21th of June 2019

INAR-hidden Markov models to detect and quantify misreported diagnosis in ADHD

Amanda Fernández-Fontelo[1], Alejandra Cabaña[2], David Moriña[2], Anna Giménez Palomo[3], Pedro Puig[2]

[1]Humboldt-Universität zu Berlin, [2]Universitàt Autònoma de Barcelona, [3]Hospital Clínic de Barcelona

Plan

1. Introduction
   - Misreporting
   - Motivation: ADHD
   - Data
   - Goals
2. The model
   - Model versions
   - Parameter estimation
   - Viterbi algorithm: most likely latent (true) sequence
   - Model performance
3. Preliminary results

Introduction

Misreporting

Misreporting in data refer to some incident responsible for reporting less (under-reporting) or more (over-reporting) than the actual level of data.

- Many consequences can derive from misreporting: e.g., inferences emerged from misreported data might be severely biased, drawing an unrealistic picture of the actual problem.
- Focusing on the public health context, it is well known that some diseases are traditionally under-reported as well as issues such as gender-based violence. However, the over-reporting in some disorders (e.g., ADHD) is extensively documented.

Attention Deficit Hyperactivity Disorder

- ADHD is a mental disorder usually diagnosed for the first time in school-aged children. Although this is one of the most common mental disorders in children, it remains poorly understood.
- Some of the more frequently documented reasons related to this systematic misdiagnosis in ADHD are:
  - relative age of school-age children.
  - symptoms are drastically different between boys and girls.
  - disagreement between diagnosis protocols.
  - etc.

Ford-Jones, P.C. (2015). Misdiagnosis of attention deficit hyperactivity disorder: "Normal behaviour" and relative maturity, Paediatric Child Health, 20(4): 200–202.

Data

Data are the monthly number of visits to CSMA from 2013 to 2018, and come from "Agència de Qualitat i Avaluació Sanitàries de Catalunya (AQuAS)"[1]:

- Geographic area (comarca and municipi of Catalonia).
  - → Barcelonès, Vallès Occidental and Baix Llobregat.
- Age range (<1 year, then 5-year intervals from 1 year to 79).
  - → Re-categorized in children (<14), adolescents ($15-24$) and adults (>24).
- Gender.
- Relevance of ADHD in the visit: is this disease the first reason for visiting?
  - → Considered cases where ADHD is the primary reason for visiting.

[1] http://aquas.gencat.cat/ca/inici

Goals

The goal of this work consists of providing a novel tool able to detect and quantify the misreporting in both directions (under- and over-reporting):

- Apply the tool to detect misreporting in ADHD cases in Catalonia according to the geographical area, gender, and age to provide clinicians a more objective measure to such misdiagnosis problem.

The model

Model I: considering only under-reporting

Consider a **latent process** X_n with the following Poisson(λ)-INAR(1) structure:

$$X_n = \alpha \circ X_{n-1} + W_n(\lambda),$$

where $\alpha \in (0, 1)$. $\mathrm{E}(X_n) = \mathrm{V}(X_n) = \lambda/(1-\alpha) = \mu_X$. The operator $\circ$ is the binomial thinning such that: $\alpha \circ X_{n-1} = \sum_{i=1}^{X_{n-1}} Z_i$,, where Z_i are i.i.d Bernoulli(α).

Consider the following **observed and potentially under-reported process** Y_n: X_n with probability $1-\omega$, or $q \circ X_n$ with probability ω.

Fernández-Fontelo, A., Cabaña, A., Puig, P. and Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. Statistics in Medicine, 35(26): 4875-4890.

Model II: considering under-reporting under a more sophisticated correlation structure

Imagine now that the under-reporting indicators $\{\mathbf{1_n}, n \geq 1\}$ are also serially dependent.

- Take the simplest scheme: a binary discrete-time Markov chain.

Now, the **observed and potentially under-reported process** $\{Z_n\}$ is: X_n with probability $1 - \omega$, or $q \circ X_n$ with probability ω, where $\mathrm{P}(\mathbf{1}_{n+k} = 1_{n+k} | \mathbf{1}_n = 1_n) \neq \mathrm{P}(\mathbf{1}_{n+k} = 1_{n+k})$.

- One additional parameter is included in the model to accommodate the correlation among the under-reporting states.

Fernández-Fontelo, A., Cabaña, A., Joe, H., Puig, P. and Moriña, D. (2019). Untangling serially dependent under-reported count data for gender-based violence. Accepted in Statistics in Medicine.

Model III: Considering both under-reporting and over-reporting

Still consider that the latent process X_n is a Poisson(λ)-INAR(1) model, but now the processes Y_n (and Z_n) can be misreported (under- or over-reported):

$$Y_n = \begin{cases} X_n & 1-\omega \\ \theta \Diamond X_n & \omega, \end{cases}$$

where $\theta = (\phi_1, \phi_2)$ and $\theta \Diamond X_n$ is called the fattening-thinning operator: $\theta \Diamond X_n | X_n = x_n = \sum_{j=1}^{x_n} W_j$:

$$\mathrm{P}(W = j) = \begin{cases} 0 & 1 - \phi_1 - \phi_2 \\ 1 & \phi_1 \\ 2 & \phi_2 \end{cases}.$$

Under-reporting or over-reporting ?

- To distinguish between under-reporting, no misreporting or over-reporting, the following can easily be computed once the model is estimated:

under-reporting	$\phi_1 + 2\phi_2 < 1$
no misreporting	$\phi_1 + 2\phi_2 = 1$
over-reporting	$\phi_1 + 2\phi_2 > 1$

- Notice that when $\phi_2 = 0$, the model results in the versions I and II, which only accounts for under-reporting or no misreporting.

Parameter estimation: moment-based method

The marginal distribution of the observed process (Y_n or Z_n) is essential to compute the moment-based estimates of the model:

$$Y_n = \begin{cases} \text{Poisson}(\mu_X) & 1-\omega, \\ \text{2nd-order Hermite}\,(\mu_X\phi_2, \mu_X(1-\phi_1-\phi_2)) & \omega. \end{cases}$$

1. Fit the mixture above to obtain estimates of $\widehat{\omega}$, $\widehat{\mu_X}$, $\widehat{\phi_1}$ and $\widehat{\phi_2}$.
2. Use the theoretical expression of the ACF (ρ_Y and ρ_Z) to estimate α.
3. Using $\widehat{\mu_X}$ (step 1) and $\widehat{\alpha}$ (step 2), λ can be easily estimated.

If G_1 and G_2 are independent Poisson distributions with parameters a and b, $H = G_1 + 2G_2 \sim \text{2nd-order Hermite}(a, b)$.

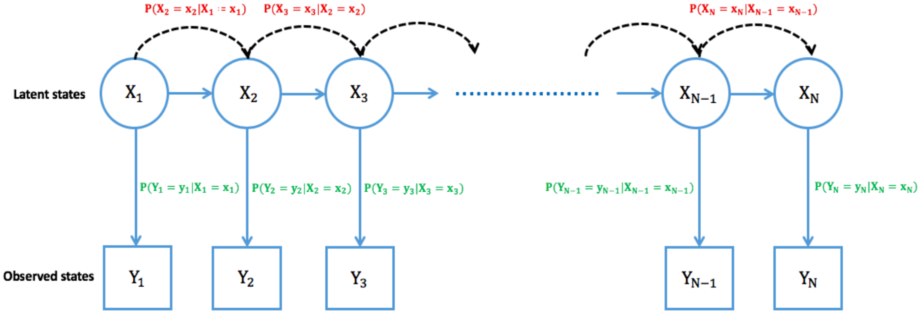
Parameter estimation: likelihood-based method

Since the likelihood functions of the processes Y_n and Z_n are directly intractable (HMC with an infinite number of states), the forward algorithm is a reasonable choice to compute such functions.

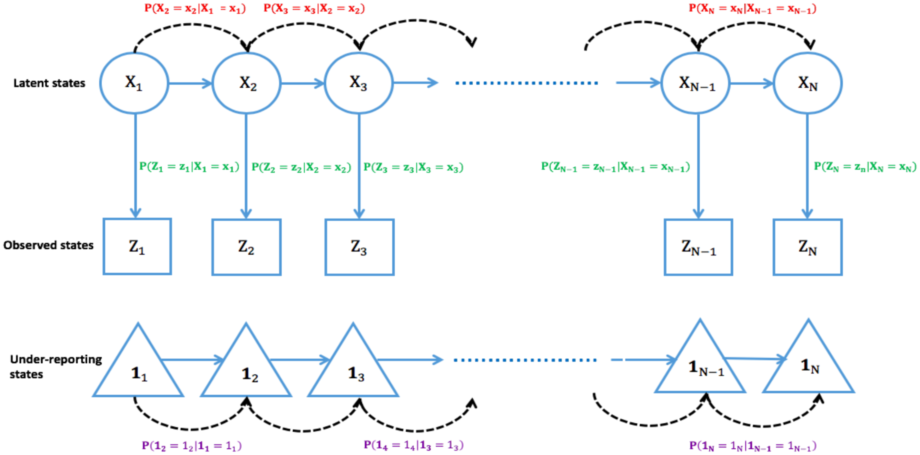
The likelihood is then achieved by $\mathrm{P}(Y_{1:N} = y_{1:N}) = \sum_{x_N=\frac{y_N}{2}, 1_N}^{\infty} \gamma_N(y_{1:N}, x_N, 1_n)$, where:

$$\gamma_n(y_{1:n}, x_n, 1_n) = \overbrace{\mathrm{P}(Y_n = y_n | X_n = x_n, 1_n)}^{\text{emission probabilities}} \sum_{x_{n-1}=\frac{y_{n-1}}{2}, 1_{n-1}}^{\infty} \underbrace{\mathrm{P}(X_n = x_n | X_{n-1} = x_{n-1})}_{\text{transition probabilities}}$$
$$\times \underbrace{\mathrm{P}(1_n = 1_n | 1_{n-1} = 1_{n-1})}_{\text{transition probability matrix}} \gamma_{n-1}(y_{1:n-1}, x_{n-1}, 1_{n-1}).$$

Visual model



Visual model



Forward probabilities

- While the transition probabilities are easily computed through the Poisson(λ)-INAR(1) cpmf, the emission probabilities are trickier:

model	$\mathrm{P}(Y_n = y_n \mid X_n = x_n, \mathbf{1}_n = 1_n)$
Y_n	$\begin{cases} 0 & \text{if} \quad x_n < \frac{y_n}{2} \\ (1-\omega) + \omega p_n & \text{if} \quad y_n = x_n \\ \omega p_n & \text{if} \quad x_n \geq \frac{y_n}{2} \end{cases}$
Z_n	$\begin{cases} 0 & \text{if} \quad x_n < \frac{y_n}{2} \\ 0 & \text{if} \quad 1_n = 0, x_n < \frac{y_n}{2} \\ p_n & \text{if} \quad 1_n = 1, x_n \geq \frac{y_n}{2} \\ 1 & \text{if} \quad 1_n = 0, x_n = y_n \end{cases}$

- p_n can be computed through the following recursive relation:

$$p_n = \frac{1}{n(1-\phi_1-\phi_2)} \left[\phi_1(x_n-(n-1))p_{n-1} + \phi_2(x_n-(n-2))p_{n-2}\right]$$

Viterbi algorithm: most likely latent (true) sequence

The Viterbi algorithm [2] is used to know the latent chain that maximises $\mathrm{P}(X_{1:n}|Y_{1:n}) = \frac{\mathrm{P}(X_{1:n}, Y_{1:n})}{\mathrm{P}(Y_{1:n})}$ (assuming all parameters are known):

- Since $\mathrm{P}(Y_{1:n})$ does not depend on X_n, it is enough to maximise the probability $\mathrm{P}(X_{1:n}, Y_{1:n})$.
- The most likely chain of latent states is obtained as:

$$X^* = \arg\max_X \mathrm{P}(X_{1:n}, Y_{1:n}).$$

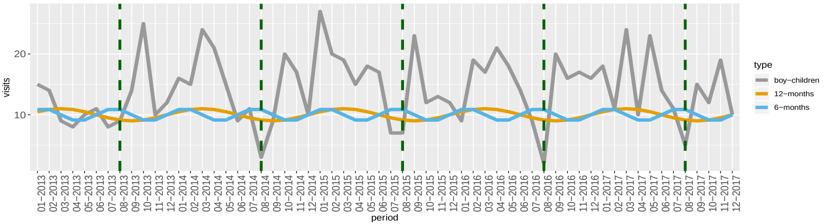
[2] Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* Apr 1967, 13, 260-269.

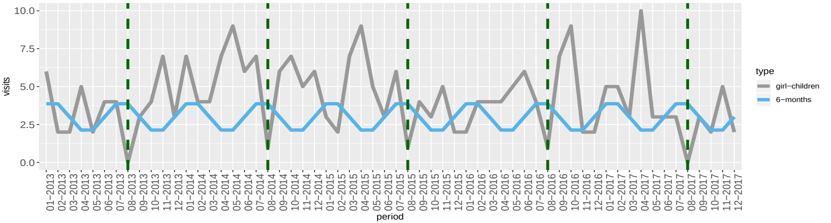
Preliminary results

Exploratory analysis

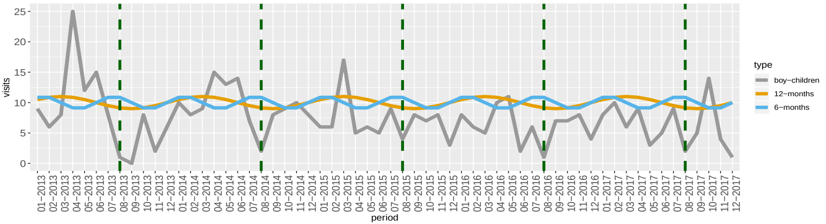
- Nearly 60 series among Barcelonès, Vallès Occidental and Baix Llobregat. Some of them are rejected because of low counts and high frequency of zeros (convergence problems in MLE).
- Many of them (especially boy-children) show clear annual and semiannual seasonal patterns (e.g., vacation periods).
- Example of boy-children and girl-children in Hospitalet de Llobregat (Barcelonès) and Terrassa (Vallès Occidental).

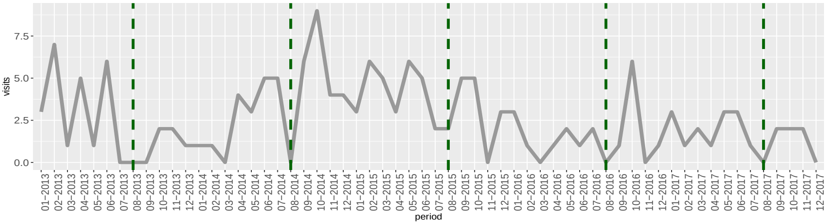
Seasonal patterns in Hospitalet de Llobregat





Seasonal patterns in Terrassa





Proposed models

area	gender	model	misreporting
Hospitalet de Llob.	boy	$X_n \sim \mathrm{Poisson}\left(e^{2.3224+0.1742 \sin \frac{2\pi n}{12} + 0.1314 \cos \frac{2\pi n}{12}}\right)$	
		$Y_n = \begin{cases} X_n & 0.1325 \\ \widehat{\theta} = (0.3122, 0.5632)\Diamond X_n & 0.8675 \end{cases}$	$0.3122 + 2\dot{0}.5632 > 1$
	girl	$Y_n \sim \mathrm{Poisson}\left(e^{1.4272-0.2812 \sin \frac{2\pi n}{6} + 0.0118 \cos \frac{2\pi n}{6}}\right)$	-
Terrassa	boy	$X_n \sim \mathrm{Poisson}\left(e^{2.8208+0.2687 \sin \frac{2\pi n}{12} + 0.2040 \cos \frac{2\pi n}{12}}\right)$	
		$Y_n = \begin{cases} X_n & 0.0275 \\ \widehat{\theta} = (0.0629, 0.1776)\Diamond X_n & 0.9725 \end{cases}$	$0.0629 + 2\dot{0}.1776 < 1$
	girl	$Y_n \sim \mathrm{Poisson}(4.2500)$	-

To Do

- Series should be conveniently validated through pseudo-residuals, and the most likely sequence of latent state should be provided (with bands).
- The reminder series of Barcelonès, Vallès Occidental, and Baix Llobregat should be analyzed appropriately.
- Seasonal patterns have to be validated with clinicians. Furthermore, further discussions with psychiatrists are needed to confirm the differences in misreporting depending on geographical areas.

Thank
You
Merci
Gracias
TerimaKasih
Hvala
Arigatô
Spacibo
Danke
KamSahHamnida
Mèsi
Xièxie
Grazie
Chokrane
Dziakuju
Kiitos
Takk
aDank
BarakAllahuFiik
Rahmat
ThintKo
Buznyg
Blagodaram
Matondo
Mercé
Taiku
Dakujem
Gràcies
Faleminderit
Efharisto
NajisTuke
Grandmerce
Tanemirt
Tenki
Aabhar
TapadhLeat
Köszönöm
Murakoze
WaadMahadsantahay
Grazzi
Mauruuru
GratiasAgimus
Barkal
Ačiū
TänanVäga
Dankie
Tak
Motashakkeram
Sagolun
GratiasAgo
KopKhunKha
Gracie
CámOn
Multumesc
Saha
Salamat
Obrigado
Bayarlalaa
Nirringrazzjak
Chnorakaloutioun
KhobChaiDeu
Gratzias
Dhanyavaad
Mamnun