the dataset with BA2 graph structure resulted in a posterior for θ of 0.51 with 94% credible interval of 0.2 to 0.84, which not included the true value of θ. This fact show that ignoring RDS structure can lead to misleading results.

We compare the prevalence estimates when strong priors are posed on θ in Figure 31. Notice that the estimates are much better for the prevalence, but the graph structure BA1 leads to inferences that are much closer to the true value in comparison to ER2 and BA2 that add homophily effects in the recruitment.

Figure 31 – Comparing posterior mean and 94% credibility intervals for β and θ for model (3.4) from three different graph structures with strong priors.



Source: Prepared by the author (2021) ans based on Stan result.

3.4.4 Including uncertainty about the recruitment graph

RDS has the constraint of being without replacement. For that reason, we do not observe all links among the sampled individuals. Considering the model developed by Crawford (2016), we can include the uncertainty regarding the recruitment graph G_R. We modify the model (3.4) as follows:

a) we suppose having the necessary data for Crawford (2016)'s model, which includes the Coupon Matrix C, the recruitment times $\boldsymbol{t}$ and the informed degrees $\boldsymbol{d}$, besides the recruitment graph G_R;

b) we sample from the posterior distribution given by equation (2.9) using a Metropolis-within-Gibbs procedure detailed in Appendix B;

c) with the sampled compatible subgraphs $\hat{G}_S^{(i)}$ and their corresponding adjacency matrices $A^{(i)}$, we have the distribution of the recruitment-induced subgraph G_S;

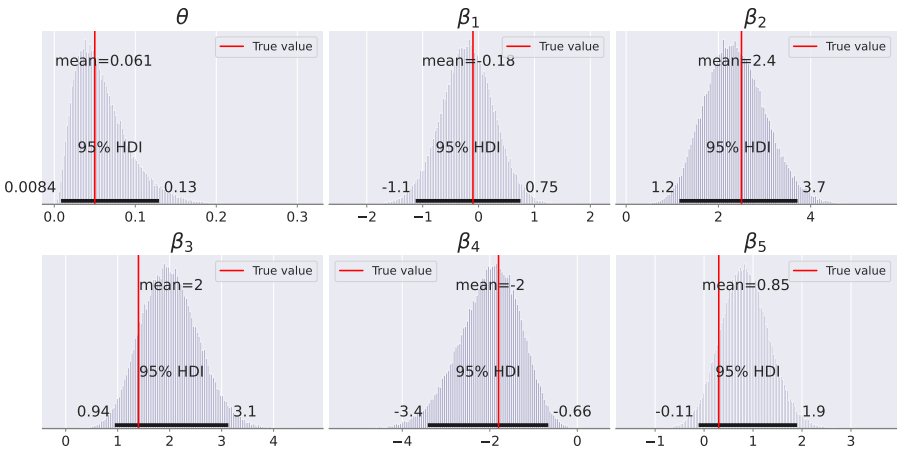
d) for each sampled $A^{(i)}$, we can obtain samples from model (3.4). The collection of all samples form a complete sampling considering the uncertainty over G_R.

There are several limitations in Crawford (2016)'s work. The recruitment are independent of the recruiter identity, the neighbors, and all the other waiting times in the model, but in real applications, this is hardly the case. It supposes that $\boldsymbol{d}$ is a fixed data without considering the uncertainty about it. Moreover, the sampled graphs are subject to inaccuracies since the uniform prior does not offer information to distinguish edge $\{i, j\}$ from $\{i, k\}$ when both do not appear in the recruitment graph. In general, there are many more wrong options to i connect than right ones. Inclusion of recruiter's characteristics can be helpful in future work.

As a validation of this method, with the same specifications of the experiment 1 of Table 6, we generate RDS data through Crawford (2016)'s model obtaining $G_R, \boldsymbol{d}, \boldsymbol{t}$ and matrix C. Moreover, we can get the recruitment-induced subgraph G_S and its adjacency matrix. We generate the results of the tests Y using model (3.4). We first sampled from the distribution $p(G_S \mid \boldsymbol{Z})$ using Metropolis-within-Gibbs described in Appendix B. We used 10000 warmup and 100 sampling iterations $\hat{G}_S$, which took around five minutes. With each graph, we performed HMC to obtain samples for the other parameters with 2200 warmup and 400 sampling iterations for two chains. This marginalize the effect of the graphs. This process took about four hours.

Informative priors were placed on θ and β, since even without considering the uncertainty of the graph, the credible intervals already are too wide. Counter intuitively, the credible intervals were not too broad. In special, the intervals were almost equal, with a difference of about 0.1 for the parameters. There are two possible reasons: a hundred graphs are a small quantity for integrating the graph uncertainty, or the estimated graph $\hat{G}_S^{(i)}$ contains more information since it predicts missing links, which individually reduces the uncertainty about the parameters. Therefore, well defined priors for the prevalence and effects can mitigate our uncertainty about the graph.

Figure 32 – Posterior distribution of θ and β in model (3.4) with uncertainty in the graph.



Source: Prepared by the author (2021) and based on Stan and ArviZ results.

3.5 Model extensions

Several characteristics of RDS were not included in the previous model, such as homophily, bottlenecks and sampling weights. This section aims to outline avenues of further work.

a) *homophily model*: ignoring homophily effects can lead to biased estimates, as we simulated in this work. Yauck et al. (2021, p. 9) defined a similar model where the spatial effect follows a SAR model and a parameter γ is included to measure homophily effects. In our model it would change the formulation of θ_i to

$$g(\theta_i) = g(\theta) + \boldsymbol{X}\beta + \gamma\frac{1}{n_i}A_i z + \omega_i,$$

where n_i is the number of connections of individual and i, A_i is the i-th row of A and z is a variable that influences the recruitment choice. The parameter γ has serious problems with identification as pointed out by Yauck et al. (2021, p. 11), which, in our case, could be dealt with a strong prior.

b) *sampling weights*: one objective of introducing the RDS-VH estimator for prevalence was to consider the inclusion probabilities of the individuals in the research. Our model does not consider these probabilities, which can contribute to wrong inferences. Bastos, Bastos, et al. (2018) used the pseudo-likelihood strategy to ponder the likelihood of each individual. For point estimate of θ, other approaches are used, such as

$$\hat{\theta} = \sum_{i=1}^{n} \frac{w_i\hat{\theta}_i}{\sum_{i=1}^{n} w_i},$$

where w_i are weights and $\hat{\theta}_i$ are point estimates of the posterior distribution of θ_i.

c) *Bottlenecks*: bottleneck effects create clusters of individuals based on their characteristics. Fortunately, hierarchical models can address this situation including an additional index j for the corresponding cluster. Then, the expression of θ_i turns into

$$g(\theta_{ij}) = g(\theta) + \boldsymbol{X}\beta + \omega_{ij},$$

where ω_{ij} can take into consideration to different effects, within cluster and between clusters.

These modifications are possible future works. There are other limitations we do not discuss, such as differential activity (GILE; BEAUDRY, et al., 2018, p. 68), limited number of sample waves, random recruitment, and small size networks (GILE; JOHNSTON; SALGANIK, 2015).

4 Data applications

In this chapter, we describe two applications to data not directly generated by our model. In fact, real applications are very difficult to obtain due to ethical concerns. Several datasets do not fit in our problem by several reasons: unavailable diagnostic tests (PERESTROIKA; PRABANDARI; WILOPO, 2021; KHOURY, 2020; SALGANIK; FAZITO, et al., 2011), unavailable RDS structure (COUTINHO et al., 2019; KENDALL et al., 2019), and unavailable covariates (WU et al., 2017). Based on that fact, we use two datasets to verify our inferences: Faux datasets in package RDS (HANDCOCK; FELLOWS; GILE, 2021) and Project 90 study (CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC), 1992).

4.1 Faux dataset

Faux dataset is simulated data designed to "demonstrate RDS functions and analysis" (HANDCOCK; FELLOWS; GILE, 2021, p. 15). It contains information about the individual identification, the recruiter identification, the informed degree, and three covariates, two being binary and one assuming three possible values. The summary statistics are in Table 8. Figure 33 presents the degree distribution for this sample. The RDS structure is pictured in Figure 34. In total, there are 389 samples.

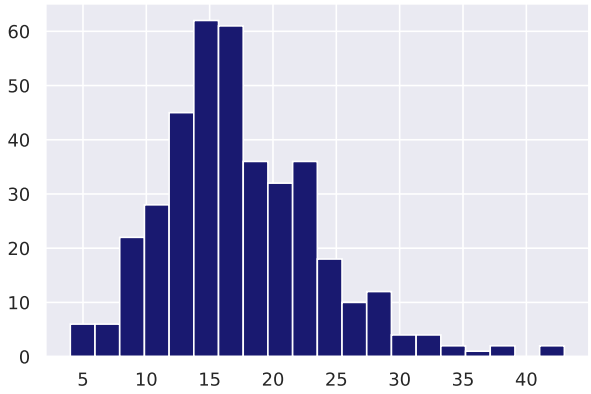
Table 8 – Summary statistics of Faux dataset.

Variable	Proportions
X	
red	0.7
blue	0.3
Y	
blue	0.44
green	0.38
black	0.18
Z	
red	0.57
blue	0.43

Source: The table was generated with data from Handcock, Fellows, and Gile (2021).

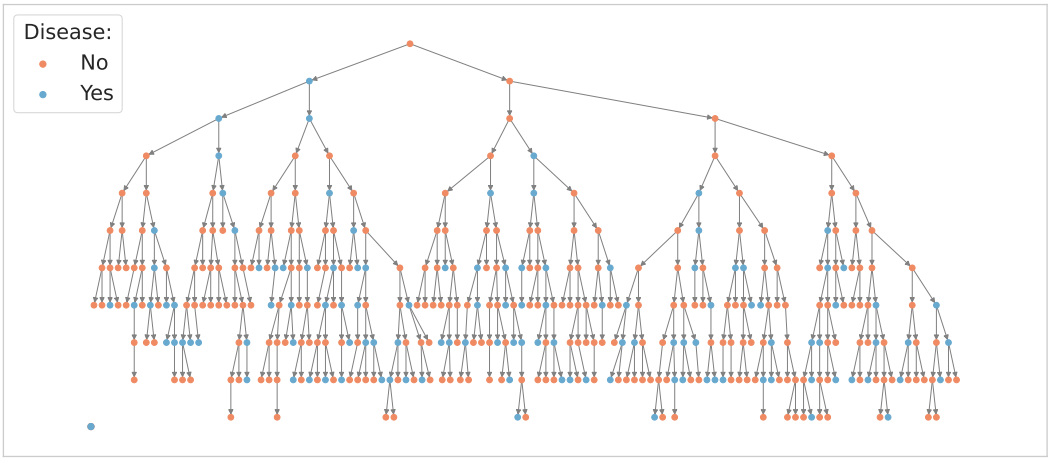
We interpret the variable X as the presence of the disease, in which blue indicates positive and red negative. We define variable $\hat{X}$ to be the diagnostic test result. We omit X and change it by $\hat{X}$ with sensitivity $\gamma_s = 0.9$ and specificity $\gamma_e = 0.85$. The prevalence is set to be 0.28, a little lower value than the detected in the dataset. The variables Y and Z are regressors. Table 9 presents the biased results when misclassification is not

Figure 33 – Histogram of the simulated informed degrees.



Source: The figure was generated with data from Handcock, Fellows, and Gile (2021).

Figure 34 – RDS structure in Faux dataset.



Source: The figure was generated with data from Handcock, Fellows, and Gile (2021) in NetworkX package. The colors indicate the result of the test, yes being positive and no being negative.

regarded. The 95% bootstrap confidence interval calculated for RDS-SS estimator was of $(0.352, 0.44)$. RDS package offers calculations of the main point estimates for prevalence and their corresponding variance estimators through bootstrapping. The only estimator we had to program was RDS-B.

Applying our model, we first verify that a gamma prior for τ is not indicated, since it puts to much mass on the assumption of correlation. It caused divergences in the HMC algorithm. Placing a Gumbel prior has the advantage of allowing τ be higher. In this case, the posterior mean was of order 10^5. The parameter ρ had posterior mean of 0.3. These results indicate absence of correlation among recruitments. The prevalece estimate was of 0.25, a much closer value to 0.28 than the others, even with weakly informative priors for the parameters. All the effects include 0 in the centered 50% credible interval, which indicates that none of them had effect on θ_i for the individuals.

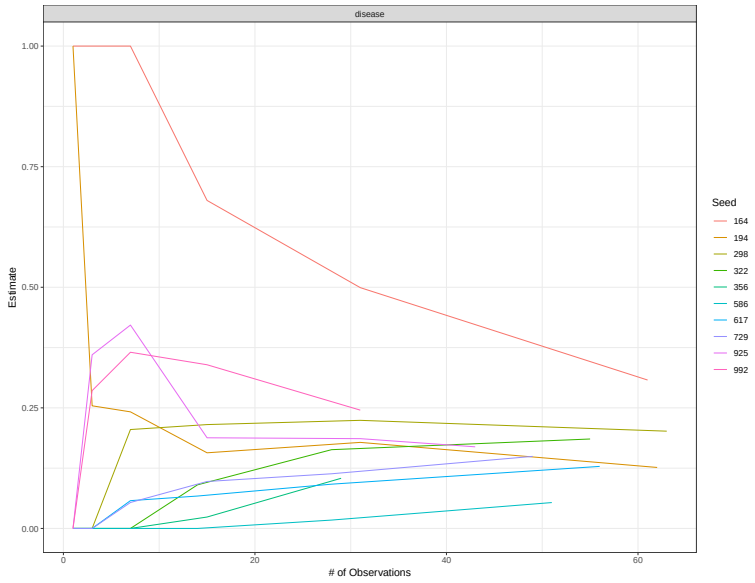
There are two variations of this dataset: madrona and sycamore. The difference

Table 9 – Prevalence point estimation of disease X by different approaches in faux dataset

Estimator	Ignoring misclassification	Frequentist correction
Naive	0.385	0.314
RDS-SH	0.4	0.333
RDS-VH	0.399	0.332
RDS-SS (N=1000)	0.396	0.331
RDS-SS (N=10000)	0.398	0.314
RDS-B	0.4	0.334

Source: Prepared by the author (2021) and based on the results of (HANDCOCK; FELLOWS; GILE, 2021), except for RDS-B, which was self-made. The second columns indicate the point estimate without considering the misclassification of the test, while the third corrects it with equation (2.5).

Figure 35 – Bottleneck plot for faux madrona dataset.

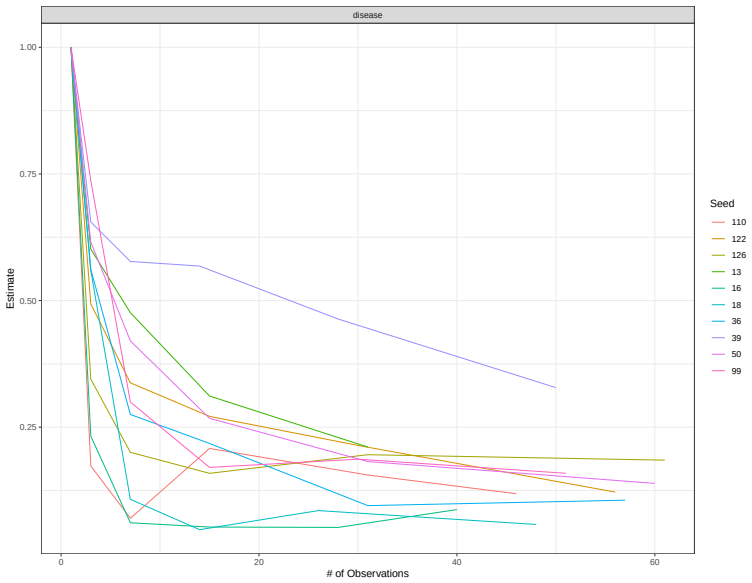


Source: Output of Handcock, Fellows, and Gile (2021)'s package.

between them is that the latter has extreme seed dependence, while the former not. The seed dependence is caused by sampling all the initial individuals within the infected population. Figure 35 and Figure 36 presents the bottleneck plots for each dataset. This plotting diagnostic introduced by Gile, Johnston, and Salganik (2015) measures the RDS-VH estimator for each chain generated by a different seed. Notice that in sycamore data, all seeds start in the highest part of the graphic and converge more slowly to the final estimate.

Both datasets are built with true prevalence of 0.2 and no covariate is available. They both have 500 samples. We make the inferences considering a diagnostic test with sensitivity $\gamma_s = 0.9$ and specificity $\gamma_e = 0.85$. Table 10 shows the resulting inferences. It seams that the naive estimator got closer to the correct value. This ocurred because the problems with the naive estimator compensate each other. All estimators had bad

Figure 36 – Bottleneck plot for faux sycamore dataset.



Source: Output of Handcock, Fellows, and Gile (2021)'s package.

performance in estimating the prevalence because the known sensitivity of the test is different from the calculated within the dataset, i.e., from the individuals without the disease, 90% tested negative, which is a greater value than the true value 85%. This little difference in Rogan and Gladen (1978)'s bias adjustment leads to a high difference in the estimates, which shows a non-robustness of the estimator. Without considering the correction for misclassification, the estimated bootstrap 95% confidence interval was of $(0.226, 0.279)$ for RDS-SS.

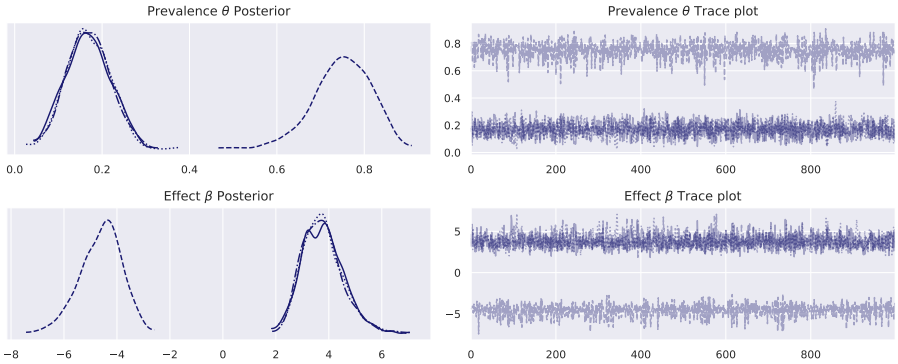
Table 10 – Prevalence point estimation of disease by different approaches in faux madrona dataset

Estimator	Ignoring misclassification	Frequentist correction
Naive	0.298	0.197
RDS-SH	0.228	0.104
RDS-VH	0.231	0.108
RDS-SS (N=1000)	0.253	0.137
RDS-SS (N=10000)	0.233	0.111
RDS-B	0.23	0.111

Source: Prepared by the author (2021) and based on the results of (HANDCOCK; FELLOWS; GILE, 2021), except for RDS-B, which was self-made. The second columns indicate the point estimate without considering the misclassification of the test, while the third corrects it with equation (2.5).

We first consider the informed degree as a covariate of the problem. Proposing weakly informative priors for θ and β, we observe a problem with $\hat{R}$, which indicates mixing problems among the chains. Increasing the number of warmup iterations did not solve the problem. Figure 37 shows an identifiability problem with this covariate in the

Figure 37 – Posterior distribution and trace plot for θ and β regarding model (3.4).



Source: Prepared by the author (2021) and based on Stan and ArviZ results.

model. In particular, the posterior distribution of β seems symmetric around 0, which is not a good inference. Disregarding the degree as covariate, the posterior mean for θ was of 0.19 (95% HDI of $0.07 - 0.31$), which is a pretty good estimate for the prevalence of 0.2.

For the faux sycamore dataset, the effect of the bias adjustment was less problematic because the true values for sensitivity and specificity were very close to the observed in the dataset. In special, RDS-SS estimator with known population size is the most close estimator. Since the seeds influence too much the inferences, the naive estimator was very far from the true value.

Table 11 – Prevalence point estimation of disease by different approaches in faux sycamore dataset

Estimator	Ignoring misclassification	Frequentist correction
Naive	0.354	0.272
RDS-SH	0.246	0.129
RDS-VH	0.27	0.161
RDS-SS (N=1000)	0.297	0.197
RDS-SS (N=10000)	0.273	0.164
RDS-B	0.273	0.164

Source: Prepared by the author (2021) and based on the results of (HANDCOCK; FELLOWS; GILE, 2021), except for RDS-B, which was self-made. The second columns indicate the point estimate without considering the misclassification of the test, while the third corrects it with equation (2.5).

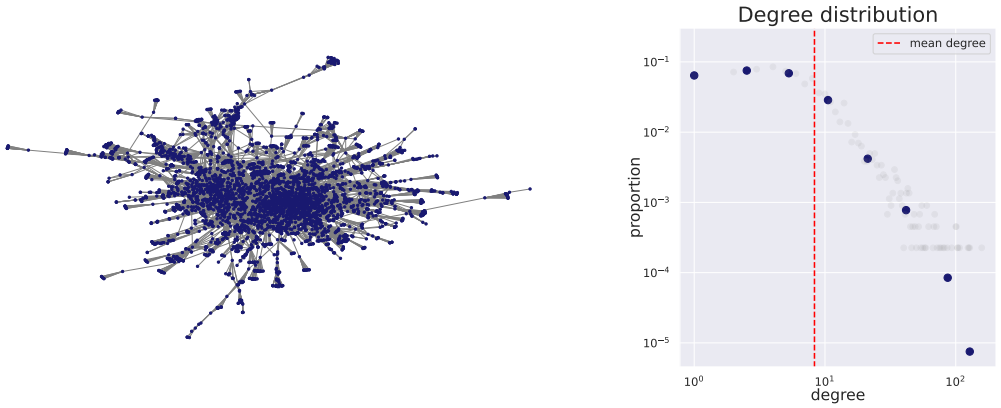
Our model had posterior mean of 0.25 for θ with 95% HDI of $0.06 - 0.41$. Since there is more seed dependence and the model does not accommodate sampling weights, the estimate gets higher than the true value and the credible interval gets wider.

4.2 Project 90 dataset

Colorado Springs Project 90 was the first prospective assessment of the impact caused by the network structure in the propagation of infectious diseases. The Centers for Disease Control and Prevention (CDC) funded this project in 1987 to investigate the dynamics of the human immunodeficiency virus (HIV) in high-risk heterosexual populations. From 1988 to 1992, the scientists enumerated 5492 individuals and their connections within the sample. The populations of the study included female sex workers, men who have sex with female sex workers, people who injected illicit drugs, and people who have sex with some injecting drug user (WOODHOUSE et al., 1994, p. 1332). To form a network of contacts between the participants, the researchers asked them to describe the relationships with every other named contact. Only partial information is available for privacy concerns.

There are 17 nodes without connections and 108 connected components in the dataset. The greater has 4430 nodes, while the second has 50 nodes. Therefore, we focus on this bigger connected subgraph. Figure 38 presents the graph structure and the degree distribution of the nodes. The degree distribution graph follows the instructions of Barabási (2013, Advanced Topic 3.A) with a log-log plot and logarithmic binning. The slow decay of the distribution indicates a scale-free network (BARABÁSI, 2013).

Figure 38 – Graph structure and degree distribution of the individuals from Project 90 study.



Source: Data from Centers for Disease Control and Prevention (CDC) (1992) and graphics prepared by the author (2021). The degree distribution is in the log-log scale. The grey dots represent the degree histogram, while the blue dots calculate the mean for each interval defined by logarithmic binning. The construction follows Barabási (2013).

Table 12 summarizes the binary variables of the dataset. In addition to these variables, the data inform the race of each individual with 75% White, 21% Black, 1% Asian/Pacific Islander, 1% Native americans, and 0.1% being others. Around 93% of the participants possess the complete data. We remove the rest from the dataset for simplicity, but in real applications, the uncertainty about this data should be considered.

Since no information about HIV test is available, we use the column disabled as

Table 12 – Proportion distribution for each binary variable in Project 90 dataset.

Covariate	Percentage (%)			Covariate	Percentage (%)		
	No	Yes	NA		No	Yes	NA
Female	56.8	43.2	0.0	Thief	91.8	2.2	6.0
Sex worker	88.7	5.2	6.1	Retired	91.1	2.9	6.0
Pimp	92.5	1.5	6.0	Housewife	88.0	6.0	5.9
Sex work client	91.0	8.9	0.1	Disabled	89.9	4.1	6.0
Drug dealer	87.6	6.4	6.0	Unemployed	77.8	16.2	6.0
Drug cook	93.2	0.8	6.0	Homeless	92.8	1.2	6.0

Source: Data from Centers for Disease Control and Prevention (CDC) (1992).

the outcome of interest. We define the sensitivity $\gamma_s = 0.9$ and specificity $\gamma_e = 0.85$ for the diagnostic test and perform a random testing. Moreover, we perform RDS in this graph following Baraff, McCormick, and Raftery (2016), as in Section 3.4.3 and Crawford (2016), as in Subsection 3.4.4. We compare the results from our model to the value in the whole network of 0.047.

Placing priors $\theta \sim \text{Unif}(0, 1)$ and $\beta_i \sim \text{Normal}(0, 1)$, we get a posterior mean for θ of 0.075 (94% HDI of $0 - 0.227$) for Baraff, McCormick, and Raftery (2016)'s simulation and 0.059 (94% HDI of $0 - 0.093$) for Crawford (2016)'s model. All of the effect parameters β had 80% HDI credible intervals including the value 0, which indicates no evidence of relation between the variables and the disabled condition. The prior for τ was Gumbel with parameter $\lambda_\tau = \log(10)$, while $\rho \sim \text{Unif}(0, 1)$. The posterior mean of τ was more than 20000 and the posterior mean of ρ was 0.476. In light of the dataset's size and the analysis made in Section 2.3.2 about ρ, we conclude that there is not much graphical correlation between recruiters and recruit regarding the disabled condition. Including gamma prior for τ did not change the parameter estimates, except for τ, as expected.

Choosing the seeds within the disabled subpopulation modified the estimate to 0.141 (94% HDI of $0.0 - 0.325$). The seed dependence bias the results, but the credible interval is larger, which is a good sign for uncertainty quantification. This is explained by the less diverse information contained in the sample, which reduces the process of inductive learning.

5 Conclusions

Respondent-driven sampling is a useful tool when the researchers cannot directly enumerate the population since building a dual incentive system encourages the individuals from the target population to engage in the research and convince others to do the same. Analysis of RDS started with Heckathorn (1997) and much work has been done as a wall of bricks. The resulting structure of the process needs many assumptions that need to be considered in our analysis. This work proposed a method to quantify the uncertainty about the characteristics of these populations, which is vital for robust decision making.

Regression analysis is a powerful toolbox of statistical procedures that help the understanding of the populations, but "with great power comes great responsibility" (LEE; DITKO, 1962, p. 13). When correct analyses are not done, problems of estimation, such as identifiability and biases, can lead to wrong inferences. In this work, we showed the importance of identifiability and solutions in the Bayesian paradigm. CAR models showed to be a good representation of the correlation between recruitments, but with hidden problems. The parameter of correlation is very difficult to interpret and even high values may not generate the correct expected dependencies. These models complicate the sampling by increasing the parameter space dimension and leading to an intricate geometry. In addition, this model can be a non-useful model when correlation is not so strong.

We also analysed the uncertainty of the graph through Crawford (2016)'s model sampling from its posterior distribution using a Metropolis-within-Gibbs method. This method showed to not wide the credible intervals, at least with our experiment settings. Furthermore, model extensions from the literature addressing distinct problems were described as future work.

For unbiased estimation of prevalence, this work presented the relevant role of sensitivity and specificity, in special when considering their uncertainties. RDS estimators, without including the accuracy of the diagnostic test, usually failed to give reasonable results. We also analysed different prior specifications of sensitivity and specificity, since strong priors are required. We highlighted each one's advantages and disadvantages. In particular, we detailed several characteristics of the bivariate beta distribution derived by Olkin and Trikalinos (2015). This distribution showed to have some specification problems since not all information can be directly converted to a well-defined distribution.

Another key aspect of Bayesian inference is the prior specification. It can save our life from identification problems, but it can derive wrong results when badly specified. In particular, when there is not much spatial correlation, a finite mean prior specification leads to divergences in the sampling method and incorrect inferences. The Gumbel type-II,

derived as a penalized complexity prior for the precision when data comes from normal distribution, showed to fit better in general cases. Moreover, prior information is not always easily converted to prior distributions as we studied in the logit normal case. Prior and posterior predictive checks are possible ways to understand the process.

Analysis of HMC results and its diagnostics are very important for the learning process. Observing divergences, mixing of the chain, energy of the model, among other diagnoses, can enlighten problems with the parametrization of the model and indicate possible solution paths. These diagnostics proved to be useful in implementing a better reparametrization of the model.

Finally, the simulations showed two important characteristics of the presented model: the computational burden is high with very complicated geometry, and the inferences are consistent when HMC is well diagnosed and prior specification is well thought. Therefore, it can be a valid option in respondent-driven samples, especially when compared to the most used point estimates for prevalence, such as RDS-VH and RDS-SS, and bootstrap distributions.

References

AVERY, Lisa. **Statistical Methods for Studies Using Respondent Driven Sampling with Applications to Urban Indigenous Health**. 2020. PhD thesis – York University, Toronto, Ontario.

BANERJEE, Sudipto; CARLIN, Bradley P; GELFAND, Alan E. **Hierarchical modeling and analysis for spatial data**. [S.l.]: Chapman and Hall/CRC, 2003.

BARABÁSI, Albert-László. Network science. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 371, n. 1987, p. 20120375, 2013.

BARABÁSI, Albert-László; ALBERT, Réka. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BARAFF, Aaron J; MCCORMICK, Tyler H; RAFTERY, Adrian E. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 113, n. 51, p. 14668–14673, 2016.

BASTOS, Francisco I; BASTOS, Leonardo Soares, et al. HIV, HCV, HBV, and syphilis among transgender women from Brazil: assessing different methods to adjust infection rates of a hard-to-reach, sparse population. **Medicine**, Wolters Kluwer Health, v. 97, 1 Suppl, 2018.

BASTOS, Leonardo S.; CARVALHO, Luiz M.; GOMES, Marcelo F.C. Modelling misreported data. In: GAMERMAN, Dani et al. **Building a Platform for Data-Driven Pandemic Prediction**. Boca Raton: CRC Press, 2021. chap. 7, p. 113–139.

BASTOS, Leonardo S.; PINHO, Adriana A., et al. **Binary regression analysis with network structure of respondent-driven sampling data**. [S.l.: s.n.], 2012. arXiv: 1206.5681 [stat.AP].

BESAG, Julian. Spatial interaction and the statistical analysis of lattice systems. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 36, n. 2, p. 192–225, 1974.

BETANCOURT, Michael. A conceptual introduction to Hamiltonian Monte Carlo. **arXiv preprint arXiv:1701.02434**, 2017.

_____. Cruising the simplex: Hamiltonian Monte Carlo and the Dirichlet distribution. In: AMERICAN INSTITUTE OF PHYSICS, 1. AIP Conference Proceedings 31st. [S.l.: s.n.], 2012. v. 1443, p. 157–164.

BETANCOURT, Michael. Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. **arXiv preprint arXiv:1604.00695**, 2016.

BETANCOURT, Michael; GIROLAMI, Mark. Hamiltonian Monte Carlo for hierarchical models. **Current trends in Bayesian methodology with applications**, CRC Press Boca Raton, FL, v. 79, n. 30, p. 2–4, 2015.

BRANSCUM, AJ; GARDNER, IA; JOHNSON, WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. **Preventive veterinary medicine**, Elsevier, v. 68, n. 2-4, p. 145–163, 2005.

BROOK, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. **Biometrika**, JSTOR, v. 51, n. 3/4, p. 481–483, 1964.

CARPENTER, Bob et al. Stan: A probabilistic programming language. **Journal of statistical software**, v. 76, n. 1, p. 1–32, 2017.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC). **Project 90 (Partial Data)**. [S.l.: s.n.], 1992. https://opr.princeton.edu/archive/p90/.

CHU, Haitao; COLE, Stephen R. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. **Journal of clinical epidemiology**, Elsevier Limited, v. 59, n. 12, p. 1331, 2006.

COUTINHO, Carolina et al. The risks of HCV infection among Brazilian crack cocaine users: incorporating diagnostic test uncertainty. **Scientific reports**, Nature Publishing Group, v. 9, n. 1, p. 1–9, 2019.

CRAWFORD, Forrest W; WU, Jiacheng; HEIMER, Robert. Hidden population size estimation from respondent-driven sampling: a network approach. **Journal of the American Statistical Association**, Taylor & Francis, v. 113, n. 522, p. 755–766, 2018.

CRAWFORD, Forrest W. The Graphical Structure of Respondent-driven Sampling. **Sociological Methodology**, v. 46, n. 1, p. 187–211, 2016. Available from: <https://doi.org/10.1177/0081175016641713>.

DAMACENA, Giseli Nogueira et al. Application of the Respondent-Driven Sampling methodology in a biological and behavioral surveillance survey among female sex workers, Brazil, 2016. **Revista Brasileira de Epidemiologia**, SciELO Brasil, v. 22, 2019.

DANIEL, J. **Sampling Essentials: Practical Guidelines for Making Sampling Choices**. [S.l.]: SAGE Publications, 2011. ISBN 9781452238401. Available from: <https://books.google.com.br/books?id=RJC87h4hCvIC>.

DEAUX, Edward; CALLAGHAN, John W. Key Informant Versus Self-Report Estimates of Health-Risk Behavior. **Evaluation Review**, v. 9, n. 3, p. 365–368, 1985. Available from: <https://doi.org/10.1177/0193841X8500900308>.

DONEGAN, Connor. Spatial Conditional Autoregressive Models in Stan. OSF Preprints, 2020.

ERDOS, Paul; RÉNYI, Alfréd, et al. On the evolution of random graphs. **Publ. Math. Inst. Hung. Acad. Sci**, Citeseer, v. 5, n. 1, p. 17–60, 1960.

FELLOWS, Ian E. Respondent-driven sampling and the homophily configuration graph. **Statistics in medicine**, Wiley Online Library, v. 38, n. 1, p. 131–150, 2019.

FISHER, Ronald A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character**, The Royal Society London, v. 222, n. 594-604, p. 309–368, 1922.

GAMERMAN, D.; LOPES, H.F. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition**. [S.l.]: Taylor & Francis, 2006. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781584885870. Available from: <https://books.google.com.br/books?id=yPvECi%5C_L3bwC>.

GELFAND, Alan E; SAHU, Sujit K. Identifiability, improper priors, and Gibbs sampling for generalized linear models. **Journal of the American Statistical Association**, Taylor & Francis, v. 94, n. 445, p. 247–253, 1999.

GELMAN, Andrew. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). **Bayesian analysis**, International Society for Bayesian Analysis, v. 1, n. 3, p. 515–534, 2006.

_____. Scaling regression inputs by dividing by two standard deviations. **Statistics in medicine**, Wiley Online Library, v. 27, n. 15, p. 2865–2873, 2008.

GELMAN, Andrew; CARPENTER, Bob. Bayesian analysis of tests with unknown specificity and sensitivity. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 69, n. 5, p. 1269–1283, 2020.

GELMAN, Andrew; JAKULIN, Aleks, et al. A weakly informative default prior distribution for logistic and other regression models. **The annals of applied statistics**, Institute of Mathematical Statistics, v. 2, n. 4, p. 1360–1383, 2008.

GILE, Krista J. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. **Journal of the American Statistical Association**, Taylor & Francis, v. 106, n. 493, p. 135–146, 2011.

GILE, Krista J; BEAUDRY, Isabelle S, et al. Methods for inference from respondent-driven sampling data. **Annual Review of Statistics and Its Application**, Annual Reviews, v. 5, p. 65–93, 2018.

GILE, Krista J; HANDCOCK, Mark S. Network model-assisted inference from respondent-driven sampling data. **Journal of the Royal Statistical Society. Series A,(Statistics in Society)**, NIH Public Access, v. 178, n. 3, p. 619, 2015.

GILE, Krista J; HANDCOCK, Mark S. Respondent-driven sampling: An assessment of current methodology. **Sociological methodology**, Sage Publications Sage CA: Los Angeles, CA, v. 40, n. 1, p. 285–327, 2010.

GILE, Krista J; JOHNSTON, Lisa G; SALGANIK, Matthew J. Diagnostics for respondent-driven sampling. **Journal of the Royal Statistical Society. Series A,(Statistics in Society)**, NIH Public Access, v. 178, n. 1, p. 241, 2015.

GOEL, Sharad; SALGANIK, Matthew J. Assessing respondent-driven sampling. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 107, n. 15, p. 6743–6747, 2010.

_____. Respondent-driven sampling as Markov chain Monte Carlo. **Statistics in medicine**, Wiley Online Library, v. 28, n. 17, p. 2202–2229, 2009.

GOODMAN, Leo A. Snowball Sampling. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 32, n. 1, p. 148–170, 1961. Available from: <https://doi.org/10.1214/aoms/1177705148>.

GREEN, AKB; MCCORMICK, TH; RAFTERY, AE. Consistency for the tree bootstrap in respondent-driven sampling. **Biometrika**, Oxford University Press, v. 107, n. 2, p. 497–504, 2020.

GUO, Jingyi; RIEBLER, Andrea; RUE, Håvard. Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors. **Statistics in medicine**, Wiley Online Library, v. 36, n. 19, p. 3039–3058, 2017.

HAGBERG, Aric; SWART, Pieter; S CHULT, Daniel. **Exploring network structure, dynamics, and function using NetworkX**. [S.l.], 2008.

HANDCOCK, Mark S.; FELLOWS, Ian E.; GILE, Krista J. **RDS: Respondent-Driven Sampling**. Los Angeles, CA, 2021. R package version 0.9-3. Available from: <https://CRAN.R-project.org/package=RDS>.

HECKATHORN, Douglas D. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. **Social problems**, Oxford University Press, v. 49, n. 1, p. 11–34, 2002.

_____. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. **Social Problems**, [Oxford University Press, Society for the Study of Social Problems], v. 44, n. 2, p. 174–199, 1997. Available from: <http://www.jstor.org/stable/3096941>.

HOFF, Peter D. **A first course in Bayesian statistical methods**. [S.l.]: Springer, 2009. v. 580.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing In Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

JIN, Xiaoping; CARLIN, Bradley P; BANERJEE, Sudipto. Generalized hierarchical multivariate CAR models for areal data. **Biometrics**, Wiley Online Library, v. 61, n. 4, p. 950–961, 2005.

KENDALL, Carl et al. The 12 city HIV surveillance survey among MSM in Brazil 2016 using respondent-driven sampling: a description of methods and RDS diagnostics. **Revista Brasileira de Epidemiologia**, SciELO Public Health, v. 22, e190004, 2019.

KHOURY, Rana B. Hard-to-Survey Populations and Respondent-Driven Sampling: Expanding the Political Science Toolbox. **Perspectives on Politics**, Cambridge University Press, v. 18, n. 2, p. 509–526, 2020.

KÜCHENHOFF, H. The identification of logistic regression models with errors in the variables. **Statistical Papers**, Springer, v. 36, n. 1, p. 41–47, 1995.

KUMAR, Ravin et al. ArviZ a unified library for exploratory analysis of Bayesian models in Python. **Journal of Open Source Software**, The Open Journal, v. 4, n. 33, p. 1143, 2019. DOI: 10.21105/joss.01143. Available from: <https://doi.org/10.21105/joss.01143>.

LEE, Duncan. A comparison of conditional autoregressive models used in Bayesian disease mapping. **Spatial and spatio-temporal epidemiology**, Elsevier, v. 2, n. 2, p. 79–89, 2011.

_____. CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. **Journal of Statistical Software**, American Statistical Association, v. 55, n. 13, p. 1–24, 2013.

LEE, Stan; DITKO, Steve. Spider Man: Amazing Fantasy 15. **Marvel comics**, 1962.

LEEFLANG, Mariska MG et al. Variation of a test's sensitivity and specificity with disease prevalence. **Cmaj**, Can Med Assoc, v. 185, n. 11, e537–e544, 2013.

LEHMANN, Eric L. Model specification: the views of Fisher and Neyman, and later developments. In: SELECTED Works of EL Lehmann. [S.l.]: Springer, 2012. P. 955–963.

LEVIN, David A; PERES, Yuval. **Markov chains and mixing times**. [S.l.]: American Mathematical Soc., 2017. v. 107.

LIN, Jiayu. On the dirichlet distribution. **Mater's Report**, Queen's University Kingston Ontario, Canada, 2016.

LINDLEY, Dennis Victor. **Bayesian statistics: A review**. [S.l.]: SIAM, 1972.

MAX JOSEPH. **Exact sparse CAR models in Stan**. [S.l.: s.n.], 2021. Stan documentation. Available from: <https://mc-stan.org/users/documentation/case-studies/mbjoseph-CARStan.html>. Visited on: 16 Aug. 2021.

MCCULLAGH, Peter; NELDER, John A. **Generalized linear models**. [S.l.]: Routledge, 2019.

MCINTURFF, Pat et al. Modelling risk when binary outcomes are subject to error. **Statistics in medicine**, Wiley Online Library, v. 23, n. 7, p. 1095–1109, 2004.

MCLAUGHLIN, Katherine R. A Bayesian framework for modelling the preferential selection process in respondent-driven sampling. **Statistical Modelling**, SAGE Publications Sage India: New Delhi, India, p. 1471082x211043945, 2021.

METACADEMY. **Hamiltonian flows**. [S.l.: s.n.], 2021. https://metacademy.org/graphs/concepts/hamiltonian_flows.

MEURER, Aaron et al. SymPy: symbolic computing in Python. **PeerJ Computer Science**, v. 3, e103, Jan. 2017. ISSN 2376-5992. DOI: 10.7717/peerj-cs.103. Available from: <https://doi.org/10.7717/peerj-cs.103>.

MITCHELL, Stephanie L et al. Performance of SARS-CoV-2 antigen testing in symptomatic and asymptomatic adults: a single-center evaluation. **BMC Infectious Diseases**, Springer, v. 21, n. 1, p. 1–7, 2021.

MOTA, Rosa Maria Salani. **Respondent driven sampling (RDS) aplicado à população de homens que fazem sexo com homens no Brasil**. 2012. PhD thesis – Universidade Federal do Ceará. Faculdade de Medicina, Fortaleza.

NOORDZIJ, Marlies et al. Measures of disease frequency: prevalence and incidence. **Nephron Clinical Practice**, Karger Publishers, v. 115, n. 1, p. c17–c20, 2010.

OGLE, Kiona; BARBER, Jarrett J. Ensuring identifiability in hierarchical mixed effects Bayesian models. **Ecological Applications**, Wiley Online Library, v. 30, n. 7, e02159, 2020.

OLKIN, Ingram; TRIKALINOS, Thomas A. Constructions for a bivariate beta distribution. **Statistics & Probability Letters**, Elsevier, v. 96, p. 54–60, 2015.

OTT, Miles Q et al. Reduced bias for respondent-driven sampling: accounting for non-uniform edge sampling probabilities in people who inject drugs in Mauritius. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 68, n. 5, p. 1411–1429, 2019.

PARIKH, Rajul et al. Understanding and using sensitivity, specificity and predictive values. **Indian journal of ophthalmology**, Wolters Kluwer–Medknow Publications, v. 56, n. 1, p. 45, 2008.

PERESTROIKA, Grhasta Dian; PRABANDARI, Yayi Suryo; WILOPO, Siswanto Agus. Sexual Intercourse Among Early Adolescents in Semarang, Central Java, Indonesia: Survey Using RDS. **Asia Pacific Journal of Public Health**, SAGE Publications Sage CA: Los Angeles, CA, p. 10105395211053157, 2021.

REITSMA, Johannes B et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. **Journal of Clinical Epidemiology**, Elsevier, v. 58, n. 10, p. 982–990, 2005.

RIDDELL, Allen; HARTIKAINEN, Ari; CARTER, Matthew. **pystan (3.0.0)**. [S.l.: s.n.], Mar. 2021. PyPI.

ROBERT, Christian. **The Bayesian choice: from decision-theoretic foundations to computational implementation**. [S.l.]: Springer Science & Business Media, 2007.

ROBERT, Christian P; CASELLA, George; CASELLA, George. **Monte Carlo statistical methods**. [S.l.]: Springer, 2004. v. 2.

ROGAN, Walter J; GLADEN, Beth. Estimating prevalence from the results of a screening test. **American journal of epidemiology**, Oxford University Press, v. 107, n. 1, p. 71–76, 1978.

ROTHMAN, Kenneth J; GREENLAND, Sander; LASH, Timothy L, et al. **Modern epidemiology**. [S.l.]: Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008. v. 3.

RUTJES, AWS et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. **HEALTH TECHNOLOGY ASSESSMENT-SOUTHAMPTON-**, National Coordinating Centre for Health Technology Assessment, v. 11, n. 50, 2007.

SALGANIK, Matthew J. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. **Journal of Urban Health**, Springer, v. 83, n. 1, p. 98, 2006.

SALGANIK, Matthew J; FAZITO, Dimitri, et al. Assessing network scale-up estimates for groups most at risk of HIV/AIDS: evidence from a multiple-method study of heavy drug users in Curitiba, Brazil. **American journal of epidemiology**, Oxford University Press, v. 174, n. 10, p. 1190–1196, 2011.

SALGANIK, Matthew J; HECKATHORN, Douglas D. Sampling and estimation in hidden populations using respondent-driven sampling. **Sociological methodology**, Wiley Online Library, v. 34, n. 1, p. 193–240, 2004.

SCHERVISH, Mark J. **Theory of statistics**. [S.l.]: Springer Science & Business Media, 2012.

SIMPSON, Daniel et al. Penalising model component complexity: A principled, practical approach to constructing priors. **Statistical science**, Institute of Mathematical Statistics, v. 32, n. 1, p. 1–28, 2017.

ŠIMUNDIĆ, Ana-Maria. Measures of diagnostic accuracy: basic definitions. **Ejifcc**, International Federation of Clinical Chemistry and Laboratory Medicine, v. 19, n. 4, p. 203, 2009.

SPILLER, Michael. **Regression modeling of data collected using respondentdriven sampling**. 2009. PhD thesis – Cornell University.

STAN DEVELOPMENT TEAM. **Vectorization**. [S.l.: s.n.], 2021. Stan documentation. Available from: <https://mc-stan.org/docs/2_18/stan-users-guide/vectorization.html>. Visited on: 12 Sept. 2021.

STATISTICAT, LLC. LaplacesDemon: A Complete Environment for Bayesian Inference within R. **R Package version**, v. 17, p. 2016, 2016.

TOLEDO, Lidiane et al. Putting respondent-driven sampling on the map: insights from Rio de Janeiro, Brazil. **JAIDS Journal of Acquired Immune Deficiency Syndromes**, LWW, v. 57, s136–s143, 2011.

VEHTARI, Aki et al. Rank-normalization, folding, and localization: An improved R hat for assessing convergence of MCMC. **arXiv preprint arXiv:1903.08008**, 2019.

VERSI, E. " Gold standard" is an appropriate term. **BMJ: British Medical Journal**, BMJ Publishing Group, v. 305, n. 6846, p. 187, 1992.

VOLZ, Erik; HECKATHORN, Douglas D. Probability based estimation theory for respondent driven sampling. **Journal of Official Statistics**, Statistics Sweden (SCB), v. 24, n. 1, p. 79, 2008.

WATTERS, John K.; BIERNACKI, Patrick. Targeted Sampling: Options for the Study of Hidden Populations. **Social Problems**, Oxford University Press, Society for the Study of Social Problems, v. 36, n. 4, p. 416–430, 1989. Available from: <http://www.jstor.org/stable/800824>.

WILL KURT. **The Logit-Normal: A ubiquitous but strange distribution!** [S.l.: s.n.], 2021. Blog Count Bayesie. Available from: <https://www.countbayesie.com/blog/2021/9/30/the-logit-normal-a-ubitiqutious-but-strange-distribution>. Visited on: 12 Nov. 2021.

WOODHOUSE, Donald E et al. Mapping a social network of heterosexuals at high risk for HIV infection. **Aids**, v. 8, n. 9, p. 1331–1336, 1994.

WORLD HEALTH ORGANIZATION. **Introduction to HIV/AIDS and sexually transmitted infection surveillance: Module 4: Introduction to respondent-driven sampling**. [S.l.], 2013. 389 p., 30 cm. Available from: <https://apps.who.int/iris/handle/10665/116864>.

WU, Jiacheng et al. Using data from respondent-driven sampling studies to estimate the number of people who inject drugs: Application to the Kohtla-Järve region of Estonia. **PloS one**, Public Library of Science San Francisco, CA USA, v. 12, n. 11, e0185711, 2017.

XIE, Yang; CARLIN, Bradley P. Measures of Bayesian learning and identifiability in hierarchical models. **Journal of Statistical Planning and Inference**, Elsevier, v. 136, n. 10, p. 3458–3477, 2006.

YAUCK, Mamadou et al. General regression methods for respondent-driven sampling data. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 30, n. 9, p. 2105–2118, 2021.

Appendix

APPENDIX A – A bivariate beta distribution

Olkin and Trikalinos (2015) describe a bivariate distribution with beta marginal distributions, positive probability over the space $[0,1] \times [0,1]$, and correlations over the full range $(-1,1)$. In this section, we derive it and analyse some of its consequences as prior distribution.

A.1 Construction of the distribution

Let $U = (U_1, U_2, U_3, U_4) \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ with $\alpha_i > 0, i = 1, \ldots, 4$ and $U_4 = 1 - U_1 + U_2 + U_3$. The joint density of U with respect to the Lebesgue measure is given by

$$f_U(u_1, u_2, u_3) = \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} u_3^{\alpha_3 - 1} (1 - u_1 - u_2 - u_3)^{\alpha_4 - 1}, \tag{A.1}$$

when $u_i \in [0,1], i = 1,2,3$, $u_1 + u_2 + u_3 \leq 1$, and 0 otherwise. The normalizing constant is defined for $\boldsymbol{v} \in \mathbb{R}^n$ as

$$B(\boldsymbol{v}) = \frac{\prod_{i=1}^n \Gamma(v_i)}{\Gamma\left(\sum_{i=1}^n v_i\right)}.$$

Definition A.1.1. Let

$$X = U_1 + U_2 \text{ and } Y = U_1 + U_3. \tag{A.2}$$

The distribution of (X, Y) is *bivariate beta* with parameter $\boldsymbol{\alpha}$.

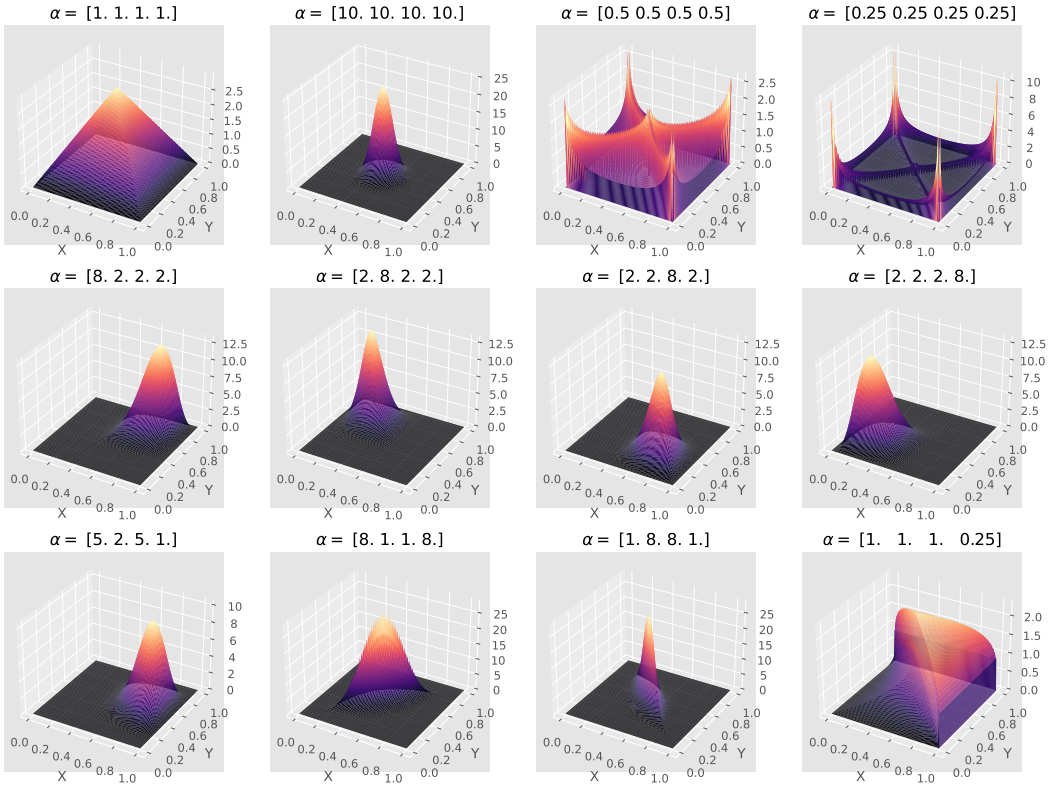
Figure 39 presents the joint density of X and Y for different values of $\boldsymbol{\alpha}$. The following two propositions describe the marginal and joint densities of bivariate beta distribution.

Proposition A.1.1 (Marginal distributions)**.** *The marginal distribution of X is Beta with parameters $\alpha_1 + \alpha_2$ and $\alpha_3 + \alpha_4$. Similarly, the marginal distribution of Y is Beta with parameters $\alpha_1 + \alpha_3$ and $\alpha_2 + \alpha_4$.*

Proof. First we derive the probability density of (U_1, U_2).

$$\begin{aligned} f_{U_1,U_2}(u_1, u_2) &= \int_{-\infty}^{\infty} f_U(u_1, u_2, u_3)\, du_3 \\ &= \frac{1}{B(\boldsymbol{\alpha})} \int_0^1 u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} u_3^{\alpha_3 - 1} (1 - u_1 - u_2 - u_3)^{\alpha_4 - 1}\, du_3 \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} \int_0^1 u_3^{\alpha_3 - 1} (1 - u_1 - u_2 - u_3)^{\alpha_4 - 1}\, du_3. \end{aligned} \tag{A.3}$$

Figure 39 – Joint density of the variables X and Y for different choices of $\boldsymbol{\alpha}$.



Source: Prepared by the author (2021). The four plots

in the first plot are symmetric and have no correlation between the variables. When $\alpha = [0.5, 0.5, 0.5, 0.5]$

Let $u_3 = (1 - u_1 - u_2)z$. Then,

$$\begin{aligned} f_{U_1,U_2}(u_1,u_2) &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1} u_2^{\alpha_2-1} \\ &\qquad \times \int_0^1 (1-u_1-u_2)^{\alpha_3-1} z^{\alpha_3-1} (1-u_1-u_2)^{\alpha_4} (1-z)^{\alpha_4-1}\, dz. \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1} u_2^{\alpha_2-1} (1-u_1-u_2)^{\alpha_3+\alpha_4-1} \int_0^1 z^{\alpha_3-1}(1-z)^{\alpha_4-1}\, dz. \qquad \text{(A.4)} \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1-1} u_2^{\alpha_2-1} (1-u_1-u_2)^{\alpha_3+\alpha_4-1} \frac{\Gamma(\alpha_3)\Gamma(\alpha_4)}{\Gamma(\alpha_3+\alpha_4)} \\ &= \frac{1}{B(\alpha_1,\alpha_2,\alpha_3+\alpha_4)} u_1^{\alpha_1-1} u_2^{\alpha_2-1} (1-u_1-u_2)^{\alpha_3+\alpha_4-1}. \end{aligned}$$

We conclude that

$$(U_1, U_2, 1 - U_1 - U_2) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3 + \alpha_4).$$

Define

$$H(v) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} v, \text{ for } v \in \mathbb{R}^2.$$

Then $(U_1, X) = H(U_1, U_2)$ and $H(\cdot)$ is bijective and differentiable function. By the Change of Variable Formula,

$$\begin{aligned} f_{U_1,X}(u_1, x) &= f(H^{-1}(u_1, x)) \left| \det \left[\left. \frac{dH^{-1}(v)}{dv} \right|_{v=(u_1,x)} \right] \right| \\ &= f(u_1, x - u_1) = \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} u_1^{\alpha_1 - 1} (x - u_1)^{\alpha_2 - 1} (1 - x)^{\alpha_3 + \alpha_4 - 1}, \end{aligned} \tag{A.5}$$

where (u_1, x) belongs to the triangle defined by the points (0,0), (0,1), and (1,1). The distribution of X for $x \in [0, 1]$ is

$$\begin{aligned} f_X(x) &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} \int_0^x u_1^{\alpha_1 - 1} (x - u_1)^{\alpha_2 - 1} (1 - x)^{\alpha_3 + \alpha_4 - 1} \, du_1 \\ &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} \int_0^x u_1^{\alpha_1 - 1} (x - u_1)^{\alpha_2 - 1} \, du_1. \\ &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} \\ &\qquad \times \int_0^x x^{\alpha_1 - 1} \left(\frac{u_1}{x} \right)^{\alpha_1 - 1} x^{\alpha_2 - 1} \left(1 - \frac{u_1}{x} \right)^{\alpha_2 - 1} \, du_1. \end{aligned} \tag{A.6}$$

Setting $u = u_1/x$ (if $x = 0, f_X(x) = 0$, then suppose $x > 0$), we have,

$$\begin{aligned} f_X(x) &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} x^{\alpha_1 + \alpha_2 - 1} \int_0^1 u^{\alpha_1 - 1} (1 - u)^{\alpha_2 - 1} \, du. \\ &= \frac{1}{B(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} x^{\alpha_1 + \alpha_2 - 1} B(\alpha_1, \alpha_2) \\ &= \frac{1}{B(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4)} (1 - x)^{\alpha_3 + \alpha_4 - 1} x^{\alpha_1 + \alpha_2 - 1} \end{aligned} \tag{A.7}$$

Therefore $X \sim \text{Beta}(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4)$. Similarly $Y \sim \text{Beta}(\alpha_1 + \alpha_3, \alpha_2 + \alpha_4)$. □

From the marginal distributions, we already know the expected values and variances of the random variables X and Y. Denote $\tilde{\alpha} = \sum_{i=1}^4 \alpha_i$ and we have

$$\begin{aligned} \mathbb{E}[X] &= \frac{\alpha_1 + \alpha_2}{\tilde{\alpha}}, & \mathbb{E}[Y] &= \frac{\alpha_1 + \alpha_3}{\tilde{\alpha}}, \\ \text{Var}[X] &= \frac{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}, & \text{Var}[Y] &= \frac{(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}. \end{aligned} \tag{A.8}$$

Proposition A.1.2 (Bivariate beta density)**.** *The joint density of (X, Y) with respect to the Lebesgue measure is given by*

$$f_{X,Y}(x, y) = \frac{1}{B(\boldsymbol{\alpha})} \int_\Omega u_1^{\alpha_1 - 1} (x - u_1)^{\alpha_2 - 1} (y - u_1)^{\alpha_3 - 1} (1 - x - y + u_1)^{\alpha_4 - 1} \, du_1, \tag{A.9}$$

where

$$\Omega = (\max(0, x + y - 1), \min(x, y)).$$

Proof. Note that

$$\begin{bmatrix} U_1 \\ X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix},$$

where the linear function is bijective and differentiable function, such that the determinant of the derivative is 1. By the Change of Variable Formula,

$$\begin{aligned} f_{U_1,X,Y}(u_1, x, y) &= f_{U_1,U_2,U_3}(u_1, x - u_1, y - u_2) \\ &= \frac{1}{B(\boldsymbol{\alpha})} u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1}(y - u_1)^{\alpha_3 - 1}(1 - x - y + u_1)^{\alpha_4 - 1}, \end{aligned} \tag{A.10}$$

where $0 \leq u_1 \leq x, u_1 \leq y$, and $0 \leq 1 - x - y + u_1$. Hence,

$$f_{X,Y}(x, y) = \frac{1}{B(\boldsymbol{\alpha})} \int_\Omega u_1^{\alpha_1 - 1}(x - u_1)^{\alpha_2 - 1}(y - u_1)^{\alpha_3 - 1}(1 - x - y + u_1)^{\alpha_4 - 1}\, du_1, \tag{A.11}$$

such that $\Omega = \{u_1 : \max(0, x + y - 1) < u_1 < \min(x, y)\}$. □

At last we derive the covariance and the correlation between X and Y.

Proposition A.1.3 (Covariance and correlation)**.** *The covariance between X and Y is*

$$\mathrm{Cov}(X, Y) = \frac{1}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}(\alpha_1\alpha_4 - \alpha_2\alpha_3)$$

and

$$\mathrm{Cor}(X, Y) = \frac{\alpha_1\alpha_4 - \alpha_2\alpha_3}{\sqrt{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)}}.$$

Proof. The covariance between U_i and U_j is (LIN, 2016, p. 11)

$$\mathrm{Cov}(U_i, U_j) = -\frac{\alpha_i\alpha_j}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}, i, j = 1, ..., 4, i \neq j \tag{A.12}$$

and the variance of U_i is

$$\mathrm{Var}(U_i) = \frac{\alpha_i(\tilde{\alpha} - \alpha_i)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}, \tag{A.13}$$

since $U_i \sim \mathrm{Beta}(\alpha_i, \tilde{\alpha} - \alpha_i)$. Therefore

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(U_1 + U_2, U_1 + U_3) = \frac{1}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}(\alpha_1\alpha_4 - \alpha_2\alpha_3). \tag{A.14}$$

□

Now we present an example where the full range of correlation is covered. Suppose X and Y have uniform distribution over $[0, 1]$, that is, they have beta distribution with parameter $1, 1$. Then, we have that

$$\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4 = \alpha_1 + \alpha_3 = \alpha_2 + \alpha_4 = 1,$$

whose solution is $\alpha_1 = \alpha_4 \in (0, 1)$ and $\alpha_2 = \alpha_3 = 1 - \alpha_4$. The correlation formula boils down to

$$\mathrm{Cor}(X, Y) = \alpha_4^2 - (1 - \alpha_4)^2 = 2\alpha_4 - 1 \in (-1, 1).$$

A.2 Implementation of the dirichlet distribution in Stan

The Dirichlet distribution is defined on the simplex of lower dimension. Therefore the sampler has to consider the restriction of $\sum_{i=1}^{4} U_i = 1$. Betancourt (2012) presents a simplification in the structure of the simplex. The propose is (BETANCOURT, 2012, p. 2)

$$z_i \sim \text{Beta}(\tilde{\alpha}_i, \alpha_i), \text{ where } \tilde{\alpha}_i = \sum_{k=i+1}^{4} \alpha_k, \quad i = 1, 2, 3$$

$$U_i = \left(\prod_{k=1}^{i-1} z_k \right) \cdot \begin{cases} 1 - z_i, & i < 4 \\ 1, & i = 4 \end{cases},$$

which removes the constraint.

A.3 Comments about integration

The density of (X, Y) is $f_{X,Y}(x, y)$ as in equation (A.11) and it can be undefined in sets of null Lebesgue measure in $\mathbb{R}^2$ and these sets may be important when plotting in a grid, for instance. This section illustrates one of these sets. If $\alpha_i \geq 1$, $i = 1, ..., 4$, the integral is clearly well defined for every $x, y \in [0, 1]$. Let $0 < \alpha_2 = \alpha_3 = a \leq 0.5$ and $x = y < 0.5$. Then

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{B(\boldsymbol{\alpha})} \int_0^x u_1^{\alpha_1 - 1} (x - u_1)^{a-1} (x - u_1)^{a-1} (1 - 2x + u_1)^{\alpha_4 - 1} \, du_1 \\ &= \frac{1}{B(\boldsymbol{\alpha})} \int_0^{x/2} u_1^{\alpha_1 - 1} (x - u_1)^{2a-2} (1 - 2x + u_1)^{\alpha_4 - 1} \, du_1 + \\ &\quad + \frac{1}{B(\boldsymbol{\alpha})} \int_{x/2}^x u_1^{\alpha_1 - 1} (x - u_1)^{2a-2} (1 - 2x + u_1)^{\alpha_4 - 1} \, du_1 \end{aligned}$$

Note that the first integral is well defined and non-negative. On the other hand, the second integral is not defined:

$$\begin{aligned} &\int_{x/2}^x u_1^{\alpha_1 - 1} (x - u_1)^{2a-2} (1 - 2x + u_1)^{\alpha_4 - 1} \, du_1 \\ &\quad \geq \int_{x/2}^x \min \left(\left(\frac{x}{2} \right)^{\alpha_1 - 1}, x^{\alpha_1 - 1} \right) (x - u_1)^{2a-2} \\ &\qquad\qquad \times \min \left(\left(1 - \frac{3}{2} x \right)^{\alpha_4 - 1}, (1 - x)^{\alpha_4 - 1} \right) du_1 \\ &\quad = K(x) \int_0^{x/2} v^{2a-2} \, dv \\ &\quad = \begin{cases} \dfrac{K(x)}{2a - 1} \lim_{t \to 0^+} \left[(x/2)^{2a-1} - t^{2a-1} \right] & \text{if } a < 0.5 \\ K(x) \lim_{t \to 0^+} \left[\log(x/2) - \log(t) \right] & \text{if } a = 0.5 \end{cases} \\ &\quad \to +\infty, \end{aligned}$$

where $K(x)$ is a function of x.

Based on this divergence, we conclude that if $0 < \alpha_2 = \alpha_3 \leq 0.5$ and $x = y < 0.5$, $f_{X,Y}(x, y)$ is not defined. Notice that if $x = y \geq 0.5$, divergence problems still happens, since the problems appear when u_1 approximates x. Similar calculations show that if $x + y = 1$ and $0 < \alpha_1 = \alpha_4 \leq 0.5$, the density is also not defined. More generally, $f_{X,Y}(x, y)$ is not defined if $\alpha_1 + \alpha_4 \leq 1$ and $x + y = 1$; $\alpha_2 + \alpha_3 \leq 1$ and $x = y$.

A.4 Elicitation of a bivariate beta

In this section, we develop a method to elicit the parameters of the bivariate beta distribution, which means to define an approximation $\hat{\boldsymbol{\alpha}}$ for the parameter $\boldsymbol{\alpha}$. This is an important step for the characterization of the prior distribution of model (3.2). If the researcher does not have information about the parameters previous seeing the data, two approaches are common in the independent beta setting and are adapted for the bivariate case:

a) both parameters receive a uniform distribution: in this case, as mentioned in Proposition A.1.3, $\alpha_1 = \alpha_4 \in (0, 1)$ and $\alpha_2 = \alpha_3 = 1 - \alpha_4$. The parameter α_4 is defined in a way that $\alpha_4 = \frac{1}{2}(1 + \text{Cor}(X, Y))$. If no information about the variables' correlation is available, it is recommended to use the independent setting since it is more flexible;

b) both parameters receive a Jeffreys' prior distribution (Beta$(1/2, 1/2)$): in this case, $\alpha_1 = \alpha_4 \in (0, 1/2)$ and $\alpha_2 = \alpha_3 = 1 - \alpha_4$. The parameter α_4 is defined in a way that $\alpha_4 = \frac{1}{4}(1 + \text{Cor}(X, Y))$.

Now, suppose that the researcher has information about following moments of the bivariate beta distribution: $m_1 = \mathbb{E}[X], m_2 = \mathbb{E}[Y], v_1 = \text{Var}(X), v_2 = \text{Var}(Y)$, and $\rho = \text{Cor}(X, Y)$. Notice that $v_1 + m_1^2 = \text{Var}(X_1) + \mathbb{E}[X_1]^2 = \mathbb{E}[X_1^2]$ and

$$\mathbb{E}[X_1^2] - \mathbb{E}[X_1] = \frac{(\alpha_1 + \alpha_2 + 1)(\alpha_1 + \alpha_2)}{(\tilde{\alpha} + 1)\tilde{\alpha}} - \frac{\alpha_1 + \alpha_2}{\tilde{\alpha}} = -\frac{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)}{\tilde{\alpha}(\tilde{\alpha} + 1)} < 0,$$

that is, $v_1 + m_1^2 - m_1 < 0 \implies v_1 < m_1 - m_1^2$ and similarly, $v_2 < m_2 - m_2^2$. After fixing these quantities, we will have a non-linear system with five equations and four unknown variables. Hence, we want to solve the following

$$\begin{cases} m_1 = \dfrac{\alpha_1 + \alpha_2}{\tilde{\alpha}}, \\ m_2 = \dfrac{\alpha_1 + \alpha_3}{\tilde{\alpha}}, \\ v_1 = \dfrac{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}, \\ v_2 = \dfrac{(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)}{\tilde{\alpha}^2(\tilde{\alpha} + 1)}, \\ \rho = \dfrac{\alpha_1\alpha_4 - \alpha_2\alpha_3}{\sqrt{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)}}. \end{cases} \tag{A.15}$$

Notice that we can simplify the third and fourth equations since

$$\frac{\alpha_3 + \alpha_4}{\tilde{\alpha}} = \frac{\tilde{\alpha} - (\alpha_1 + \alpha_2)}{\tilde{\alpha}} = 1 - m_1,$$

and analogously,

$$\frac{\alpha_2 + \alpha_4}{\tilde{\alpha}} = 1 - m_2.$$

Therefore,

$$\begin{aligned} v_1 &= \frac{m_1(1-m_1)}{\tilde{\alpha}+1}, \\ v_2 &= \frac{m_2(1-m_2)}{\tilde{\alpha}+1}. \end{aligned}$$

This already tells us that the system do not have a solution if

$$\frac{m_1(1-m_1)}{v_1} \neq \frac{m_2(1-m_2)}{v_2}.$$

The following proposition builds a solution excluding the fourth equation, given the above comment.

Proposition A.4.1. *System* (A.15) *without the fourth equation has a unique solution given by*

$$\begin{aligned} \alpha_1 &= (m_1 + m_2 - 1)\tilde{\alpha} + \alpha_4, \\ \alpha_2 &= (1 - m_2)\tilde{\alpha} - \alpha_4, \\ \alpha_3 &= (1 - m_1)\tilde{\alpha} - \alpha_4, \\ \alpha_4 &= \rho\tilde{\alpha}\sqrt{m_1 m_2 (1-m_1)(1-m_2)} + (1-m_1)(1-m_2), \end{aligned} \tag{A.16}$$

where $\tilde{\alpha}$ is given by the expression

$$\tilde{\alpha} = \frac{(m_1 - m_1^2 - v_1)}{v_1}.$$

Proof. The first two equations of the system (A.15) can be rewritten as a linear system:

$$\begin{aligned} (m_1 - 1)\alpha_1 + (m_1 - 1)\alpha_2 + m_1\alpha_3 + m_1\alpha_4 = 0, \\ (m_2 - 1)\alpha_1 + m_2\alpha_2 + (m_2 - 1)\alpha_3 + m_2\alpha_4 = 0, \end{aligned}$$

which is equivalent to

$$\begin{aligned} \alpha_1 + \alpha_2 + \frac{m_1}{m_1 - 1}\alpha_3 + \frac{m_1}{m_1 - 1}\alpha_4 = 0, \\ \alpha_2 + \frac{1 - m_2}{m_1 - 1}\alpha_3 + \frac{m_1 - m_2}{m_1 - 1}\alpha_4 = 0. \end{aligned}$$

Then, we can write α_1 and α_2 as functions of α_3 and α_4:

$$\alpha_1 = \frac{m_1 + m_2 - 1}{1 - m_1}\alpha_3 + \frac{m_2}{1 - m_1}\alpha_4 \tag{A.17}$$

$$\alpha_2 = \frac{1 - m_2}{1 - m_1}\alpha_3 + \frac{m_1 - m_2}{1 - m_1}\alpha_4. \tag{A.18}$$

Based on that expression, denote $\alpha_1 = a_3\alpha_3 + a_4\alpha_4$, $\alpha_2 = b_3\alpha_3 + b_4\alpha_4$, $c_3 = a_3 + b_3 + 1$, and $c_4 = a_4 + b_4 + 1$. Then, the third equation can be written as

$$a_3\alpha_3 + a_4\alpha_4 + b_3\alpha_3 + b_4\alpha_4 + \alpha_3 + \alpha_4 + 1 = c_3\alpha_3 + c_4\alpha_4 = \frac{m_1(1 - m_1)}{v_1} - 1,$$

which implies that

$$\alpha_3 = \frac{m_1(1 - m_1) - v_1 - c_4 v_1 \alpha_4}{c_3 v_1},$$

that is a linear function of α_4. We summarize the expressions in function of α_4 with some simplifications:

$$\begin{aligned}
\alpha_1 &= (m_1 + m_2 - 1)\frac{(m_1 - m_1^2 - v_1)}{v_1} + \alpha_4, \\
\alpha_2 &= (1 - m_2)\frac{(m_1 - m_1^2 - v_1)}{v_1} - \alpha_4, \\
\alpha_3 &= (1 - m_1)\frac{(m_1 - m_1^2 - v_1)}{v_1} - \alpha_4,
\end{aligned}$$

which implies that

$$\tilde{\alpha} = \frac{m_1 - m_1^2 - v_1}{v_1}.$$

Now rewrite the fifth equation using the first two equations from system (A.15) as follows

$$\begin{aligned}
\rho &= \frac{\alpha_1\alpha_4 - \alpha_2\alpha_3}{\sqrt{(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)}} \\
&= \frac{\alpha_1\alpha_4 - \alpha_2\alpha_3}{\tilde{\alpha}^2\sqrt{m_1 m_2(1 - m_1)(1 - m_2)}} \\
&= \frac{(m_1 + m_2 - 1)\tilde{\alpha}\alpha_4 + \alpha_4^2 - ((1 - m_2)\tilde{\alpha} - \alpha_4)((1 - m_1)\tilde{\alpha} - \alpha_4)}{\tilde{\alpha}^2\sqrt{m_1 m_2(1 - m_1)(1 - m_2)}} \\
&= \frac{\alpha_4 - (1 - m_1)(1 - m_2)\tilde{\alpha}}{\tilde{\alpha}\sqrt{m_1 m_2(1 - m_1)(1 - m_2)}}
\end{aligned} \tag{A.19}$$

and the solution is, therefore,

$$\alpha_4 = \rho\tilde{\alpha}\sqrt{m_1 m_2(1 - m_1)(1 - m_2)} + (1 - m_1)(1 - m_2).$$

□

There is an additional restriction to the sum given by the marginal distributions. Let $Z \sim \text{Beta}(a, b)$. Then:

$$\frac{\mathbb{E}[Z](1-\mathbb{E}[Z])}{\text{Var}[Z]} - 1 = \frac{\frac{ab}{(a+b)^2}}{\frac{ab}{(a+b)^2(a+b+1)}} - 1 = a + b,$$

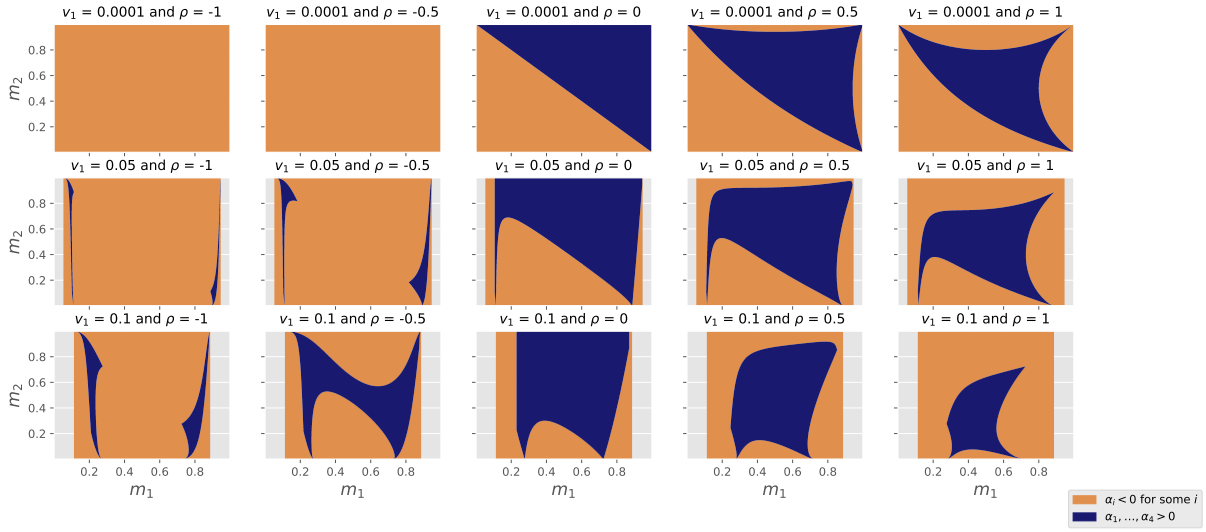
then

$$\sum_{i=1}^{4} \alpha_i = \frac{m_1(1-m_1)}{v_1} - 1 = \frac{m_2(1-m_2)}{v_2} - 1. \tag{A.20}$$

Besides solving the system (A.15), the bivariate beta distribution needs that $\alpha_1, \ldots, \alpha_4 > 0$. However, this is not always achievable. Since it is difficult to find the subset $D \subset [0,1]^4$ in which the solution for (A.16) is strictly positive for $\alpha_1, \ldots, \alpha_4$, we present some examples in Figure 40. For each subplot, the values of v_1 and ρ are fixed, while $m_1, m_2 \in [0,1]^2$. The grey area corresponds to the set where $v_1 \geq m_1 - m_2$, which is impossible. The orange area means that the solution to system (A.15) is not strictly positive. At last, the blue region is the set of interest.

When $\rho = -1$ for instance, only a few specifications of m_1 and m_2 generate a strictly positive solution. These examples show that several interesting specifications for the researchers can lead to a non positive solution, which is not desirable.

Figure 40 – Verification of positivity of the solution for different and fixed values of v_1 and ρ, and $m_1, m_2 \in [0,1]^2$.



Source: Prepared by the author (2021).

In light of this, we can only have an approximation using some optimization solver. From now on, we suppose the researcher has knowledge about m_1, m_2 and ρ. Through equations (A.17), (A.18), and solving the correlation equation for α_3 with help of the

symbolic solver SymPy (MEURER et al., 2017), we have the following three expressions:

$$\alpha_1 = \alpha_4 \frac{m_1 m_2 + \rho\sqrt{m_1 m_2 (1-m_1)(1-m_2)}}{(1-m_1)(1-m_2) + \rho\sqrt{m_1 m_2 (1-m_1)(1-m_2)}},$$
$$\alpha_2 = \alpha_4 \frac{m_1 (1-m_2) - \rho\sqrt{m_1 m_2 (1-m_1)(1-m_2)}}{(1-m_1)(1-m_2) + \rho\sqrt{m_1 m_2 (1-m_1)(1-m_2)}},$$
$$\alpha_3 = \alpha_4 \frac{m_2 (1-m_1) - \rho\sqrt{m_1 m_2 (1-m_1)(1-m_2)}}{(1-m_1)(1-m_2) + \rho\sqrt{m_1 m_2 (1-m_1)(1-m_2)}},$$

and $\alpha_4 > 0$ is a free parameter. In order to have $\boldsymbol{\alpha} > 0$, we have two situations:

a) the denominator of $\alpha_1, \alpha_2, \alpha_3$ is negative: in this case, it is not possible to have both α_2 and α_3 positives;

b) the denominator of $\alpha_1, \alpha_2, \alpha_3$ is positive: in this case, we have that

$$\rho \in \left(-\frac{\min(m_1 m_2, (1-m_1)(1-m_2))}{\sqrt{m_1 m_2 (1-m_1)(1-m_2)}}, \frac{\max(m_1, m_2) - m_1 m_2}{\sqrt{m_1 m_2 (1-m_1)(1-m_2)}} \right).$$

When $m_1 = m_2 = m$, the upper bound is 1 and the lower bound is

$$\begin{cases} -\dfrac{m}{1-m}, & m < 1/2 \\ -\dfrac{1-m}{m}, & m > 1/2. \end{cases}$$

Suppose that ρ belongs to this interval. Then we have to choose $\alpha_4 > 0$. Using a symbolic solver, we see that

$$\sum_{i=1}^{4} \alpha_i = \frac{\alpha_4}{(1-m_1)(1-m_2) + \rho\sqrt{m_1 m_2 (1-m_1)(1-m_2)}},$$

therefore v_1 and v_2 are inversely proportional to α_4. To have a higher variance, pick a small α_4. To have a lower variance, pick a large one. If ρ does not belong to the interval, taking a suitable value in it is a possibility.

Suppose the researcher has also knowledge about v_1 and v_2. By Proposition A.4.1 and Figure 40, there is no viable solution in several situations. Because of that, two approaches are suggested:

a) no variable is fixed: solve the optimizing problem given by Olkin and Trikalinos (2015, p. 7). The problem with this approach is that ρ and the means m_1, m_2 get far from the given values. Weights can be specified for each parameter to incorporate some preference;

b) fix m_1 and m_2 and let ρ, v_1 and v_2 vary: it is the limit of the above method, with the weights of m_1 and m_2 going to the infinity. It is more suitable when the researcher has stronger beliefs or information in the means than the other moments.

In this work, we use the second approach, which gives less importance to the correlation in comparison to the means.

Remark A.4.1. If the researcher has information about a credibility interval, this information needs to be converted in terms of the variance for our framework.

A.5 Simulate data

In this section, we experiment the estimation process of $\hat{\boldsymbol{\alpha}}$ through two different simulations:

a) simulating from bivariate beta: fix the parameter $\alpha = (0.5, 0.5, 0.5, 0.5)$ and generate 1000 different datasets from bivariate beta distribution of size 100; calculate m_1, m_2, v_1, v_2, and ρ and (i) solve the equations through Proposition (A.4.1), (ii) complete optimization problem, and (iii) optimization problem with m_1 and m_2 fixed, which we call *mixed solver*. The mean squared error is calculated;

b) simulating from bivariate logit normal distribution: fix the means and covariance matrix and follow the same instructions from the previous item.

Since the comparison is in terms of mean squared error, it is clear that the minimization problem will have the least value. When time is important, it is a very expensive method. Solving equations should be the best method, but under uncertainty, its results can have biases. Table 13 summarizes the results. Notice that when simulating from the bivariate beta or from the bivariate logit normal with parameters corresponding to the beta, the estimation process has little error.

Table 13 – Comparing the different methods for each simulation strategy.

Simulation	Method	MSE	s/ite
Bivariate beta	Solving equations	0.04	$3.47 \cdot 10^{-5}$
	Minimization problem	0.032	1.43
	Mixed solver	0.033	0.03
Logit bivariate normal	Solving equations	0.034	$5.21 \cdot 10^{-5}$
	Minimization problem	0.026	1.53
	Mixed solver	0.026	0.03

Source: Prepared by the author (2021). The MSE is the mean squared error, where the mean is taken with respect to the iterations. The s/ite is the number of seconds per iteration.

Using the logit bivariate normal simulation with parameters $\mu = (5, 2.3)$ and $\Sigma = [[12, -2.5], [-2.5, 4]]$ in order to yield $\mathbb{E}[X] \approx 0.9, \mathbb{E}[Y] \approx 0.8, \text{Var}(X) = \text{Var}(Y) \approx 0.05$, and $\text{Cor}(X, Y) \approx -0.2$, 77% of the simulations has no exact solution strictly positive and the error of the solvers were 0.82 for the mixed solver and 0.91 for the minimization problem, which is much higher than the previous simulation.

APPENDIX B – Sampling from the posterior distribution of the graph

Here we derive the method developed by Crawford (2016) for sampling from $p(G_S, \lambda \mid \boldsymbol{Z})$. The implementation used Python language. We first define the prior density for (G_s, λ) as

$$\pi(G_S, \lambda) = \pi(G_S) \times \pi(\lambda) = \frac{1}{|\mathcal{C}(\boldsymbol{Z})|} \times \frac{\beta_r^{\alpha_r}}{\Gamma(\alpha_r)} \lambda^{\alpha_r - 1} e^{-\beta_r \lambda},$$

where α_r and β_r are positive hyperparameters. Notice that G_S as uniform distribution over all compatible graphs and $\lambda \sim \text{Gamma}(\alpha_r, \beta_r)$.

To perform Gibbs sampling, we need to sample from $p(G_S \mid \lambda, \boldsymbol{Z})$ and $p(\lambda \mid G_S, \boldsymbol{Z})$. Since we do not know to sample from both conditional distributions, Crawford (2016) proposes to use Metropolis-Hastings for each one. Then, we need to specify a proposal distribution to both conditionals:

a) the proposal for $p(G_S|\lambda, \boldsymbol{Z})$ generates a new compatible graph G_S^* with a simple algorithm. It samples two random nodes i, j without replacement. If both nodes have less connections in G_S than their informed degrees, i.e., $u_i \geq 1$ and $u_j \geq 1$, and they are not connected in G_S, we include the edge $\{i, j\}$ in G_S^* and copy all the other edges. If the edge already is in G_S and it is not in G_R, we remove it from G_S^*. Otherwise we draw another two nodes and continue. We call this proposal of $P(G_S^* \mid G_S)$;

b) the proposal for $p(\lambda \mid G_S, \boldsymbol{Z})$ assumes a normal distribution for the Maximum likelihood estimator (MLE) $\hat{\lambda}$ of λ from the likelihood (2.8). Deriving it, we obtain

$$\begin{aligned}\frac{d}{d\lambda} \log L(w \mid G_S, \lambda) &= \frac{d}{d\lambda} \left[(n-m)\log(\lambda) + \sum_{k \text{ isn't seed}} \log(s_k) - \lambda \boldsymbol{s}^T w \right] \\ &= \frac{n-m}{\lambda} - \boldsymbol{s}^T w,\end{aligned}$$

where m is the number of seeds. At the optimal $\hat{\lambda}$, we have that

$$\frac{n-m}{\hat{\lambda}} - \boldsymbol{s}^T w = 0 \implies \hat{\lambda} = \frac{n-m}{\boldsymbol{s}^T w}.$$

The Fisher information is, under some regularity conditions,

$$I(\lambda) = -\mathbb{E}\left[-\frac{n-m}{\lambda^2} \right] = \frac{n-m}{\lambda^2}.$$

For a large n, we can assume

$$\sqrt{n-m}(\hat{\lambda} - \lambda) \xrightarrow{d} N(0, \lambda^2).$$

Then the proposal distribution is the normal distribution of mean $\hat{\lambda}$ and variance $\sigma^2 = \lambda^2/(n-m)$.

Now, it remains to calculate the probability of acceptance for both Metropolis-Hastings samplings. Crawford (2016) establishes several results to make the calculations more efficient. We omit them here. Using relation (2.17), the probabilities are:

a) for the graph transition:

$$\begin{aligned}\alpha(G_S^* \mid G_S) &= \min\left(1, \frac{p(G_S^*|\lambda, \boldsymbol{Z})P(G_S \mid G_S^*)}{p(G_S|\lambda, \boldsymbol{Z})P(G_S^* \mid G_S)}\right) \\ &= \min\left(1, \frac{L(w \mid G_S^*, \lambda)\pi(G_S^*)\pi(\lambda)P(G_S \mid G_S^*)}{L(w \mid G_S, \lambda)\pi(G_S)\pi(\lambda)P(G_S^* \mid G_S)}\right) \\ &= \min\left(1, \frac{L(w \mid G_S^*, \lambda)P(G_S \mid G_S^*)}{L(w \mid G_S, \lambda)P(G_S^* \mid G_S)}\right);\end{aligned}$$

b) for the rate λ transition:

$$\begin{aligned}\alpha(\lambda^* \mid \lambda) &= \min\left(1, \frac{p(\lambda^* \mid G_S, \boldsymbol{Z})g(\lambda \mid G_S)}{p(\lambda \mid G_S, \boldsymbol{Z})g(\lambda^* \mid G_S)}\right) \\ &= \min\left(1, \frac{L(w \mid G_S, \lambda^*)\pi(G_S)\pi(\lambda^*)g(\lambda \mid G_S)}{L(w \mid G_S, \lambda)\pi(G_S)\pi(\lambda)P(\lambda \mid G_S)}\right) \\ &= \min\left(1, \frac{L(w \mid G_S, \lambda^*)\pi(\lambda^*)g(\lambda \mid G_S)}{L(w \mid G_S, \lambda)\pi(\lambda)g(\lambda^* \mid G_S)}\right).\end{aligned}$$

This probability is better calculated using the logarithmic transformations, since

$$\log\frac{L(w \mid G_S, \lambda^*)}{L(w \mid G_S, \lambda)} = (n-m)\left(\log(\lambda^*) - \log(\lambda)\right) - \boldsymbol{s}^T w(\lambda^* - \lambda),$$

$$\log\frac{\pi(\lambda^*)}{\pi(\lambda)} = (\alpha_r - 1)(\log(\lambda^*) - \log(\lambda)) - \beta(\lambda^* - \lambda), \text{ and}$$

$$\log\frac{g(\lambda \mid G_S)}{g(\lambda^* \mid G_S)} = \log(\lambda^*) - \log(\lambda) - \frac{1}{2}\left[\frac{(\lambda - \hat{\lambda})^2}{\sigma^2} - \frac{(\lambda^* - \hat{\lambda})^2}{(\sigma^*)^2}\right].$$

Establishing the Metropolis-Hastings, the Gibbs sampling is well-defined.

APPENDIX C – Stan codes

C.1 Perfect tests

This is the Stan code for model (3.1)

```
data {
    int<lower=0> n_samples;
    int<lower=0> n_predictors;

    int<lower=0, upper=1> Y[n_samples];
    matrix[n_samples, n_predictors] X;

    cov_matrix[n_predictors] Sigma;
    vector[n_predictors] mu;
    real<lower=0> alpha_p;
    real<lower=0> beta_p;
}
transformed data {
  matrix[n_predictors, n_predictors] sigma_beta;
  sigma_beta = cholesky_decompose(Sigma);
}
parameters {
    vector[n_predictors] normal_raw;
    real<lower=0, upper=1> prev;
}
transformed parameters {
    vector[n_predictors] effects = mu + sigma_beta * normal_raw;
}
model {
    normal_raw ~ std_normal();
    prev ~ beta(alpha_p, beta_p);
    Y ~ bernoulli_logit(logit(prev) + X * effects);
}
generated quantities {
  vector[n_predictors] effects_prior = multi_normal_rng(mu, Sigma);
  real<lower = 0, upper = 1> prev_prior = beta_rng(alpha_p, beta_p);
}
```

C.2 Sensitivity and specificity

This is the Stan code for model (3.2)

Logit normal prior

```
data {
     int<lower = 0> n_pos;
     int<lower = 0> n_neg;
     int Y_p;
     int Y_n;
     vector[2] mu_gamma;
     cov_matrix[2] Sigma_gamma;
}
transformed data {
    matrix[2,2] sigma_gamma;
    sigma_gamma = cholesky_decompose(Sigma_gamma);
}
parameters {
     vector[2] normal_raw;
}
transformed parameters {
    vector[2] logit_sens_spec;
    logit_sens_spec = mu_gamma + sigma_gamma * normal_raw;
}
model {
     normal_raw ~ std_normal();
     Y_p ~ binomial_logit(n_pos, logit_sens_spec[1]);
     Y_n ~ binomial_logit(n_neg, logit_sens_spec[2]);
}
generated quantities {
    real<lower = 0, upper = 1> sens;
    real<lower = 0, upper = 1> spec;
    sens = inv_logit(logit_sens_spec[1]);
    spec = inv_logit(logit_sens_spec[2]);
}
```

Bivariate beta prior with constant α

```
data {
     int<lower = 0> n_pos;
     int<lower = 0> n_neg;
     int Y_p;
     int Y_n;
     vector<lower = 0>[4] alpha_data;
}
transformed data {
    vector<lower = 0>[3] alpha_sum;
     alpha_sum[3] = alpha_data[4];
     alpha_sum[2] = alpha_data[3] + alpha_sum[3];
     alpha_sum[1] = alpha_data[2] + alpha_sum[2];
}
parameters {
```

```
    vector<lower = 0, upper = 1>[3] Z;
}
transformed parameters{
    real<lower = 0, upper = 1> sens;
    real<lower = 0, upper = 1> spec;
    sens = 1 - Z[1] * Z[2]; //(1 - Z[1]) + Z[1] * (1 - Z[2])
    spec = 1 - Z[1] + Z[1] * Z[2] * (1 - Z[3]);
}
model {
    Z ~ beta(alpha_sum, alpha_data[1:3]);
    Y_p ~ binomial(n_pos, sens);
    Y_n ~ binomial(n_neg, spec);
}
generated quantities {
  real<lower = 0, upper = 1> Z_prior[3];
  real<lower = 0, upper = 1> sens_prior;
  real<lower = 0, upper = 1> spec_prior;
  int Y_p_prior;
  int Y_n_prior;

  Z_prior = beta_rng(alpha_sum, alpha_data[1:3]);
  sens_prior = 1 - Z_prior[1] * Z_prior[2];
  spec_prior = 1 - Z_prior[1] + Z_prior[1] * Z_prior[2] * (1 - Z_prior[3]);
  Y_p_prior = binomial_rng(n_pos, sens_prior);
  Y_n_prior = binomial_rng(n_neg, spec_prior);
}
```

Bivariate beta prior with random α

```
data {
    int<lower = 0> n_pos;
    int<lower = 0> n_neg;
    int Y_p;
    int Y_n;
    vector<lower = 0>[4] a;
    vector<lower = 0>[4] b;
}
parameters {
    vector<lower = 0, upper = 1>[3] Z;
    vector<lower = 0>[4] alpha;
}
transformed parameters{
    real<lower = 0, upper = 1> sens;
    real<lower = 0, upper = 1> spec;
    vector<lower = 0>[3] alpha_sum;
    alpha_sum[3] = alpha[4];
    alpha_sum[2] = alpha[3] + alpha_sum[3];
    alpha_sum[1] = alpha[2] + alpha_sum[2];
```

```
    sens = 1 - Z[1] * Z[2]; //(1 - Z[1]) + Z[1] * (1 - Z[2])
    spec = 1 - Z[1] + Z[1] * Z[2] * (1 - Z[3]);
}
model {
    alpha ~ gamma(a, b);
    Z ~ beta(alpha_sum, alpha[1:3]);
    Y_p ~ binomial(n_pos, sens);
    Y_n ~ binomial(n_neg, spec);
}
```

C.3 Imperfect tests

This is the Stan code for model (3.3)

```
data {
    int<lower=0> n_samples;
    int<lower=0> n_predictors;

    int<lower=0, upper=1> Y[n_samples];
    matrix[n_samples, n_predictors] X;

    cov_matrix[n_predictors] Sigma;
    vector[n_predictors] mu;
    real<lower = 0> alpha_p;
    real<lower = 0> beta_p;
    real<lower = 0> alpha_s;
    real<lower = 0> beta_s;
    real<lower = 0> alpha_e;
    real<lower = 0> beta_e;
}
transformed data {
   matrix[n_predictors, n_predictors] sigma;
   sigma = cholesky_decompose(Sigma);
}
parameters {
    vector[n_predictors] normal_raw;
    real<lower = 0, upper = 1> prev;
    real<lower = 0, upper = 1> sens;
    real<lower = 0, upper = 1> spec;
}
transformed parameters {
    vector[n_samples] p;
    vector[n_predictors] effects;
    effects = mu + sigma * normal_raw;
    p = (1 - spec)
        + (spec + sens - 1)
        * inv_logit(logit(prev) + X * effects + (1/sqrt(tau)) * omega);
}
```

```
model {
    normal_raw ~ std_normal();
    prev ~ beta(alpha_p, beta_p);
    sens ~ beta(alpha_s, beta_s);
    spec ~ beta(alpha_e, beta_e);
    Y ~ bernoulli(p);
}
generated quantities {
    vector[n_samples] theta;
    theta = inv_logit(logit(prev) + X * effects);
}
```

C.4 Imperfect tests and respondent-driven sampling

This is the Stan code for model (3.4)

```
functions {
  /**
  * Return the log probability of a proper (CAR) prior
  * with a sparse representation for the adjacency matrix
  *
  * @param omega Vector containing the parameters with a CAR prior
  * @param tau Precision parameter for the CAR prior (real)
  * @param rho Dependence parameter for the CAR prior (real)
  * @param W_sparse Sparse representation of adjacency matrix (int array)
  * @param n Length of omega (int)
  * @param W_n Number of adjacent pairs (int)
  * @param D_sparse Number of neighbors for each location (vector)
  * @param lambda Eigenvalues of D^{-1/2}*W*D^{-1/2} (vector)
  *
  * @return Log probability density of CAR prior up to additive constant
  */
  real sparse_car_lpdf(vector omega, real rho,
    int[,] W_sparse, vector D_sparse, vector lambda, int n, int W_n) {
      row_vector[n] omegat_D; // omega' * D
      row_vector[n] omegat_W; // omega' * W
      vector[n] ldet_terms;

      omegat_D = (omega .* D_sparse)';
      omegat_W = rep_row_vector(0, n);
      for (i in 1:W_n) {
        omegat_W[W_sparse[i, 1]] = omegat_W[W_sparse[i, 1]]
                                   + omega[W_sparse[i, 2]];
        omegat_W[W_sparse[i, 2]] = omegat_W[W_sparse[i, 2]]
                                   + omega[W_sparse[i, 1]];
      }
```

```
      for (i in 1:n) ldet_terms[i] = log1m(rho * lambda[i]);
      return 0.5 * (sum(ldet_terms)
              - omegat_D * omega + rho * (omegat_W * omega));
  }
  real gumbel_type2_lpdf(real tau, real lambda){
    return -(3.0/2.0 * log(tau) + lambda / sqrt(tau));
  }
  real gamma_a_lpdf(real tau, real M_sigma){
      return -2.0 * log(tau);
  }
}
data {
    int<lower=0> n_samples;
    int<lower=0> n_predictors;

    int Y[n_samples];
    matrix[n_samples, n_predictors] X;

    cov_matrix[n_predictors] Sigma;
    vector[n_predictors] mu;
    real<lower = 0> alpha_p;
    real<lower = 0> beta_p;
    real<lower = 0> alpha_s;
    real<lower = 0> beta_s;
    real<lower = 0> alpha_e;
    real<lower = 0> beta_e;
    real<lower = 0> alpha_rho;
    real<lower = 0> beta_rho;

    real<lower = 0> alpha_tau;
    real<lower = 0> beta_tau;
    real<lower = 0> lambda_tau;
    real<lower = 0> M_sigma;
    int<lower = 0, upper = 2> tau_prior;

    matrix<lower = 0, upper = 1>[n_samples, n_samples] adj_matrix;
    int adj_pairs;
}
transformed data{
  int adj_sparse[adj_pairs, 2];    // adjacency pairs
  vector[n_samples] D_sparse;      // diagonal of D
  vector[n_samples] lambda;        // eigenvalues of invsqrtD * A * invsqrtD
  matrix[n_predictors, n_predictors] sigma;
  real max_lambda;
  real lower_bound_tau;

  { // generate sparse representation for A
  int counter;
```

```
  counter = 1;
  // loop over upper triangular part of A to identify neighbor pairs
    for (i in 1:(n_samples - 1)) {
      for (j in (i + 1):n_samples) {
        if (adj_matrix[i, j] == 1) {
          adj_sparse[counter, 1] = i;
          adj_sparse[counter, 2] = j;
          counter = counter + 1;
        }
      }
    }
  }
  for (i in 1:n_samples) D_sparse[i] = sum(adj_matrix[i]);
  {
    vector[n_samples] invsqrtD;
    for (i in 1:n_samples) {
      invsqrtD[i] = 1 / sqrt(D_sparse[i]);
    }
    lambda = eigenvalues_sym(quad_form(adj_matrix, diag_matrix(invsqrtD)));
    max_lambda = max(lambda);
  }
  sigma = cholesky_decompose(Sigma);
  if (tau_prior == 2){
      lower_bound_tau = 1/M_sigma;
  }else {
    lower_bound_tau = 0;
  }
}
parameters {
    vector[n_predictors] normal_raw;
    real<lower = 0, upper = 1> prev;
    real<lower = 0, upper = 1> sens;
    real<lower = 0, upper = 1> spec;
    real<lower = 0, upper = 1/max_lambda> rho;

    vector[n_samples] omega;
    real<lower = lower_bound_tau> tau;
}
transformed parameters {
    vector<lower = 0, upper = 1>[n_samples] p;
    vector[n_predictors] effects;
    effects = mu + sigma * normal_raw;
    p = (1 - spec)
        + (spec + sens - 1)
        * inv_logit(logit(prev) + X * effects + (1/sqrt(tau)) * omega);
}
model {
    if (tau_prior == 0){
```

```
        tau ~ gamma(alpha_tau, beta_tau);
    } else if (tau_prior == 1) {
        tau ~ gumbel_type2(lambda_tau);
    } else {
        tau ~ gamma_a(M_sigma);
    }

    rho ~ beta(alpha_rho, beta_rho);
    omega ~ sparse_car(rho, adj_sparse, D_sparse,
                       lambda, n_samples, adj_pairs);

    normal_raw ~ std_normal();
    prev ~ beta(alpha_p, beta_p);
    sens ~ beta(alpha_s, beta_s);
    spec ~ beta(alpha_e, beta_e);

    Y ~ bernoulli(p);
}
generated quantities {
    vector[n_samples] theta;
    theta = inv_logit(logit(prev) + X * effects + (1/sqrt(tau)) * omega);
}
```