BAYESIAN EVIDENCE SYNTHESIS:OPIOID CRISIS

HYEONGCHEOL PARK* & PAUL GUSTAFSON* & MICHEAL A IRVINE*[1]

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

* *Department of Statistics, University of British Columbia, Vancouver, Canada*
[1] *Department of Mathmatics, University of British Columbia, Vancouver, Canada*

1 INTRODUCTION

All examples here were performed in Python 3.7 using the library pyMC [reference] and JAGS. Training was performed using No U-Turn Sampling (NUTS) over two chains with 10000 iterations and initial conditions for the chain were selected through the use of ADVI. Fitting was performed on a GHz Intel Core i5 with 8GM of LPDDD3 RAM and typically had wall times under ten minutes. Data processing was carried out using the Pandas and SciPy library [reference] Data visualization was performed using the libraries Seaborn and Matplotlib [ref]. Code for all examples in this study are provided.

1. Intro of Intro: Early Writing (Below from tamplete) opioid crisis is one of major issues in North America continents including Canada. There were 1,490 deaths and 15,598 paramedic- attended overdose events during 2017 alone. [1] (need to know about bib in latex, change statistics to 2018 later) The goal of this project is to apply Bayesian evidence synthesis to help reduce the effect of opoid crisis in Vancouver, Canada.

   There are multiple sources of data sets.

   We estimates the number of drug addicts. A statement requiring citation [2].

2. Project Goal: scrapped from e mails

   <Frame>

   This project is about the area of Bayesian evidence synthesis methodology oriented toward applications in public health and epidemiology. The project will involve one or both of the following lines of enquiry:

   1. Developing new evidence synthesis models to address specific public health challenges.

   2. Developing theoretical understanding of what a priori assumptions are required to yield partial or full identification of target quantities in prototype evidence synthesis models.

   Mike builds Bayesian evidence synthesis models for public health and epidemiology applications. Typically in public health many aspects of the system of study remain hidden and surveillance data is derived from downstream effects. (e.g. think of a disease that is symptomless where only diagnoses are observed) These models also require flexibility including incorporating dynamic rates and geographic variation through the use of hierarchical components. Some examples include:

   1. Assessing the impact of the take home naloxone program on the number of overdose deaths in BC.

   2. Estimating the number of overdoses occurring within a given month and region based on ambulance call-outs, hospitalizations and other surveillance data.

   3. Estimating the size of the population with substance use disorder, opioid use disorder, individuals who inject opioids etc.

   4. The prevalence of sexually-transmitted infections based on diagnosis and testing data, combined with sexual behaviour survey data.

   Generally there are scientifically relevant questions of "information flow" in such models, very roughly put: how good do the inputs needs to be in order to get useful outputs. And the inputs could be multiple datasets plus multiple prior distributions.

   An example of this can be seen when estimating the total number of overdoses occurring based on different data sources. If an individual overdoses there is

some associated probability that it is witnessed by either a passerby or emergency services. There's a further probability an ambulance may be called or the individual is transported to hospital. Data on ambulance-attended overdoses, ED visits, and other uses of intervention each provide different information about this underlying process and can help to refine the total estimate.

More specifically, it may not be obvious how the width of the posterior distribution on a given target parameter depends on widths of the prior distributions on various parameters plus the amount of data in each data source. But it would be scientifically relevant to know this. For example, estimating populations at risk are required for service planning. An estimate with large uncertainty (wide posterior width) would most likely not be useful.

Of course one way to address this question is in a simulation context. Just try changing the sample sizes and the prior specification, and see what happens. And this project would likely involve some of this.

But hopefully though there is also scope to get some clean and general answers at the cost of working with stripped-down, simplified versions of models.

A specific avenue to investigate is whether any simplified models are amenable to the sort of "partial identification" analysis that Paul has used in other contexts. Likely the most accessible intro to this is Chapters 1 and 2 of Paul's recent book:

Another aspect of this project can be to develop model assessment, validation, and selection within the context of partial identification. Previous work within the field of public health has used information criteria such as the Deviance Information Criterion to assess model fit and select between candidate models. Understanding the evolution of these criteria under different data sources, increased number of observations, and changes to the structure of the priors would also be valuable to investigate.

The project is at a very early stage, quite nascent and amorphous. So the RA work is on trying to flesh out more specific research plans. For instance, can we identify which of Mike's models, and with what simplifications, are good candidates for further study?

3. Scratch Paragraphs: initial writing attempts We have certain goals

   1. developing new evidence synthesis models for a certain public health challenges

   2. developing theorietical understanding of what a priori assumptions are required to yield identification of target quantities

   3. Estimate some possible variables of interest: The number of overdoses, the number of infected people and so on.

   4. Information flow: how good should the input data be in order to have fairly good enough output?

   5. how priors and data set affect the posteriors of interest?

      - 5.a. by simulation in complicated context
      - 5.b. in simplified model, some general understanding mathmatically.

   6. Model assessment, selection and validation: Paul's partially identified model application

   7. For further study, which of Mike's models are good with some simplifications?

Process

The number of overdose is our interest of estimation. Let O_t the number of overdose in a given month t. Suppose there was a survey conducted to see the proportion of ambulance call p_A among the subjects of overdose. Let n_A the sample size of the survey and x_A to be the total number who confirmed they did call an ambulance. Then x_A is assumed to follow Binomial distribution as follows.

$$x_A \sim Bin(n_A, p_A)$$

The total overdoses need to be modeled. The simplest conceptual model is to take an underlying log-rate z_t that is independent and identically distributed according to a normal distribution with mean μ and variance σ^2. [1] Denote λ_t the rate of overdose at time t. It is assumed that the total overdose O_t follows Poission distribution where the population of the region of interest is N.

$$z_t \sim N(\mu, \sigma^2)$$
$$\lambda_t^{OD} = \exp(z_t)$$
$$O_t \sim Poi(\lambda_t^{OD} N)$$

Notice that O_t cannot be estimated directly since none of the variables determining O_t (μ, σ, N) is known. Henc O_t should be inferred from p_A, the ambulance call out rate.

It is assumed that the data of ambulance-attended overdoses is given. Let U_t the number of ambulance-attended overdoses at a time point t. U_t follows Binomial distribution:

$$U_t \sim Bin(O_t, p_A)$$

Now O_t can be estimated since p_A is infered by survey data and the data U_t is given. A simple model was suggested. The model combines Ambulance Call-outs Model and Overdose Model.

The next step is to run some simulations and to figure out how different types of inputs lead the changes of output. To do so, simplest model is proposed as a start point.

Simulation

The first simulation simplifies variables as much as possible. For simplicity, assume $N = 10000, n_A = 1000$. Later the total number of population for a region N could vary over time and also be staritified for better realization of the real world. Likewise, n_A can be vary and it is recommended to see the impact of the likelihood over the posterior distributions of variables of interest.

Likelihood

The two data set (survey data set and ambulance attended overdose) is simulated. For the survey data, the true value was set $p_A = 0.8$. For twelve months (t=1,2,3, ..., 12), x_t values are independentally generated.

Simulation for the overdose model assumed the true values of parameters $\mu = \log 0.05, \sigma = 1$. Likewise twelve samples of O_t and U_t were generated.

Note that only U_t and x_t are known as likelihood. Also $N = 10000, n_A = 1000$ are given for now. The goal is to estimate p_A and O_t.

Prior Distributions

Noninformative prior distributions are presumed as a beginning for simplicity.

$$p(p_A) \sim Beta(1, 1)$$
$$\mu \sim U(-10, 0)$$

$$\sigma \sim U(0, 5)$$

This lead the posterior estimates of interest heavily depend on the likelihood. Later the priors will be changed and the impact of the change over posteriors will be investigated.

Early Result

2 METHODS

We first focus on the simplest sitaution that can describe the data sets and the general idea.

1. First item in a list
2. Second item in a list
3. Third item in a list

2.1 Paragraphs

PARAGRAPH DESCRIPTION

DIFFERENT PARAGRAPH DESCRIPTION

2.2 Math

$$\cos^3\theta = \frac{1}{4}\cos\theta + \frac{3}{4}\cos 3\theta \tag{1}$$

Definition 1 (Gauss)**.** To a mathematician it is obvious that $\int_{-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi}$.

Theorem 1 (Pythagoras)**.** *The square of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the other two sides.*

Proof. We have that $\log(1)^2 = 2\log(1)$. But we also have that $\log(-1)^2 = \log(1) = 0$. Then $2\log(-1) = 0$, from which the proof. □

3 RESULTS AND DISCUSSION

Reference to Figure 1 on the following page.

3.1 Subsection

3.1.1 *Subsubsection*

WORD Definition

CONCEPT Explanation

IDEA Text

- First item in a list
- Second item in a list
- Third item in a list

3.1.2 *Table*

Reference to Table 1 on the next page.

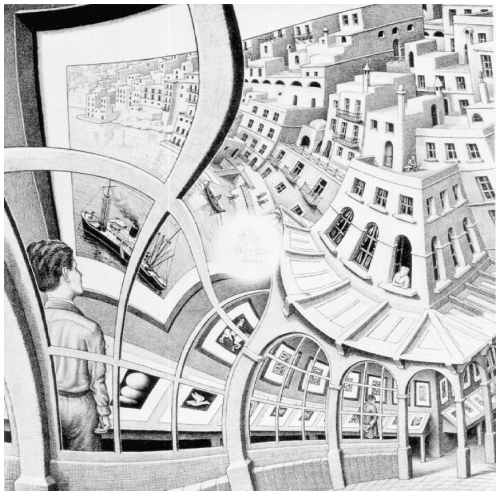


Figure 1: An example of a floating figure (a reproduction from the *Gallery of prints*, M. Escher, from http://www.mcescher.com/).

Table 1: Table of Grades

Name		
First name	Last Name	Grade
John	Doe	7.5
Richard	Miles	2

3.2 Figure Composed of Subfigures

Reference the figure composed of multiple subfigures as Figure 2 on the following page. Reference one of the subfigures as Figure 2b on the next page.

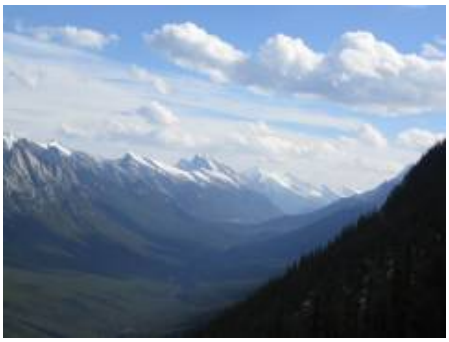
REFERENCES

[1] Buxton J Balshaw R Otterstatter M Macdougall L et al. Irvine MA, Kuo M. Modelling the combined impact of interventions in averting deaths during a synthetic-opioid overdose epidemic. *Addiction*, 2019.

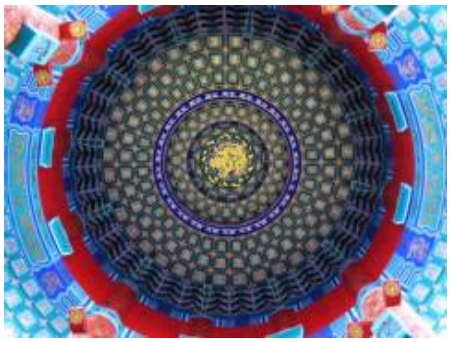
[2] A. J. Figueredo and P. S. A. Wolf. Assortative pairing and life history strategy - a cross-cultural study. *Human Nature*, 20:317–330, 2009.

(a) A city market.

(b) Forest landscape.



(c) Mountain landscape.



(d) A tile decoration.

Figure 2: A number of pictures with no common theme.