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

Contents

Todo list

List of sections to revise

1. Respondent-driven sampling;
2. Add Hierarchical modelling chapter;
3. Should I add a subsection in Bayesian Statistics revising Prevalence estimation models using Bayesian paradigm?

What to do after?

1. Notes about Bivariate Beta;
2. Study case about CAR models in bernoulli aspect.

1 Introduction

Hidden or hard-to-reach populations have two main features: no sampling frame exists, given that their size and boundaries are unknown, and there are privacy concerns because the subjects are stigmatized or have illegal behavior (HECKATHORN, 1997). Fear of exposition or prosecution complicates the enumeration of the populations and the learning about them. Moreover, if the occurrence frequency of the condition 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.

Research has been carried out with the development of some 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). (HECKATHORN) introduced the Respondent-Driven Sampling (RDS) to fill some gaps from other methods 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.

Following the sampling from the target population, a questionnaire or a disease test is conducted. This work considers binary outcomes. For instance, asking about smoking status or testing for HIV infections. However, the diagnoses are subject to measure error, and regard their accuracy is a vital step (REITSMA et al., 2005). One common way to do this is to measure jointly *sensitivity* and *specificity*. The former is the ability to detect the condition, while the latter to identify the absence of it.

Nevertheless, because of our lack of knowledge about Nature itself, it is necessary to model the uncertainty of this process, and Bayesian Statistics is the indicated area of study. In the Bayesian paradigm, the parameters are random variables, and the beliefs about them are updated given new data. The idea is to propagate uncertainty about the outcome through the network of contacts, which has its probability distribution.

This work proposes to study the survey method Respondent-Driven Sampling (RDS), a chain-referral method with the objective of sampling from hard-to-reach

populations when necessary to estimate the prevalence of some binary condition from this population. The modeling also accounts for sensibility and sensitivity since the imperfection of the detection tests. We also intend to apply this framework efficiently, comparing Monte Carlo algorithms and Laplace approximations.

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, the epidemiologists use some *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 represents 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, its evolution 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 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. 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 everyone in the population, it is necessary to random 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. 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.

Provide some reference

A perfect test would discriminate every sick individual from the non-sick ones. Given that there is not 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 golden standard being negative;

d) false negative (FN): when the test under evaluation has a negative diagnose, despite the golden 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. For more details, 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 focus 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. Observe that this correlation is related to the estimation of the parameters, then it can only be noticed in meta-analysis studies.

2.2 Respondent-driven sampling

Respondent-driven sampling (RDS) is a procedure developed by Heckathorn (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's proposal (1997) 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 peers. 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.

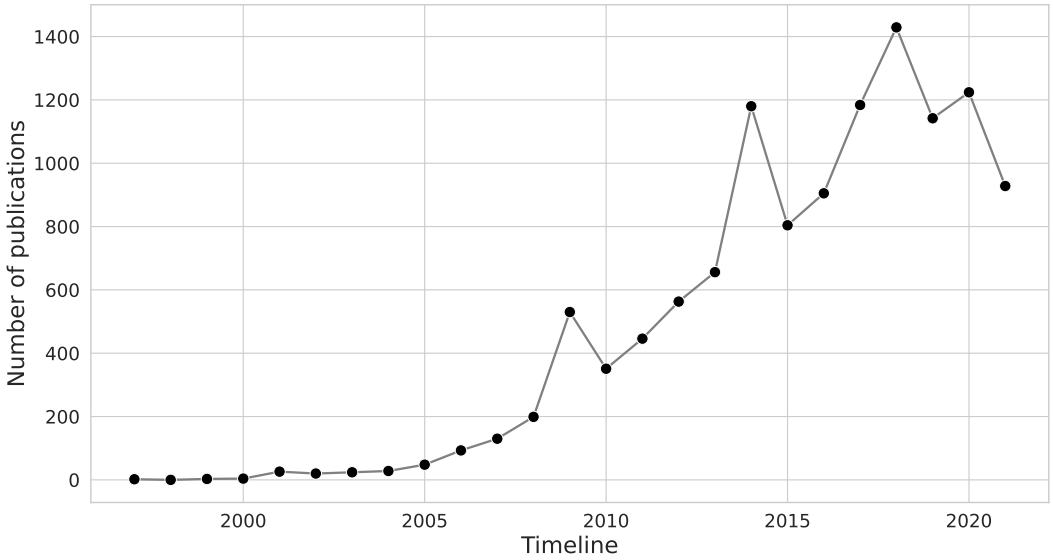
According to Gile, Beaudry, et al. (2018, p. 66), there are two main advantages of RDS over other snowball samplings. First, the fixed number of recruitment coupons enforces the network gets deeper and distant from the seeds, which reduces the dependence of the final sample from the initial chosen by 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 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 biological and behavioral surveillance 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

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.

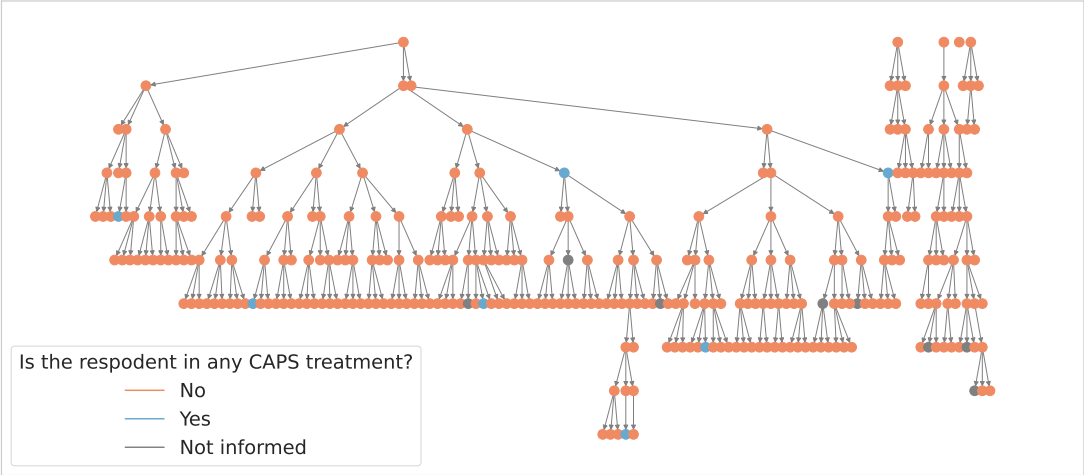
survey. According to it:

> 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 can not 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 musta 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 inferencial 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

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.

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 participant.

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 reaching 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 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) requires 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 to 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) respondent-driven sampling 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 contraction 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.

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-probability 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 obliged. The degree of a person is, therefore, $d_i = \sum_{j=1}^{N} A_{ij}$.

Include notation of RDS used posteriorly.

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 to analyse the structure created by the recruitment links. They deeply discussed the problems that bottlenecks can cause. McLaughlin (2021) developes a Bayesian model for the recruitment process considering preferential selection based on the covariates.

2.2.3.1 First-order Markov process

This approximation was the first model 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 derives three theorems:

Theorem 2.2.1 (Convergence to equilibrium)**.** *Let $\{Z_n\}_{n \in \mathbb{N}}$ be the recruitment process. Given that the space 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 the geometric convergence. □

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 with 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 argues 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. 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

Definition 2.2.3 (Coupon matrix)**.** The *coupon matrix* $C \in \{0, 1\}^{n \times n}$ 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 $\boldsymbol{Z} = (G_R, \boldsymbol{d}, \boldsymbol{t}, C)$.

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 $\boldsymbol{Z}$.

Provide an example to explain all the above definitions.

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 over 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. A Metropolis-within-Gibbs sampling scheme is used to draw pairs (G_S, λ). A simulated anneling procedure can be used to obtain a sequence that converges to the maximum a posteriori. 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 (SH) RDS estimator*: 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 $N_1 = N\theta$. 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

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

and that

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

is the prevalence estimator, such that $\hat{d}_0, \hat{d}_1, \hat{c}_{01}$, and $\hat{c}_{10}$ are estimated for these quantities.

c) *Volz-Heckathorn RDS (VH) estimator*: 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 (SS) estimator:* 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 account the finite population effect. If the sampling fraction is small, this estimator is similar to VH estimator. Otherwise, when it grows, VH is biased. The limitation of 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.

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:

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

In 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-posteriori 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) performs a binary regression to prevalence estimation through a hierarchical model where correlation structure was modelled as a Conditionally autoregressive (CAR) model (see Section XXX). 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.

2.2.6 Bootstrap methods for uncertainty quantification

(BARAFF; MCCORMICK; RAFTERY, 2016), (SALGANIK, 2006).

2.2.7 Diagnosis of RDS

(GILE; JOHNSTON; SALGANIK, 2015)

2.3 Modelling strategies

In this section, we briefly describe some modelling strategies used throughout the dissertation.

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 $\mu = \boldsymbol{X}\beta$, 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 for each Y_i is a possible variation for this model.

Generalized linear models (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 distribution (ROBERT, 2007, p. 115).

When Y_i has Bernoulli distribution with probability of success $\mu \in (0, 1)$, 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 inferencial procedure for binary response, such as having or not a disease.

2.3.2 Conditionally autoregressive models

The *Conditionally Autoregressive* (CAR) 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). We have that $\lambda_{\min}^{-1} < 0 < \lambda_{\max}^{-1}$, then this interval is not empty.

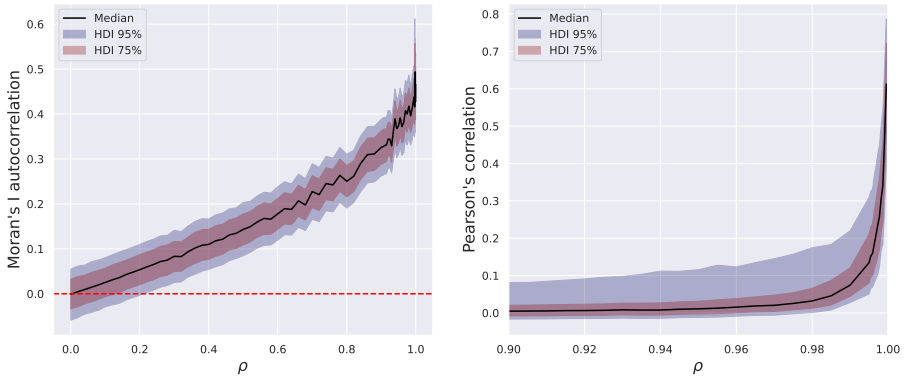
For many applications, CAR reproduces the strong spatial correlation between neighbors only when ρ is close to the limits. Moreover, its interpretation is not so clear. To verify this fact, we generate a random matrix $\tilde{B}$ with binary entrances and adjust $B = 0.5(\tilde{B} + \tilde{B}^T)$ yielding 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 3 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 generate 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.

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).

Figure 3 – 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.

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. The recent development of new computational tools brought these ideas again.

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 out the parameter conditioned All 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

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

Another important concept used thorough out 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 $f(y \mid \theta)$ 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

2.5.1 Hamiltonian Monte Carlo

We follow (BETANCOURT, 2017). This method was developed in the late 1980s as Hybrid Monte Carlo to tackle calculations in Lattice Quantum Chromodynamics. Instead of moving in the parameter space randomly with uninformed jumps, the direction from the vector field given by the gradients are used to trace out a trajectory through the *typical set*, the region which has significant contribution to the expectations. However, if only the gradient was used, the trajectory would pull towards the mode of the distribution, so more geometric constraints are needed. In order to a satellite rotate around the Earth, we have to endow ir with enough momentum to counteract the gravitational field, turning the system into a conservative one.

First, we introduce auxiliary momentum parameters p_n (lift) of the same dimension from the parameter space $\Omega \subseteq \mathbb{R}^D$. Then q_n turns to (q_n, p_n), with the use the joint probability distribution $\pi(q, p) = \pi(p \mid q)\pi(q)$. Particularly, we use

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

such that H is the *Hamiltonian*. Note that $H(q, p) = -\log \pi(p \mid q) - \log \pi(q) =: K(p, q) + V(q)$. We call K the kinetic energy, and V the potential energy. The vector field is generated by Hamilton's equations,

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

Therefore, we are able to define the Hamiltonian flows $\phi_t : (p, q) \to (p, q), \forall t \in \mathbb{R}$.

2.5.1.1 Diagnostics

The importance of diagnosing. The potential problems that it can show.

- Divergent transitions;
- Transitions that hit the maximum tree depth;
- Low E-BFMI values;

- Low effective samples sizes;
- $\hat{R} \notin (0.95, 1.05)$.

2.5.2 Metropolis-within-Gibbs

If this method is used.

3 Prevalence modelling and regression methods

Fisher (1922, p. 311) stated that the objective of statistics is to reduce the data since its volume is impossible to comprehend by the researchers. In that sense, few parameters should represent the whole phenomenon catching the most relevant information. Years later, Newman 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 to diagnose. 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 written in Appendix B. For plotting the diagnosis and the distributions, Arviz (KUMAR et al., 2019) and Matplotlib (HUNTER, 2007) Python packages were used.

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 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 $\boldsymbol{X}$ the covariate matrix.

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

A formal definition for identifiability 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);

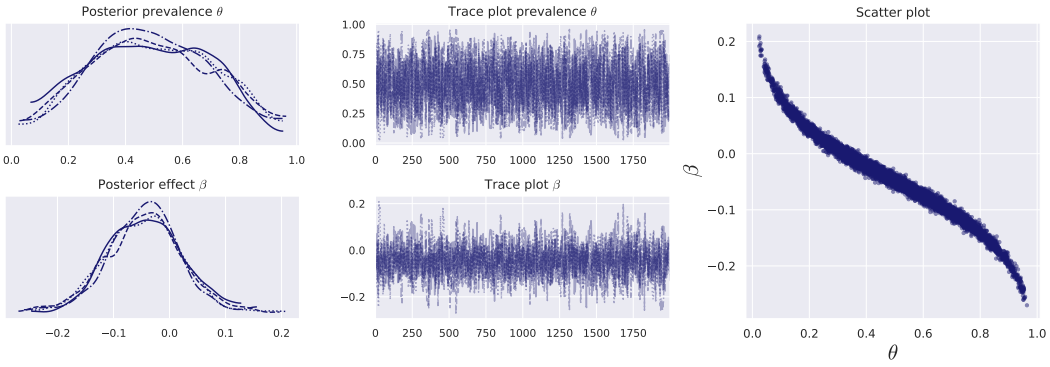
(v) make 1000 warm-up and 1000 sampling iterations using Stan given the data $(Y_1, X_1), \ldots, (Y_n, X_n)$.

(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 4 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 5.

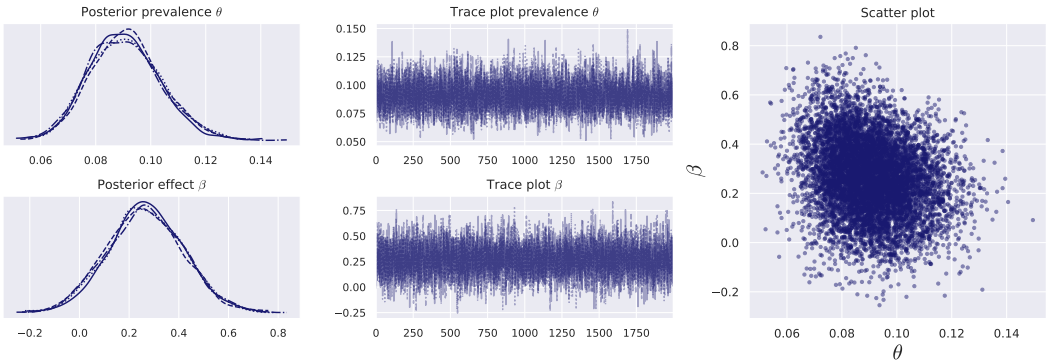
We observe that the interpretation of prevalence from Remark 3.1.1 changes from centred and uncentered since the meaning of $\boldsymbol{x}_i = 0$ is different. Along with

Figure 4 – 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).

Figure 5 – 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).

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 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.

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 parameters.

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

Exp	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

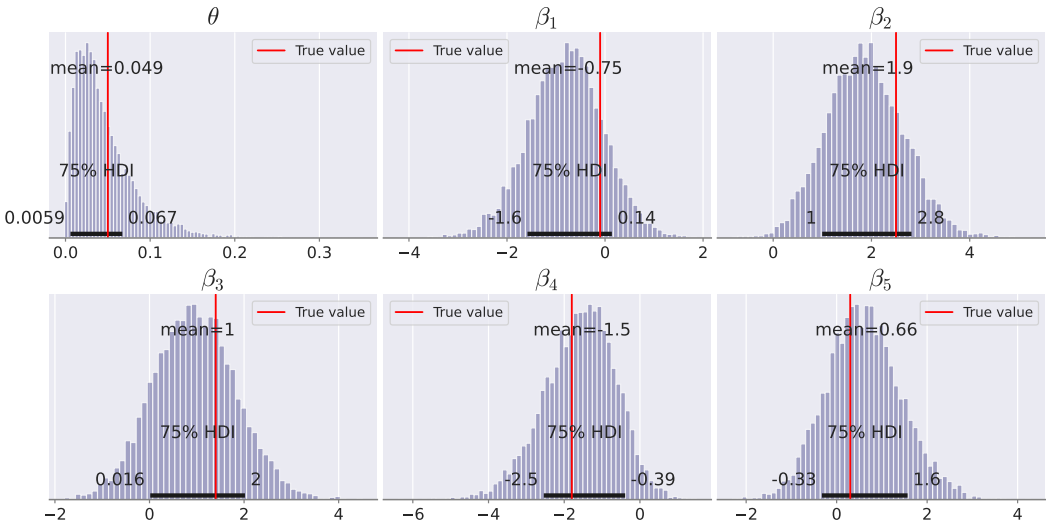
Source: Prepared by the author (2021). We denote n for number of samples, k_c for the number of continuous variables, and k_b for binary variables. Between paranthesis, *normal* means that the variables were generated from a Multivariate Normal with prespecified parameters, and *cauchy* from a Cauchy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

We primally look at the settings from experiment 1. With a non-informative prior for θ (Jeffreys prior Beta$(1/2, 1/2)$) and a weakly informative for β (zero mean and covariance matrix four times the identity matrix), Figure 6 shows the posterior distributions for the parameters. The prevalence estimate is good despite Jeffreys' 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 was greater than 3000 for all parameters, while Tails ESS were greater that 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 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.

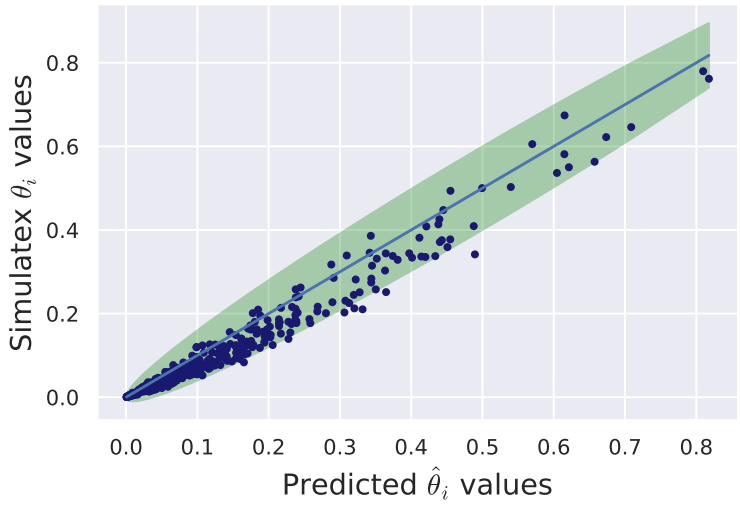
Experiment 2 is used to see if these properties repeat when the prevalence is higher. The same regressors were used for the comparison, but the input data Y were generated with different prevaleces. With prevalence being 0.9, the estimates

Figure 6 – Posterior distribution for parameters of model (3.1).



Source: Prepared by the author (2021). The red line represents the true value inputed for the simulation.

Figure 7 – Comparing predicted and simulated probabilities of having the disease from model (3.1).

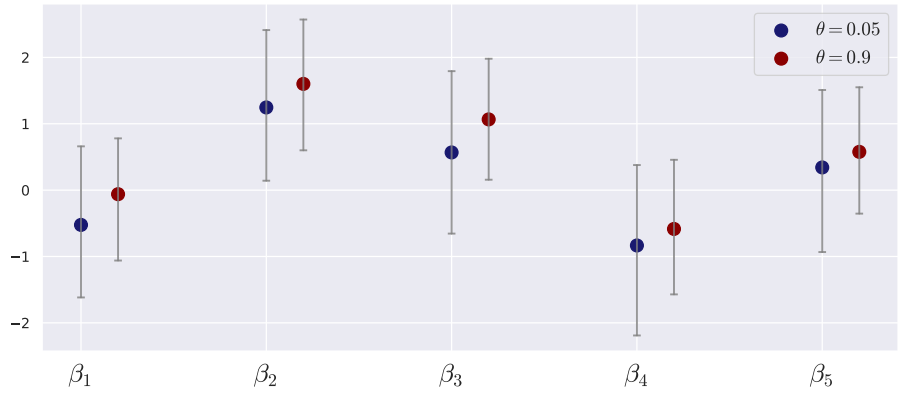


Source: Prepared by the author (2021). The green area is delimited by the curves generated by $2\sqrt{\theta_i(1-\theta_i)/n}$, where $n = 500$ is the number of points. This area is a proxy for ± 2 standard-error bounds.

were a little high for all coefficients as Figure 8 presents. This is related to the fact that the posterior mean underestimated the true value for this experiment. After increasing the number of samples, the estimates were closer, as expected.

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

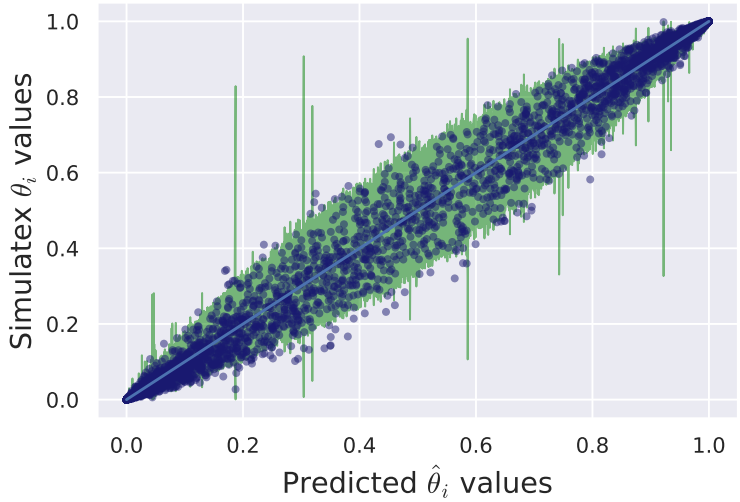
Figure 8 – Comparing posterior mean and 94% credibility intervals for β in model (3.1) with the same regressors $\boldsymbol{X}$ but different prevalences.



Source: Prepared by the author (2021).

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 3 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 9 presents how the predicted probabilities for each individual behaves in this case.

Figure 9 – Comparing predicted and simulated probabilities of having the disease from model (3.1) with high dimension.



Source: Prepared by the author (2021). 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. In the Two-by-two formulation from Chart 1, we have

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).

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:

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 $\alpha_1, \ldots, \alpha_4 \in \mathbb{R}_{>0}$.

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

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

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

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

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 that prior approach is that the moments of logit normal distribution are not in closed form and there is no available formula to derive $\mathbb{E}[\gamma_s]$ 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 research have prior information about specificity and specificity, such their mean and correlation.

To specify the prior hyperparameters using prior information, Section A.3 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 $b_i = \hat{\alpha}_i/\operatorname{Var}(\alpha_i)$ 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.

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.

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.

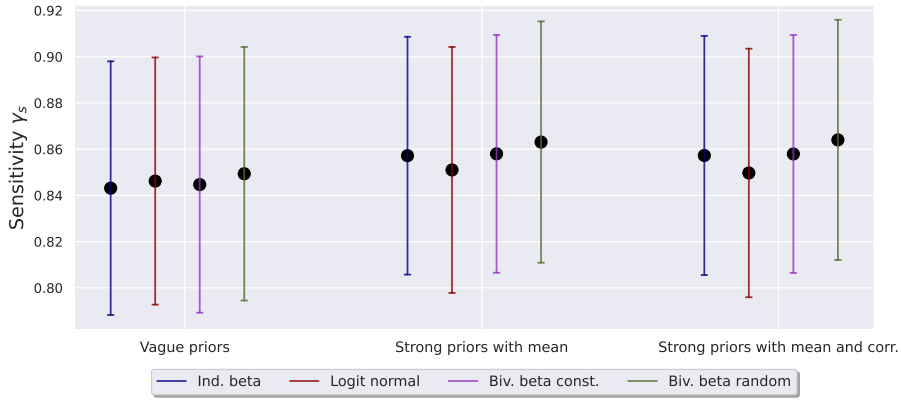
For each situation and each dataset, it was drawn 2000 samples from the posterior distribution and the HDI 75% interval and posterior mean were calculated. The Hits column counts the percentage of the times that the true values lied in the interval, while the MSE column calculates the mean squared error of the posterior mean with respect to the true value. We notice that the fourth prior approach had

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

Situation	Prior approach	Hits		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 10 – 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).

a little number of effective samples when compared to the other methods. Notice that 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 10 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.

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. This is an extension of model 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 no 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 can not generally offset this difference, and identifiability can not 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 not 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.

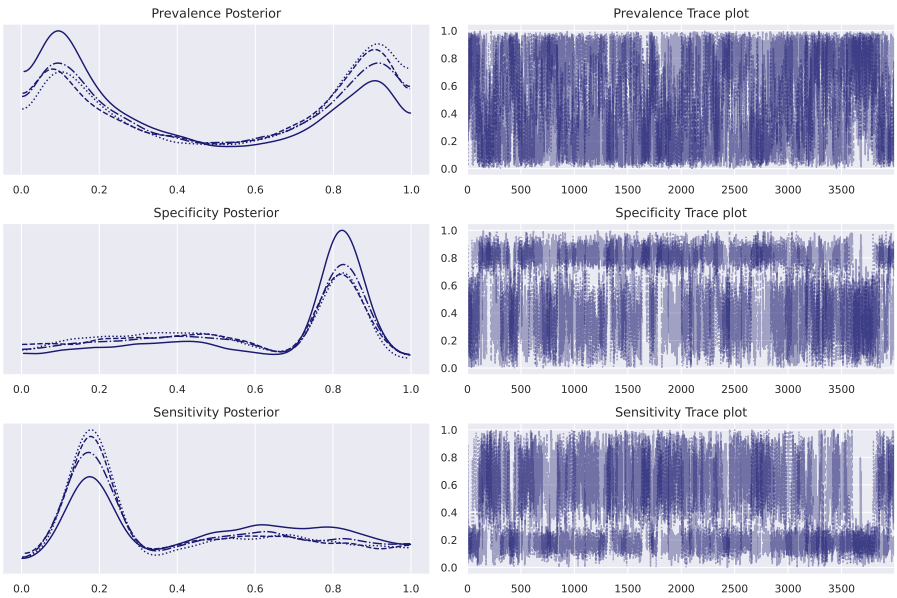
Table 4 – Results from HMC algorithm for the practical identifiability analysis in model (3.3).

	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.

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

Figure 11 – 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).

Exp	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

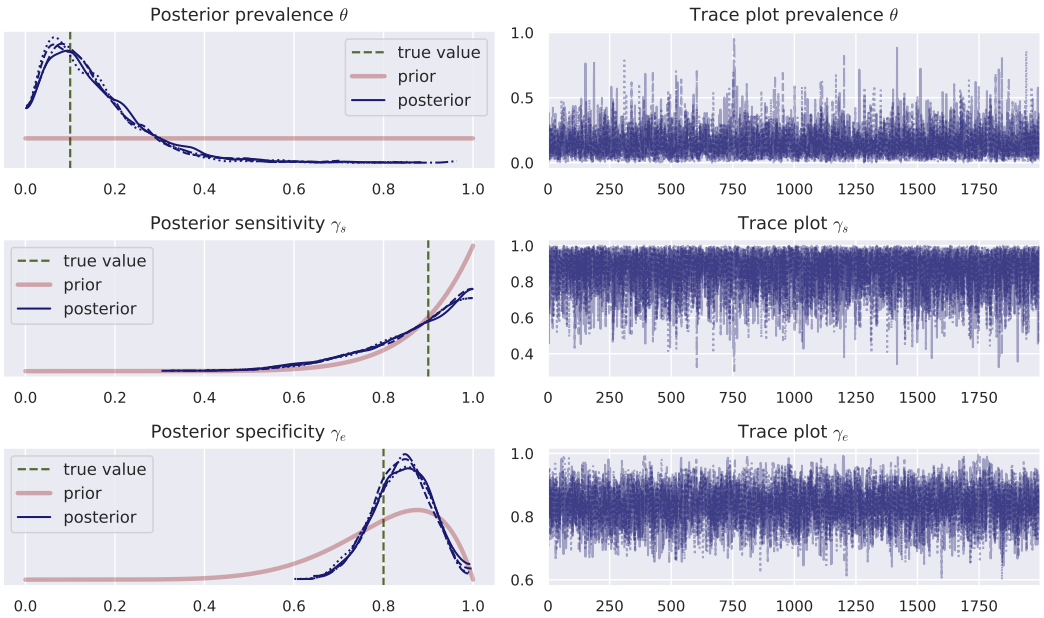
Source: Prepared by the author (2021). We denote n for number of samples.

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 diagnosis from HMC were good. Figure 12 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) concludes in its application that "uncertainty in the population prevalence is in large part driven by uncertainty in the specificity." However we did not noticed this effect in this model. Figure 13 presents the resultant scatter plot of the posterior simulations. Here we observe that all parameters drive prevalence uncertainty.

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 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 $\theta, \gamma_s, \gamma_e, \beta_1, \ldots, \beta_5$.

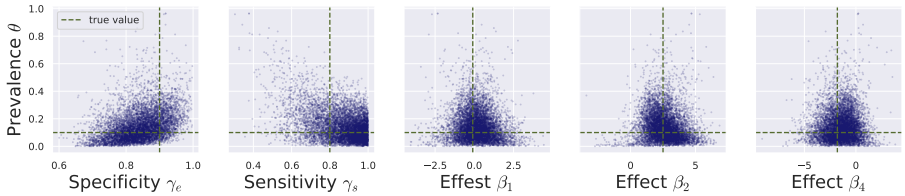
The second experiment considers the case where the number of samples is not so high, but prevalece is low. It contrasts with the third experiment where many more samples are obtained. Figure 14 presents the differences. Notice that the uncertainty

Figure 12 – 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.

Figure 13 – Scatter plot of the posterior simulations of prevalence, specificity, sensitivity and effects of model (3.4)

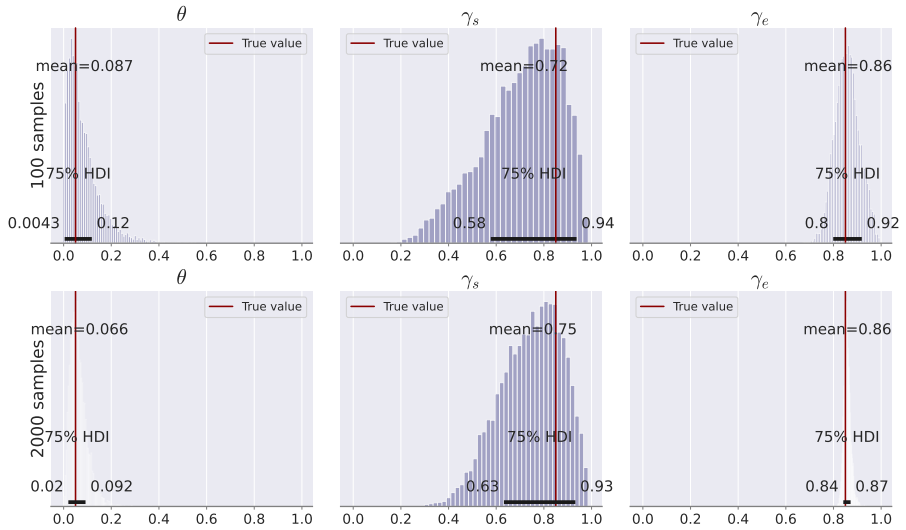


Source: Prepared by the author (2021) from the Stan sampling result.

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,

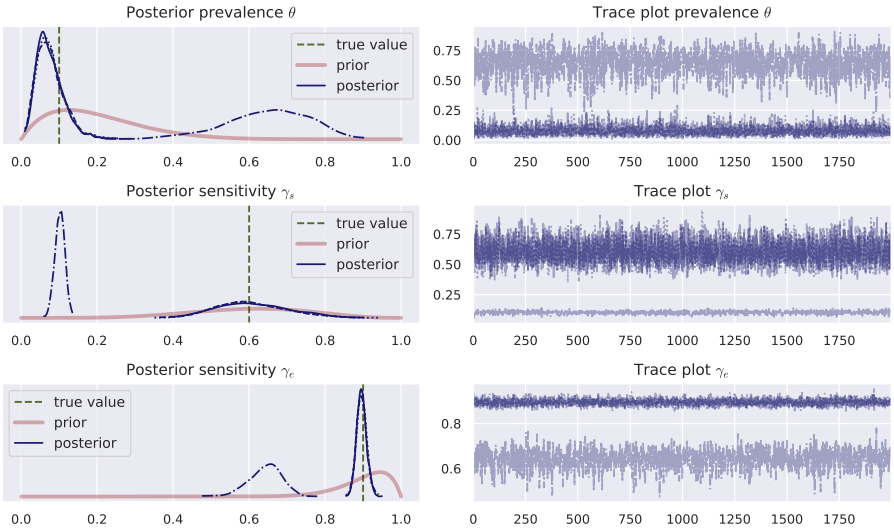
The fourth and fifth experiments compare two opposite situations. The former sets a low sensitivity and a high specificity, while the latter has high sensitivity and low specificity. These examples are convenient for detecting the problem with identifiability in high dimensional data. Specifying the hyperparameters with $\alpha_p = 2$, $\beta_p = 8$, $alpha_s = 6$, $\beta_s = 4$, $\alpha_e = 18$, $\beta_e = 2$, with 2000 iterations for warmup the resulting posterior and trace is given by Figure 15. Notice that one of the chains was very far from the tru value driving a very high $\hat{R}$. Although the result seems

Figure 14 – Posterior distribution for θ, γ_s and γ_e from model (3.4)



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

Figure 15 – 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.

awkward, it makes sense under the identifiability problem, since the chain produces a very similar probability of positive test. A more strong 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 can not 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 Conditionally autoregressive (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:

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

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 has several discussions 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) proves 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) suggests 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 satisfies 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.

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.

Vectorization of the variables

Non-centered parameterization

Efficient implementation

Scaling the precision in the efficient implementation

3.4.3 Simulated data

a) More sparse matrices (RDS data is very sparse) is generating the funil we do not want to see. This is not connected to the number of connected components. In order to see that, a simple example with the Erdos-Renyi Random Graph can answer to us. In the sparse case, the number of edges

is $O(n)$ with $p = 1/n$. If $p = 1$, the number of edges is $O(n^2)$ and the funil disappears. This problem does not appear in the poisson model.

3.4.4 Including uncertainty about the recruitment graph

3.5 Model extensions

Several characteristics of RDS were not include in the previous model, such as homophily, bottlenecks, and sampling weights. This section aims to build some options for these aspects and establish future works in that line.

a) *Homophily model*: (YAUCK et al., 2021)

b) *Sampling weights*: GLM weighted

c) *Bottlenecks*

3.6 Mispecified data simulation

4 Discussion about prior distributions and sensitivity analysis

4.1 Prior analysis of sensitivity and specificity

4.2 Prior analysis on the parameter tau

4.3 Prior analysis on theta

5 Real data applications

6 Conclusion

Parte final do trabalho, apresenta as conclusões correspondentes aos objetivos ou hipóteses.

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.

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.

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.

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.

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.

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>.

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.

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.

GELMAN, Andrew. 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.

_____. 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. 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>.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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>.

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 $v \in \mathbb{R}^n$ as

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

Definition A.1.1. Let

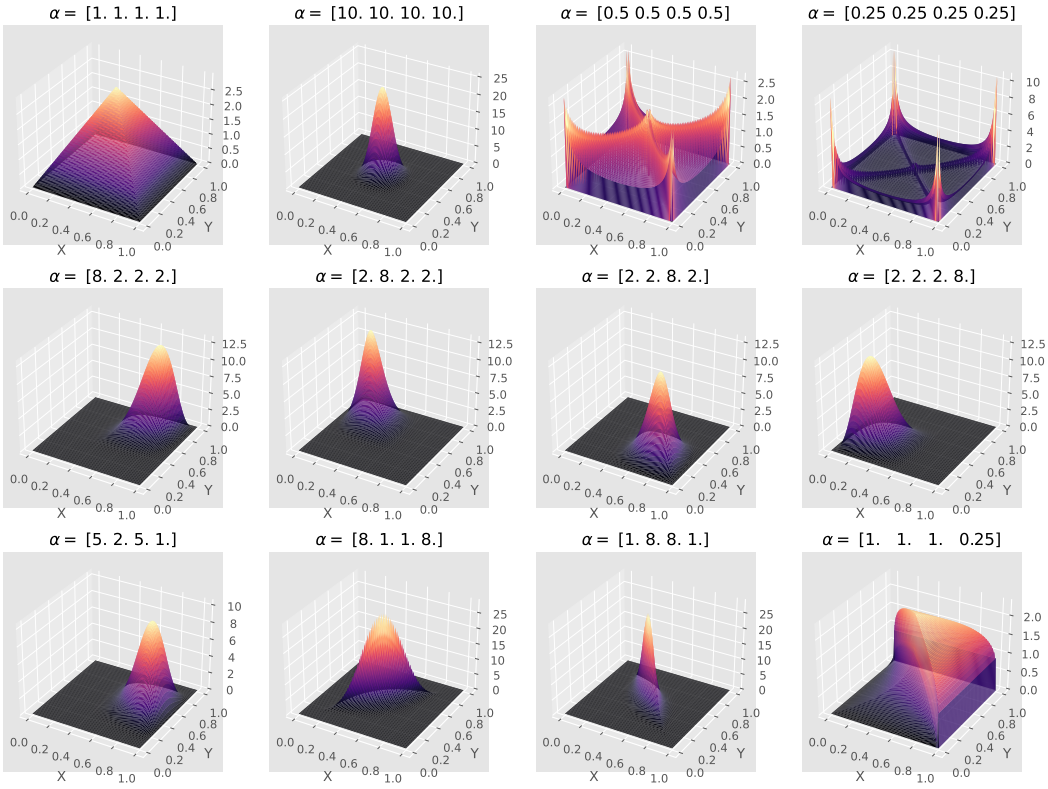
$$X = U_1 + U_2 \text{ and } Y = U_1 + U_3. \tag{A.2}$$

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

Figure 16 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. Their proofs are in the Appendix.

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 $\alpha_2 + \alpha_4$.*

Figure 16 – 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 $\alpha = [0.5, 0.5, 0.5, 0.5]$

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

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

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

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

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 $0 \le u_1 \le x, u_1 \le y$, and $0 \le 1 - x - y + u_1$. 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 Comments about integration

The density of (X,Y) is $f_{X,Y}(x,y)$ as in equation (A.11). Therefore 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:

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

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.3 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,

$$v_1 = \frac{m_1(1-m_1)}{\tilde{\alpha}+1}$$
$$v_2 = \frac{m_2(1-m_2)}{\tilde{\alpha}+1}$$

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.3.1. *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:

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

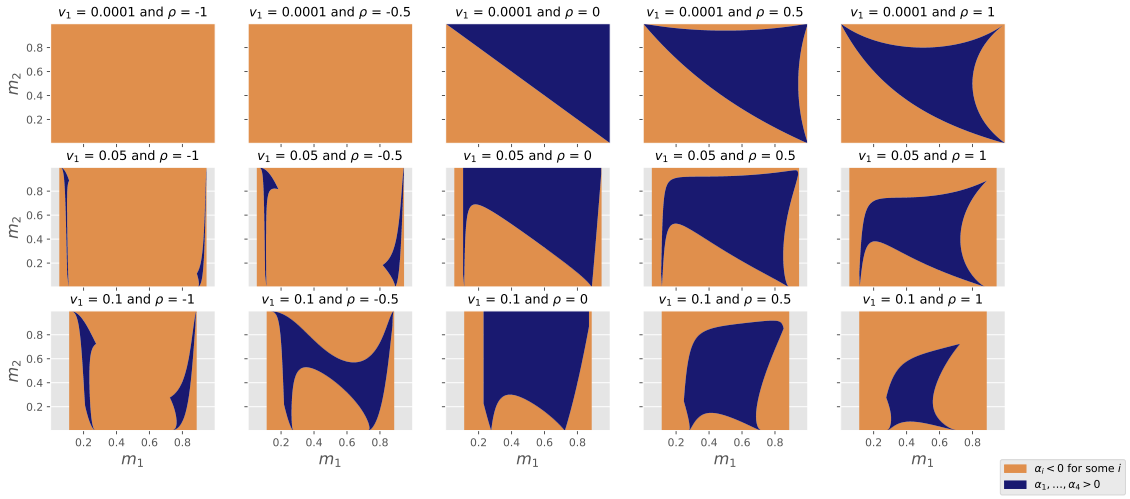
then

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

Besides solving the system (A.15), the bivariate beta distribution needs that $\alpha_1, \ldots, \alpha_4 > 0$. However, this is not necessarily true. 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 17. 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 17 – 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).

Because of that reason, 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 denominador 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 denominador 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.3.1 and Figure 17, 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 distanced 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 more beliefs in the means than the other moments.

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

Remark A.3.1. If the researcher has information about a credibility interval, this information needs to be converted in terms of the variance.

A.4 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.3.1), (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 6 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 6 – 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 – Stan codes