THE DYNAMIC MODEL TO INFECTION RATE BASED ON POOLED SAMPLES

Paola da Silva Martins

Thesis of Master presented to the Postgraduate in Mathematics Program, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

Advisor: James Dean Oliveira dos Santos Júnior

Manaus
May 2020

THE DYNAMIC MODEL TO INFECTION RATE BASED ON POOLED SAMPLES

Paola da Silva Martins

THESIS SUBMITTED TO THE FACULTY OF THE POSTGRADUATE IN MATHEMATICS PROGRAM, OF THE FEDERAL UNIVERSITY OF AMAZONAS, AS A PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF MATHEMATICS.

Examined by:

Prof. Laura Letícia Ramos Rifo, Dr.

Prof. Kelly Cristina Mota Gonçalves, Dr.

Prof. James Dean Oliveira dos Santos Júnior, Dr.

MANAUS, AM – BRAZIL
MAY 2020

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

M386d

Martins, Paola Da Silva
The dynamic model to infection rate based on pooled samples / Paola Da Silva Martins . 2020
46 f.: il. color; 31 cm.

Orientador: James Dean Oliveira dos Santos Júnior
Dissertação (Mestrado em Matemática - Estatística) - Universidade Federal do Amazonas.

1. Dynamic Generalized Linear Model. 2. West Nile. 3. Beta Distribution. 4. Disease. 5. Empirical Bayes. I. Santos Júnior, James Dean Oliveira dos. II. Universidade Federal do Amazonas III. Título

"Live as if you were to die
tomorrow. Learn as if you were
to live forever."
Mahatma Gandhi

Acknowledgement

Primarily I would thank God for giving me all the help, strength and determination to complete my thesis.

I would like to express my gratitude to my advisor James Dean for the useful comments, remarks and engagement through the learning process of this thesis, as well for the support on the way.

I take this opportunity to express gratitude to all of the Department faculty members for their help and support. I am also extremely thankful for my friends that were all the time with me either in the classroom studying, or eating açaí, or having meals at the cafeteria.

I would like to thank my loved ones, my parents and brothers, who have supported me throughout the entire process, both by keeping me harmonious and helping me put pieces together. I will be grateful forever for your love. I also thank my boyfriend, Quentin, for the unceasing encouragement, support and attention.

Abstract of Thesis presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

THE DYNAMIC MODEL TO INFECTION RATE BASED ON POOLED SAMPLES

Paola da Silva Martins

May/2020

Advisor: James Dean Oliveira dos Santos Júnior

Research lines: Statistics

In this thesis we will work on the real time estimation of infection rates in vectors. It uses the dynamic generalized linear model to estimate the rate of infection of theses vector that are put in different pools sizes. The proposed methodology used the data of the mosquitoes tested weekly during the months of June through October referring to the period of 2012 to 2019. These mosquitoes were taken from the Department of Health from Rhode Island, in the United States. The model found had a good adherence to the aforementioned data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diseases such as dengue, leishmaniasis, onchocerciasis and Nile fever are transmitted to humans by arbovirus (name used to refer to any viruses that are transmitted by arthropod vectors). In order to control diseases such as these, governments and private enterprises have set up epidemiological control agencies that collect information regularly to monitor their development and issue warnings for intervention. This is accomplished by monitoring various indicators.

Among the monitored variables, we highlight the prevalence of infection in the vectors. As this is typically low (less than 1%), it would take a large sample to find an infected vector, which makes verifying individual dissecting vectors very costly. The solution is to create a pool: a grouping of vectors that will be shredded and tested simultaneously. Pooled samples are observed over epidemiological weeks, obtaining a time series of the test results for the arbovirus in question. These results are used to create alerts: an increase in the infection rate in the vectors may imply an increase in the incidence of the disease in humans.

For this thesis, it will be used a data that the arbovirus is the West Nile virus (WNV), that causes West Nile fever. This disease is typically spread by mosquitoes. Cases of WNV occur during mosquito season, which starts in the summer and continues through fall, in the United States. There are no vaccines to prevent or medications to treat WNV in people. On the other hand, the repellent use has been associated with reduced risk of WNV infection [Mon]. Fortunately, most people infected with WNV do not feel the symptoms.

Since the population of the data is composed by a population of mosquitoes, which is difficult to be accessed, the solution that was adopted to get information was to create pools, which in each pool there is an amount of mosquitoes. The mosquitoes collected for each pool were crushed in order to form a soup. From this soup the study was made to know if the pool was infected or not. For this reason, it is impossible to know how many mosquitoes were infected. Knowing only the information of the infected pool, that is, there is at least one infected mosquito in the pool. The idea of using a pool was a good strategy, since checking mosquito a mosquito would be very laborious and costly. And it was done over a period of time. In Figure 1.1 we can see the number of infected pools, in the State of Rhode Island, between 2012 and 2019. It's interesting to observe that there were a small amount of infected pools.

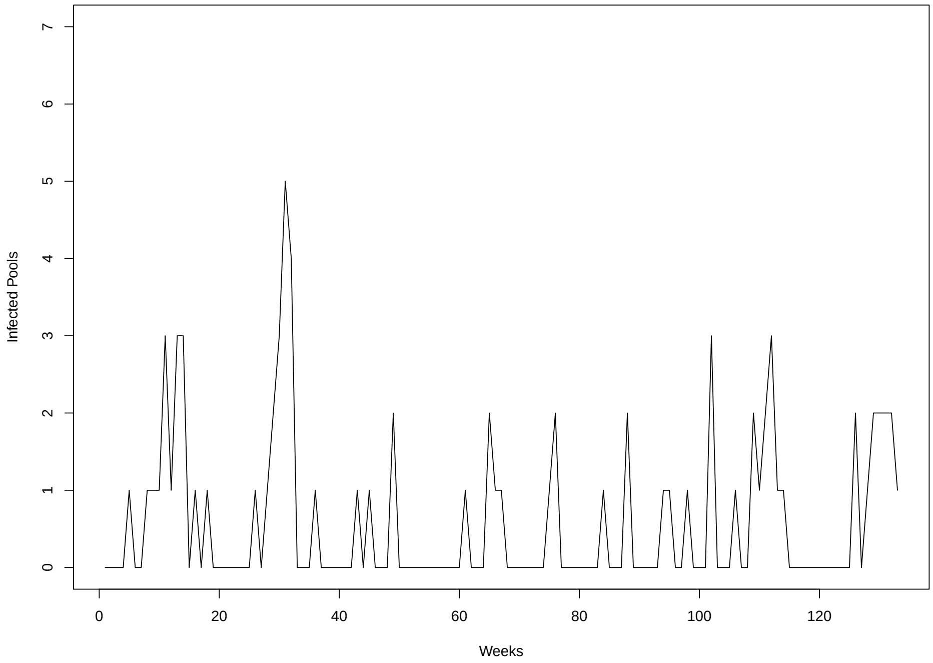

Figure 1.1: Chart of the infected pools between 2012 and 2019

Therefore, for each period of time, we had observed the rate infected of mosquitoes in the pools. Consequently, the amount of mosquitoes varies during the time. So, to analyse a data with dependence in time, we can use basic time series analysis. One possibility is to find a reasonable regularity in the behavior of the phenomenon under study: forecasting the future behavior is clearly easier if the series tends to repeat a regular path over time. Another way is using dynamic generalized linear models, estimation and forecasting can be obtained recursively by the well known Kalman filter. In this work we will explain and propose the last one.

1.1 Objective

The objective of this thesis is to build a dynamic generalized linear model to estimate the rate of infection for any arbovirus. This model will be applied to data from Rhode Island Department of Environmental Management (DEM).

1.2 Organization of the text

The Chapter 2 shows how the likelihood of the original problem can be approximated by a model in the exponential family. It was done a literature review of the methods which were used in the next chapters, highlighting the characteristics of the state space models, dynamic linear model, weak bayes, and dynamic generalized linear model that will be proposed as a model. The Chapter 3 describes the structure of the model for datas in pools. In addition, it discusses a conjugate prior distribution to approximate the likelihood function of infection rate. Chapter 4 briefly describes how the data were obtained and presents the application of the proposed modeling. First, an exploratory analysis of the data is presented. And then, it shows the predicted application of rate estimation for the West Nile Fever transmitter mosquitoes in the United States. Chapter 5 presents the conclusion of the work.

Chapter 2

Literature Review

In this chapter we show an approximate likelihood that will be necessary for the development of this thesis. Besides that, we discuss the basic notions about state space models. Dynamic linear models are presented as a special case of general state space models, being linear and Gaussian.

2.1 Approximate Likelihood

Considering a sample of k vectors, let $X_i \sim Bernoulli(\pi)$, where $X_i = 1$ implies that the i^{th} vector was infected and π is the prevalence of having an infected pool. Let

$$Y = \max\{X_1, \dots, X_k\} \tag{2.1}$$

be the result of the pool test formed by the k vectors and such that $X_i \sim Bernoulli(\pi)$ independents, with $i = 1, \dots, k$. A positive result ($Y = 1$) indicates that at least one vector in the sample was infected, while a negative result ($Y = 0$) implies that all the vectors were not infected. Considering that $X_1, \dots, X_k$ is a sample of independent and identically distributed random variables, we have that

$$Y \sim Bernoulli(1 - (1 - \pi)^k). \tag{2.2}$$

In the matter of monitoring and intervention, these pools are analyzed within a time window (usually one epidemiological week. Epidemiological week is on week starting on Sunday and ending on Saturday). Then consider that for a given week t, n_t pools were observed, with sizes $k_t = \{k_{1,t}, \dots, k_{n_t,t}\}$, where $k_{i,t}$ represents the i^{th} pool size in the epidemiological week t, resulting in the results $y_t = \{y_{1,t}, \dots, y_{n_t,t}\}$ where $y_{i,t}$ is the information about the infection or not of the i^{th} pool in the epidemiological week t, $i = 1 \dots, n_t$. So the likelihood function for π_t, which represents the prevalence in week t, can be written as

$$L(\pi_t) = \prod_{i=1}^{n_t} p(y_{i,t}|\pi_t) = (1-\pi_t)^{\sum_{i=1}^{n_t} k_{i,t}(1-y_{i,t})} \prod_{i=1}^{n_t} [1-(1-\pi_t)^{k_{i,t}}]^{y_{i,t}}. \tag{2.3}$$

It is important to stretch that π_t is variable throughout the year, since the samples have different sizes. If all pools have the same sample size, the prediction problem can be solved using dynamic generalized linear models (dglm) [16], since the total number of infected pools can be appropriately modeled by one model in the exponential family. Unfortunately, this does not apply when pools have different sizes. However, Santos and Dorgam [15] showed that when the pools have small variability between the sizes and prevalence is low, the likelihood in 2.3 can be approximated by

$$L^*(\pi_t) = (1-\pi_t)^{n_t \bar{k}_t (1-\bar{y}_t)} \left[1-(1-\pi_t)^{\bar{k}_t}\right]^{n_t \bar{y}_t}, \tag{2.4}$$

where

$$\bar{y}_t = \sum_{i=1}^{n_t} \frac{y_{i,t}}{n_t}, \tag{2.5}$$

$$\bar{k}_t = \sum_{i=1}^{n_t} \frac{k_{i,t}}{n_t}, \tag{2.6}$$

where $\bar{y}_t$ is the mean of infected pools at time t and $\bar{k}_t$ is the mean size of the pools at time t.

In this thesis, both hypotheses are satisfied, that means, the pools have small variability between the sizes and prevalence is low. Indeed, they have shown that under these conditions, common in practice, the Kullback-Leibler divergence [7] between $L(\pi)$ and $L^*(\pi)$ is given by

$$KL(L, L^*) = E_L\left(\frac{L(\pi)}{L^*(\pi)}\right) \approx \pi \sum_{i=1}^{n} k_i log\left(\frac{k_i}{\bar{k}_t}\right). \tag{2.7}$$

For a more detailed explanation of the Kullback-Leibler divergence see [7].

If k_i are close to $\bar{k}$ then the log will be close to zero. This is the condition of low variability.

From the approximation of the proposed likelihood, most classics estimators are based on the statistics $\dot{y}_t$ ($\dot{y}_t = \sum_{i=1}^{n_t} y_{i,t}$). Since $L(\pi_t)$ 2.3 can be approximated by

$L^*(\pi_t)$ 2.4, it follows that

$$Y_{i,t} \approx Bernoulli(1-(1-\pi_t)^{\bar{k}_t}). \tag{2.8}$$

Therefore, it is possible to build new estimators for π_t using existing results for a single pool size.

By maximizing L^* with respect to π_t, the approximate maximum likelihood estimator is given by

$$\hat{\pi_t}^* = 1 - \left(1 - \frac{\sum_{i=1}^{n_t} Y_i}{n_t}\right)^{1/\bar{k}_t}, \tag{2.9}$$

and its corrected version, obtained after the bias correction of Burrows [11], is given by

$$\hat{\pi_t}^*_B = 1 - \left(\frac{2\bar{k}_t(n_t - \sum_{i=1}^{n_t} Y_i) + \bar{k}_t - 1}{2\bar{k}_t n_t + \bar{k}_t - 1}\right)^{1/\bar{k}_t}. \tag{2.10}$$

In practice the pools have different sizes, but to facilitate the modeling of the problem it is assumed that the pools have equal sizes, in this case the average size of the pools in a given epidemiological week t.

Mathematically, the problem can be addressed using different pool sizes, however, the maximal estimator is biased, and to correct this, Burrows did two approaches. Another way to deal with the mathematical problem is to assume that the pools have equal sizes. This last form was chosen to deal with in the current thesis.

With this model, the purpose of this thesis is to present the construction of a dynamic generalized linear model for π_t.

2.2 State Space Models

Consider a time series $\{Y_t, t = 1, 2, \dots\}$, where Y_t is an observable $(m \times 1)$ random vector. For making inference on the time series, in particular for predicting the next value Y_{t+1} given the observations $\{Y_1, \dots, Y_t\}$, we need to specify the probability law of the process (Y_t), which means giving the dependence structure among the Y_t's variables. The Figure 2.1 represents graphically the dependency among the observed variables Y_t and the unobservable process [10].

$$\begin{tikzcd}
\theta_0 \arrow[r] & \theta_1 \arrow[r] \arrow[d] & \theta_2 \arrow[r] \arrow[d] & \cdots \arrow[r] & \theta_{t-1} \arrow[r] \arrow[d] & \theta_t \arrow[r] \arrow[d] & \theta_{t+1} \arrow[r] \arrow[d] & \cdots \\
 & Y_1 & Y_2 & & Y_{t-1} & Y_t & Y_{t+1} &
\end{tikzcd}$$

Figure 2.1: Dependence structure for a state space model

State space models are based on the idea that the time series (Y_t) is an incomplete and noisy function of some underlying unobservable process $\{\theta_t, t = 1, 2, \dots\}$, called the state process. The term state space model is used when the state variables are continuous. When they are discrete, the term hidden Markov model is used.

2.2.1 Filtering

The use of recent data to revise inferences regarding previous values of the state vector is called filtering, this information being filtered back to previous time points.

Let us denote with D_t the information provided by the first t observations, $\{Y_1, \dots, Y_t\}$. The filtered and predictive densities can be computed by a recursive algorithm. Starting from $\theta_0 \sim p_0(\theta_0) = p(\theta_0|D_0)$ one can recursively compute, for $t = 1, 2, \dots$ the following.

Proposition 2.2.1. *(Filtering recursion).*

i) The one-step-ahead predictive density for the states can be computed from the filtered density $p(\theta_{t-1}|D_{t-1})$ according to

$$p(\theta_t|D_{t-1}) = \int p(\theta_t|D_{t-1})p(\theta_{t-1}|D_{t-1})dv(\theta_{t-1}). \tag{2.11}$$

ii) The one-step-ahead predictive density for the observations can be computed from the predictive density for the states as

$$f(y_t|D_{t-1}) = \int f(y_t|\theta_t)p(\theta_t|D_{t-1})dv(\theta_t). \tag{2.12}$$

iii) The filtering density can be computed from the above densities as

$$p(\theta_t|D_t) = \frac{f(y_t|\theta_t)p(\theta_t|D_{t-1})}{p(y_t|D_{t-1})}. \tag{2.13}$$

Proof. i) Note that θ_{t+1} is independent of $(Y_1, \ldots, Y_t)|\theta_t$. Therefore

$$\begin{aligned} p(\theta_t|D_{t-1}) &= \int p(\theta_{t-1}, \theta_t|D_{t-1}) dv(\theta_{t-1}) \\ &= \int p(\theta_t|\theta_{t-1}, D_{t-1}) p(\theta_{t-1}|D_{t-1}) dv(\theta_{t-1}) \\ &= \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|D_{t-1}) dv(\theta_{t-1}). \end{aligned} \tag{2.14}$$

ii) From the conditional independence between Y_t and $(Y_1, \ldots, Y_{t-1})|\theta_t$, we have:

$$\begin{aligned} f(y_t|D_{t-1}) &= \int f(y_t, \theta_t|D_{t-1}) dv(\theta_t) \\ &= \int f(y_t|\theta_t, D_{t-1}) p(\theta_t|D_{t-1}) dv(\theta_t) \\ &= \int f(y_t|\theta_t) p(\theta_t|D_{t-1}) dv(\theta_t). \end{aligned} \tag{2.15}$$

iii) Using the Bayes rule:

$$p(\theta_t|D_t) = \frac{p(\theta_t|D_{t-1}) f(y_t|\theta_t, D_{t-1})}{f(y_t|D_{t-1})} = \frac{p(\theta_t|D_{t-1}) f(y_t|\theta_t)}{p(y_t|D_{t-1})}, \tag{2.16}$$

by the conditional independence between Y_t and $(Y_1, \ldots, Y_{t-1})|\theta_t$.

□

2.2.2 Smoothing

One of the attractive features of state space models is that estimation and forecasting can be developed sequentially. However, in time series analysis one often has observations on Y_t for a certain period, $t = 1, \ldots, T$ and wants to retrospectively reconstruct the behavior of the system, underlying the observations. While Bayesian filters in their basic form only compute estimates of the current state of the system given the history of measurements, smoothing can be used to reconstruct states that happened before the current time. Again, one has a backward-recursive algorithm for computing the conditional densities of $\theta_t|D_T$, for $t < T$, starting from the filtering density $p(\theta_T|D_T)$ and estimating backward all the states' history.

Proposition 2.2.2. *(smoothing recursion):*

i) Conditional on D_T, the state sequence $(\theta_0, \dots, \theta_T)$ has backward transition probabilities given by

$$p(\theta_t|\theta_{t+1}, D_T) = \frac{p(\theta_{t+1}|\theta_t)p(\theta_t|D_t)}{p(\theta_{t+1}|D_t)}. \tag{2.17}$$

ii) The smoothing densities of θ_t given D_T can be computed according to the following backward recursion in t (starting from $p(\theta_T|D_T)$):

$$p(\theta_t|D_T) = p(\theta_t|D_t) = \int \frac{p(\theta_{t+1}|\theta_t)}{p(\theta_{t+1}|D_t)} p(\theta_{t+1}|D_T) d\nu(\theta_{t+1}). \tag{2.18}$$

Proof. i) Using the Bayes formula:

$$p(\theta_t|\theta_{t+1}, D_T) = p(\theta_t|\theta_{t+1}, D_t) = \frac{p(\theta_t|D_t)p(\theta_{t+1}|\theta_t, D_t)}{p(\theta_{t+1}|D_t)} = \frac{p(\theta_t|D_t)p(\theta_{t+1}|\theta_t)}{p(\theta_{t+1}|D_t)}. \tag{2.19}$$

ii) Let's find the marginal for $p(\theta_t, \theta_{t+1}|D_T)$:

$$\begin{aligned}
p(\theta_t|D_T) &= \int p(\theta_t, \theta_{t+1}|D_T) d\nu(\theta_{t+1}) = \int p(\theta_{t+1}|D_T)p(\theta_t|\theta_{t+1}, D_T) d\nu(\theta_{t+1}) \\
&= \int p(\theta_{t+1}|D_T)p(\theta_t|\theta_{t+1}, D_T) d\nu(\theta_{t+1}) \\
&= \int p(\theta_{t+1}|D_T) \frac{p(\theta_{t+1}|\theta_t, D_T)p(\theta_t|D_T)}{p(\theta_{t+1}|D_T)} d\nu(\theta_{t+1}) \\
&= p(\theta_t|D_t) \int p(\theta_{t+1}|\theta_t) \frac{p(\theta_{t+1}|D_T)}{p(\theta_{t+1}|D_t)} d\nu(\theta_{t+1}). \qquad (2.20)
\end{aligned}$$

□

2.3 Dynamic linear models

The first important class of state space models is given by Gaussian linear state space models, also called dynamic linear models (dlm), that can be specified by the following definition:

Definition 2.3.1. Let

$$Y_t = F_t\theta_t + v_t, \quad v_t \sim N_m(0, V_t), \tag{2.21}$$

$$\theta_t = G_t\theta_{t-1} + w_t, \quad w_t \sim N_p(0, W_t), \tag{2.22}$$

where G_t and F_t are known matrices and v_t and w_t are two independent white noises with normal distribution, with mean zero and the known covariance matrices V_t and W_t, respectively. Besides that, 2.21 are the *observation equations* and 2.22 are the *evolution equations.*

Furthermore, it is assumed that θ_0 has a Gaussian distribution,

$$\theta_0 \sim N_p(m_0, C_0). \tag{2.23}$$

2.3.1 The Kalman Filter for dlm

The Proposition 2.2.1 applied to the dlm is called Kalman filter. The Kalman filter has long been considered the best answer to several tracking and data prediction tasks. It merely calculates these the functions "measuring" and "updating" over and over again.

The filter cyclically overrides the mean and the variance of the result. The filter can continuously be assured on wherever it is, as long as the readings do not deviate too much from the predicted value.

Since the measured values (in update) match relatively well to the predicted ones (by predict), the filter improves step by step to make sure that it is correct (normal distributions become narrower and higher), even though the values are noisy.

The Kalman filter produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each timeframe.

Theorem 2.3.1. *(Kalman Filter). Given that $\theta_{t-1}|D_{t-1} \sim N_p(m_{t-1}, C_{t-1})$, for the dlm, if*

$$\theta_0|D_0 \sim N_p(m_0, C_0), \tag{2.24}$$

then, for every $t \geq 1$,

i) the one-step-ahead state predictive density of θ_t, given D_{t-1} is Gaussian, with parameters

$$a_t = E(\theta_t|D_{t-1}) = G_t m_{t-1}, \tag{2.25}$$

$$R_t = Var(\theta_t|D_{t-1}) = G_t C_{t-1} G_t^{'} + W_t; \tag{2.26}$$

ii) the one-step-ahead predictive density of Y_t given D_{t-1} is Gaussian, with pa-

rameters

$$f_t = E(Y_t|D_{t-1}) = F_t a_t, \tag{2.27}$$
$$Q_t = Var(Y_t|D_{t-1}) = F_t R_t F_t^{'} + V_t; \tag{2.28}$$

iii) the filtering density of θ_t given D_t is Gaussian, with

$$m_t = E(\theta_t|D_t) = a_t + R_t F_t^{'} Q_t^{-1} e_t, \tag{2.29}$$
$$C_t = Var(\theta_t|D_t) = R_t - R_t F_t^{'} Q_t^{-1} F_t R_t. \tag{2.30}$$

where $e_t = Y_t - f_t$ is the forecast error.

Proof. From standard results on the multivariate Normal distribution it follows that the joint density of $(\theta_0, \theta_1, \ldots, \theta_t, Y_1, \ldots, Y_t)$ is Gaussian, for any $t \geq 1$. Consequently, the distribution of any subvector is also Gaussian, as is the conditional distribution of some components given some other components. Therefore the predictive densities and the filtering densities are Gaussian, and it suffices to compute their means and variances.

i)

$$\begin{aligned} a_t &= E(\theta_t|D_{t-1}) = E(G_t\theta_{t-1} + w_t|D_{t-1}) \\ &= E(G_t\theta_{t-1}|D_{t-1}) + E(w_t|D_{t-1}) \\ &= G_t E(\theta_{t-1}|D_{t-1}) = G_t m_{t-1}. \end{aligned} \tag{2.31}$$

$$\begin{aligned} R_t &= Var(\theta_t|D_{t-1}) = E(Var(\theta_t|D_{t-1})) + Var(E(\theta_t|D_{t-1})) \\ &= E(Var(G_t\theta_{t-1} + w_t|D_{t-1})) = E(Var(G_t\theta_{t-1}|D_{t-1})) + E(Var(w_t|D_{t-1})) \\ &= E(G_t Var(\theta_{t-1}|D_{t-1})G_t^{'}) + E(W_t) \\ &= E(G_t C_{t-1} G_t^{'}) + W_t = G_t C_{t-1} G_t^{'} + W_t. \end{aligned} \tag{2.32}$$

ii)

$$\begin{aligned} f_t &= E(Y_t|D_{t-1}) = E(F_t\theta_t + v_t|D_{t-1}) \\ &= E(F_t\theta_t|D_{t-1}) + E(v_t|D_{t-1}) \\ &= F_tE(\theta_t|D_{t-1}) = F_ta_t. \end{aligned} \tag{2.33}$$

$$\begin{aligned} Q_t &= Var(Y_t|D_{t-1}) = E(Var(Y_t|D_{t-1})) + Var(E(Y_t|D_{t-1})) \\ &= E(Var(F_t\theta_t + v_t|D_{t-1})) = E(Var(F_t\theta_t|D_{t-1})) + E(Var(v_t|D_{t-1})) \\ &= E(F_tVar(\theta_t|D_{t-1})F_t^{'}) + E(V_t) \\ &= F_tR_tF_t^{'} + V_t. \end{aligned} \tag{2.34}$$

iii) Let's say $A_t = R_tF_t^{'}Q_t^{-1}$. We know that the covariance between $y_t|D_t$ and $\theta_t - A_ty_t|D_t$ is null, and from the normality, they are independents, then,

$$\begin{aligned} E(\theta_t|D_t) &= E(\theta_t - A_ty_t + A_ty_t|D_t) \\ &= E(\theta_t - A_ty_t|D_t) + E(A_ty_t|D_t) \\ &= E(\theta_t - A_ty_t|y_t, D_{t-1}) + A_ty_t \\ &= E(\theta_t - A_ty_t|D_{t-1}) + A_ty_t = a_t - A_tf_t + A_ty_t \\ &= a_t + A_t(y_t - f_t) = m_t. \end{aligned} \tag{2.35}$$

$$\begin{aligned} Var(\theta_t|D_t) &= Var(\theta_t - A_ty_t + A_ty_t|D_t) \\ &= Var(\theta_t - A_ty_t|D_t) = Var(\theta_t - A_ty_t|D_{t-1}) \\ &= Var(\theta_t|D_{t-1}) + A_tVar(y_t|D_{t-1})A_t^{'} - 2Cov(\theta_t, A_ty_t|D_{t-1}) \\ &= R_t + A_tQ_tA_t^{'} - 2Cov(\theta_t, y_t|D_{t-1})A_t^{'} \\ &= R_t + A_tQ_tA_t^{'} - 2R_tF_t^{'}A_t^{'} \\ &= R_t + A_tQ_tA_t^{'} - 2R_tF_t^{'}Q_t^{-1}Q_tA_t^{'} = R_t - A_tQ_tA_t^{'} = C_t. \end{aligned} \tag{2.36}$$

□

2.3.2 Smoothing

For the dlm the Proposition 2.2.2 is reduced to the following.

Proposition 2.3.2. *(smoothing recursion for the dlm):*

If $\theta_{t+1}|D_T \sim N_p(s_{t+1}, S_{t+1})$, *then* $\theta_t|D_T \sim N_p(s_t, S_t)$, *where*

$$s_t = m_t + C_t G^{'}_{t+1} R^{-1}_{t+1}(s_{t+1} - a_{t+1}). \tag{2.37}$$

$$S_t = C_t + C_t G^{'}_{t+1} R^{-1}_{t+1}(S_{t+1} - R_{t+1}) R^{-1}_{t+1} G_{t+1} C_t. \tag{2.38}$$

Proof.

$$\begin{aligned} p(\theta_t|\theta_{t+1}, D_T) &\propto p(\theta_t, \theta_{t+1}, y_T, \ldots, y_{t+1}, D_t) && (2.39)\\ &\propto p(y_T, \ldots, y_{t+1}|\theta_{t+1}) p(\theta_{t+1}|\theta_t, D_t) p(\theta_t|D_t) && (2.40)\\ &\propto p(\theta_{t+1}|\theta_t, D_t) p(\theta_t|D_t). && (2.41)\end{aligned}$$

If

$$\theta_{t+1}|\theta_t \sim N_p(G_{t+1}\theta_t, W_{t+1}), \tag{2.42}$$

$$\theta_t \sim N_p(m_t, C_t), \tag{2.43}$$

then, using the Theorem A.1 from the appendix we have:

$$\begin{pmatrix} \theta_{t+1} & \Big| D_t \\ \theta_t & \end{pmatrix} \sim N_p\left(\begin{pmatrix} G_{t+1}m_t \\ m_t \end{pmatrix}, \begin{pmatrix} G_{t+1}C_tG^{'}_{t+1} + W_{t+1} & G_{t+1}C^{'}_t \\ C_tG^{'}_{t+1} & C_t \end{pmatrix}\right). \tag{2.44}$$

Now we can write using the Conditional Theorem A.2 that

$$\theta_t|\theta_{t+1}, D_T \sim N_p(m_t + B_{t+1}(\theta_{t+1} - a_{t+1}), \quad C_t - B_{t+1}R_{t+1}B^{'}_{t+1}) \tag{2.45}$$

where $W_{t+1} + G_{t+1}C_tG^{'}_{t+1} = R_{t+1}$ and $C_tG^{'}_{t+1}R^{-1}_{t+1} = B_{t+1}$. Now we have through the Theorem A.1 from the appendix the Equation 2.45 and

$$\theta_{t+1}|D_T \sim N_p(s_{t+1}, \quad S_{t+1}). \tag{2.46}$$

Therefore,

$$\begin{pmatrix} \theta_t & \Big| D_T \\ \theta_{t+1} & \end{pmatrix} \sim N_p\left(\begin{pmatrix} m_t + B_1(s_{t+1} - a_{t+1}) \\ s_{t+1} \end{pmatrix}, \begin{pmatrix} U_{t+1} \end{pmatrix}\right), \tag{2.47}$$

where

$$U_{t+1} = \begin{pmatrix} C_t + B_1(S_{t+1} - R_{t+1})B^{'}_1 & B_1S_{t+1} \\ S_{t+1}B^{'}_1 & S_{t+1} \end{pmatrix}.$$

□

2.3.3 Forecasting

The forecasting involves the supply of forecast information in terms of probability distributions that represent and summarise current uncertain information and beliefs. For the sample D_t the forecast one-step-ahead distribution is $y_{t+k}|D_t$, with $k > 0$. After having the D_t, one may be inquisitive about forecasting future values of the observations, Y_{t+k}, or of the state vectors, θ_{t+k}. For dlm, the recursive kind of computations makes it natural to calculate the one-step-ahead forecasts and to update them consecutive, as new data become obtainable. This is clearly of interest in applied issues where the data do arrive consecutive, like in the problem of this thesis.

Note below that the data only enter the predictive distributions through the mean of the filtering distribution at the time the last observation was taken.

Proposition 2.3.3. *Set $a_t(0) = m_t$ and $R_t(0) = C_t$. Then, for $k \geq 1$, the following hold:*

i) The distribution of θ_{t+k} given D_t is Gaussian, with

$$a_t(k) = G_{t+k}a_t(k-1), \tag{2.48}$$

$$R_t(k) = G_{t+k}R_t(k-1)G^{'}_{t+k} + W_{t+k}; \tag{2.49}$$

ii) The distribution of Y_{t+k} given D_t is Gaussian, with

$$f_t(k) = F_{t+k}a_t(k), \tag{2.50}$$

$$Q_t(k) = F_{t+k}R_t(k)F^{'}_{t+k} + V_{t+k}. \tag{2.51}$$

Proof. i) We need to prove what are the parameters for the Gaussian distribution of θ_{t+k} given D_t.
If $k = 0$ we know that $\theta_t|D_t \sim N_p(m_t, C_t)$.
Using induction hypothesis we have $\theta_{t+k-1}|D_t \sim N_p(a_t(k-1), R_t(k-1))$.
We also know that

$$\theta_{t+k}|\theta_{t+k-1}, D_t \sim N_p(G_{t+k}\theta_{t+k-1}, W_{t+k}), \tag{2.52}$$

$$\theta_{t+k-1}|D_t \sim N_p(a_t(k-1), R_t(k-1)). \tag{2.53}$$

Through the Theorem A.1 of the Appendix we can write

$$\left(\begin{matrix}\theta_{t+k}\\ \theta_{t+k-1}\end{matrix}\ \middle|\ D_t\right) \sim N_p\left(\begin{pmatrix}G_{t+k}a_t(k-1)\\ a_t(k-1)\end{pmatrix}, \Big(Z_{t+k}(k-1)\Big)\right). \tag{2.54}$$

where

$$Z_{t+k}(k-1) = \begin{pmatrix} G_{t+k}R_t(k-1)G'_{t+k} + W_{t+k} & G_{t+k}R_t(k-1) \\ R_t(k-1)G'_{t+k} & R_t(k-1) \end{pmatrix} \tag{2.55}$$

Therefore,

$$\theta_{t+k}|D_t \sim N_p(a_t(k), R_t(k)). \tag{2.56}$$

ii) For the forecast we want to know the parameters for the Gaussian distribution $Y_{t+k}|D_t$.
We know that

$$Y_{t+k}|\theta_{t+k}, D_t \sim N_p(F_{t+k}\theta_{t+k}, V_{t+k}), \tag{2.57}$$

$$\theta_{t+k}|D_t \sim N_p(a_t(k), R_t(k)). \tag{2.58}$$

Now, it is easy to see through the Theorem A.1 from the appendix that

$$\begin{pmatrix} Y_{t+k} \\ \theta_{t+k} \end{pmatrix} \Bigg| D_t \Bigg) \sim N_p\left(\begin{pmatrix} F_{t+k}a_t(k) \\ a_t(k) \end{pmatrix}, \begin{pmatrix} F_{t+k}R_t(k)F'_{t+k} + V_{t+k} & F_{t+k}R_t(k) \\ R_t(k)F'_{t+k} & R_t(k) \end{pmatrix}\right). \tag{2.59}$$

Therefore,

$$Y_{t+k}|D_t \sim N_p(f_t(k), Q_t(k)). \tag{2.60}$$

□

2.4 Weak Bayes' Estimation

Among several model selection criteria, we choose the Weak Bayes' estimation, which makes probabilistic statements that facilitate recurrences for various distributional characteristics. These recurrence relationships are identical to those of linear Bayes' but are based upon a precise modelling assumption rather than a loss function approach.

In this thesis will be illustrated an application of Weak Bayes' estimation in the context of dlms. Let's suppose that $\theta_{t-1}|D_{t-1} \sim [m_{t-1}, C_{t-1}]$, which means that we don't know its distribution, except by its mean m_t and its variance C_t. As long as we consider the moments of the observation and evolution equations, let

$$
\begin{align}
E(\theta_t|D_{t-1}) &= a_t, \tag{2.61}\\
Var(\theta_t|D_{t-1}) &= R_t, \tag{2.62}\\
E(Y_t|D_{t-1}) &= f_t, \tag{2.63}\\
Var(Y_t|D_{t-1}) &= Q_t. \tag{2.64}
\end{align}
$$

The covariance between θ_t and Y_t given D_{t-1} is

$$
\begin{align}
Cov(\theta_t, Y_t|D_{t-1}) &= Cov(\theta_t, F_t\theta_t + v_t|D_{t-1}) \tag{2.65}\\
&= Cov(\theta_t, v_t|D_{t-1}) + Cov(\theta_t, F_t\theta_t|D_{t-1}) \tag{2.66}\\
&= 0 + Cov(\theta_t, \theta_t|D_{t-1})F_t^{'} = Var(\theta_t|D_{t-1})F_t^{'} = R_tF_t^{'}. \tag{2.67}
\end{align}
$$

Therefore, we can write:

$$\begin{pmatrix} \theta_t & \Bigg| D_{t-1} \\ Y_t & \end{pmatrix} \sim \left[\begin{pmatrix} a_t \\ f_t \end{pmatrix}, \begin{pmatrix} R_t & R_tF_t^{'} \\ F_tR_t & Q_t \end{pmatrix} \right]. \tag{2.68}$$

Let's say $A_t = R_tF_t^{'}Q_t^{-1}$. Then, with a transformation matrix

$$L = \begin{pmatrix} \mathbf{I} & -A_t \\ 0 & \mathbf{I} \end{pmatrix}, \tag{2.69}$$

the transformed vector

$$L\begin{pmatrix} \theta_t \\ Y_t \end{pmatrix} = \begin{pmatrix} \mathbf{I} & -A_t \\ 0 & \mathbf{I} \end{pmatrix}\begin{pmatrix} \theta_t \\ Y_t \end{pmatrix} = \begin{pmatrix} \theta_t - A_tY_t \\ Y_t \end{pmatrix}, \tag{2.70}$$

has the following moments

$$E\left(L\begin{pmatrix} \theta_t \\ Y_t \end{pmatrix} \middle| D_{t-1} \right) = \begin{pmatrix} E(\theta_t|D_{t-1}) - A_tE(Y_t|D_{t-1}) \\ E(Y_t|D_{t-1}) \end{pmatrix} = \begin{pmatrix} a_t - A_tf_t \\ f_t \end{pmatrix}, \tag{2.71}$$

$$Var\left(L\begin{pmatrix}\theta_t\\ Y_t\end{pmatrix}\middle| D_{t-1}\right) = LVar\left(\begin{matrix}\theta_t\\ Y_t\end{matrix}\ \middle| D_{t-1}\right)L' = \begin{pmatrix}\mathbf{I} & -A_t\\ 0 & \mathbf{I}\end{pmatrix}\begin{pmatrix}R_t & A_tQ_t\\ Q_tA_t' & Q_t\end{pmatrix}\begin{pmatrix}\mathbf{I} & 0\\ -A_t' & \mathbf{I}\end{pmatrix} \tag{2.72}$$

$$= \begin{pmatrix}R_t - A_tQ_tA_t' & A_tQ_t - A_tQ_t\\ Q_tA_t' & Q_t\end{pmatrix}\begin{pmatrix}\mathbf{I} & 0\\ -A_t' & \mathbf{I}\end{pmatrix} = \begin{pmatrix}R_t - A_tQ_tA_t' & 0\\ 0 & Q_t\end{pmatrix}. \tag{2.73}$$

Concluding that $Cov(\theta_t - A_tY_t, Y_t|D_{t-1}) = 0$. The weak Bayes supposes that this covariance equal to zero, under normality, implies independence. Therefore, it is possible to find the moments of the *posterior* distribution of $\theta_t|D_t$ by using the Theorem 2.3.1 item *iii*) and substituting Y_t to y_t. Then we have,

$$E(\theta_t - A_tY_t|D_t) = E(\theta_t - A_ty_t|Y_t = y_t, D_{t-1}) = a_t + A_te_t \quad \text{and} \tag{2.74}$$

$$Var(\theta_t - A_tY_t|D_t) = Var(\theta_t - A_ty_t|Y_t = y_t, D_{t-1}) = R_t - A_tQ_tA_t'. \tag{2.75}$$

2.5 Dynamic generalized linear models

In the time series context, the use of time varying regression type models is appropriate, applying to define the following dynamic generalized linear model.

Definition 2.5.1. Consider the following quantities at time t.

- θ_t, an p-dimensional state vector at time t;
- F_t, a known $p \times p$ diagonal matrix;
- G_t, a known $p \times p$ diagonal matrix;
- ω_t, an p-vector of evolution errors having zero mean and known variance matrix W_t, denoted by $\omega_t \sim [0, W_t]$;
- $\lambda_t = F_t\theta_t$, a linear function of the state vector parameters;
- $(Y_t|\eta_t)$, have a distribution which belongs to the exponential family (Section ??);
- η_t, is the natural parameter;
- $g(\eta_t)$, a known, continuous and monotonic function mapping η_t to the real line.

Then the dynamic generalized linear model (dglm) for the series $\{Y_t, t = 1, 2, \dots\}$ is defined by the following components.

Observation model:

$$p(y_t|\eta_t) \quad \text{and} \quad g(\eta_t) = \lambda_t = F_t\theta_t. \tag{2.76}$$

Evolution equation:

$$\theta_t = G_t\theta_{t-1} + \omega_t \quad \text{with} \quad \omega_t \sim [0, W_t]. \tag{2.77}$$

Consider the time series of scalar observations $\{Y_t, t = 1, 2, \dots\}$. If Y_t is assumed to have a sampling distribution in the exponential family, then the density of Y_t may be described as follows. For some defining quantities η_t and V_t, and three known functions $y_t(Y_t)$, $a(\eta_t)$ and $b(Y_t, V_t)$, the density is

$$p(Y_t|\eta_t, V_t) = \exp\{V_t^{-1}[y_t(Y_t)\eta_t - a(\eta_t)]\}b(Y_t, V_t), \tag{2.78}$$

where,

1. η_t is the natural parameter of the distribution, a continuous quantity.
2. $V_t > 0$ is a scale parameter; the precision parameter of the distribution is defined as $\phi_t = V_t^{-1}$.
3. As a function of the natural parameter for fixed Y_t, Equation 2.78, viewed as a likelihood for η_t, depends on Y_t through the transformed value $y_t(Y_t)$.
4. The function $a(\eta_t)$ is assumed twice differentiable in η_t.

The definition 2.5.1 is an extension of the standard dlm in the observation model. Given the non-normality of the observational model and in general the non-linearity of the observation mean μ_t as a function of θ_t, there is no general, exact analysis. It is important to know that $p(.)$ assumes data conditionally independently drawn from distributions with common exponential family form.

The Definition 2.3.1 provides the basic observation and evolution model at time t. To complete the model specification for time t, we need to fully define two more component distributions: (a) that of the evolution error ω_t, as yet only specified in terms of mean and variance matrix; and (b) $p(\theta_{t-1}|D_{t-1})$ that sufficiently summarises the historical information and analysis prior to time t. West and Harrison [16] showed 5 steps as an alternative for the usual forecasting and updating equations at time t. We can see, briefly, the steps 1 to 5 below.

Step 1: Prior for $\boldsymbol{\lambda_t}$

$\lambda_t = F_t\theta_t$ is a linear function of the state vector. Hence, under the prior, λ_t and θ_t have a joint prior distribution that is only partially specified in terms of moments

$$\begin{pmatrix} \lambda_t \\ \theta_t \end{pmatrix} \Bigg| \mathrm{D}_{t-1} \Bigg) \sim \left[\begin{pmatrix} f_t \\ a_t \end{pmatrix}, \begin{pmatrix} q_t & F_t R_t \\ R_t F_t^{'} & R_t \end{pmatrix} \right], \tag{2.79}$$

where,

$$f_t = F_t a_t \quad \text{and} \quad q_t = F_t R_t F_t^{'}. \tag{2.80}$$

Step 2: One-step-ahead forecasting

The sampling distribution of Y_t depends on θ_t only via $\eta_t = g^{-1}(\lambda_t)$, and thus the historical information relevant to forecasting Y_t is completely summarised in the marginal prior for $(\eta_t|D_{t-1})$. However, this is now only partially specified through the mean and variance of $\lambda_t = g(\eta_t)$ from 2.79,

$$(\lambda_t|D_{t-1}) \sim [f_t, q_t]. \tag{2.81}$$

In order to calculate the forecast distribution (and to update to the posterior for η_t), further assumptions about the form of the prior for η_t are necessary. Apart from 2.81, no further restrictions have been made on the prior. Thus, there is no prior form to be calculated or approximated in any sense, the forecaster may choose any desired form consistent with the mean a_t and variance R_t. The prior may be assumed approximately normal, for example, or to take any other convenient form. The most convenient form is that of the conjugate family, and thus a conjugate prior is supposed. This requires, of course, that such a prior can be found consistent with the mean and variance of λ_t.

Given 2.81, assume that the prior for η_t has the conjugate form namely

$$p(\eta_t|D_{t-1}) = c(r_t, s_t)\exp[r_t\eta_t - s_t a(\eta_t)]. \tag{2.82}$$

Since f_t and q_t are functions of r_t and s_t, these parameters (r_t and s_t) are chosen to be consistent with the moments for λ_t in 2.79, thus implicitly satisfying the equations

$$E(g(\eta_t)|D_{t-1}) = f_t \quad \text{and} \quad Var(g(\eta_t)|D_{t-1}) = q_t. \tag{2.83}$$

The resolution of this system is called the elicitation step. The one-step-ahead forecast distribution now follows from the density

$$p(y_t|D_{t-1}) = \frac{c(r_t, s_t)b(y_t, V_t)}{c(r_t + \phi_t y_t, s_t + \phi_t)}. \tag{2.84}$$

See Section **??** to remember who are these parameters.

Step 3: Updating for $\boldsymbol{\eta_t}$

Observing Y_t, the posterior for η_t in the conjugate form,

$$p(\eta_t|D_t) = c(r_t + \phi_t Y_t, s_t + \phi_t) \exp[(r_t + \phi_t Y_t)\eta_t - (s_t + \phi_t)a(\eta_t)]. \tag{2.85}$$

By analogy with the prior, denote the posterior mean and variance of $\lambda_t = g(\eta_t)$ by

$$f_t^* = E(g(\eta_t)|D_t) \quad \text{and} \quad q_t^* = Var(g(\eta_t)|D_t). \tag{2.86}$$

Step 4: Conditional structure for $(\theta_t|\lambda_t, D_{t-1})$

The objective of the updating is to calculate the posterior for θ_t. This can be derived from the joint posterior for λ_t and θ_t. The joint density is, by Bayes' Theorem,

$$\begin{aligned} p(\lambda_t, \theta_t|D_t) &\propto p(\lambda_t, \theta_t|D_{t-1})p(y_t|\lambda_t) \\ &\propto [p(\theta_t|\lambda_t, D_{t-1})p(\lambda_t|D_{t-1})]p(y_t|\lambda_t) \\ &\propto p(\theta_t|\lambda_t, D_{t-1})[p(\lambda_t|D_{t-1})p(y_t|\lambda_t)] \\ &\propto p(\theta_t|\lambda_t, D_{t-1})p(\lambda_t|D_t). \end{aligned} \tag{2.87}$$

Hence, given λ_t, and D_{t-1}, θ_t is conditionally independent of Y_t, and it follows that

$$p(\theta_t|D_t) = \int p(\theta_t|\lambda_t, D_{t-1})p(\lambda_t|D_t)d\lambda_t. \tag{2.88}$$

The second component in the integrand $p(\lambda_t|D_t)$ may be obtained directly from the conjugate form posterior for η_t in 2.82. The first component, defining the conditional prior for θ_t given $_t$, is, of course, not fully specified. Note, however, that to complete the updating cycle, we need to calculate only the posterior mean and variance matrix of θ_t, the full posterior remaining unspecified and indeterminate. From 2.88 the key ingredients in these calculations are the prior mean and variance matrix of $(\theta_t|\lambda_t, D_{t-1})$. Unfortunately, due to the incomplete specification of the joint prior, these conditional moments are unknown, non-linear and indeterminate functions of λ_t. They cannot be calculated without imposing further structure. However, given the partial moments specification in 2.79, they can be estimated using standard Bayesian techniques.

Step 5: Updating for θ_t.

From 2.88, it follows that

$$E(\theta_t|D_t) = E(E(\theta_t|\lambda_t, D_{t-1})|D_t), \tag{2.89}$$

and

$$Var(\theta_t|D_t) = Var(E(\theta_t|\lambda_t, D_{t-1})|D_t) + E(Var(\theta_t|\lambda_t, D_{t-1})|D_t). \tag{2.90}$$

This leads to the posterior moment

$$(\theta_t|D_t) \sim [m_t, C_t], \tag{2.91}$$

the posterior moments defined as follows. Firstly,

$$\begin{aligned} m_t &= E(E(\theta_t|\lambda_t, D_{t-1})|D_t) \\ &= E(a_t + R_t F_t(\lambda_t - f_t)/q_t|D_t) \\ &= a_t + R_t F_t E(\lambda_t|D_t) - f_t)/q_t \\ &= a_t + R_t F_t(f_t^* - f_t)/q_t. \end{aligned} \tag{2.92}$$

Similarly,

$$\begin{aligned} C_t &= Var(E(\theta_t|\lambda_t, D_{t-1})|D_t) + E(Var(\theta_t|\lambda_t, D_{t-1})|D_t) \\ &= Var(a_t + R_t F_t(\lambda_t - f_t)/q_t|D_t) + E(R_t - R_t F_t F_t^{'} R_t/q_t|D_t) \\ &= R_t F_t F_t^{'} R_t Var(\lambda_t|D_t)/q_t^2 + R_t - R_t F_t F_t^{'} R_t/q_t \\ &= R_t - R_t F_t F_t^{'} R_t(1 - q_t^*/q_t)/q_t. \end{aligned} \tag{2.93}$$

Substituting the values of f_t^* and q_t^* from Step 3 completes the updating.

2.5.1 Solving the elicitation step in dynamic models for proportions

The elicitation of hyperparameters at time t in dynamic Bayesian models for proportions is performed by solving a nonlinear system. James and José [14] show that an algorithm can solve this system when the logit function is used. If the initial conditions are satisfied, it is guaranteed that the algorithm converges to the solution.

In the following it will be shown the functions $\psi(.)$ and $\psi_1(.)$ whose definitions and some properties can be found in A.3. The $\psi(.)$ is known as the digamma function and $\psi_1(.)$ is known as the trigamma function. Without loss of generality, the elicitation step can be summarized in the following nonlinear system

$$\psi(r) - \psi(s) = f, \tag{2.94}$$

$$\psi_1(r) + \psi_1(s) = Q, \tag{2.95}$$

where the index t has been suppressed from the terms (f, Q, r, s) for clarity. Let $h : \mathbb{R}_+^2 \to \mathbb{R} \times \mathbb{R}_+$ be a real function given by

$$h(r, s) = (\psi(r) - \psi(s), \psi_1(r) + \psi_1(s)). \tag{2.96}$$

They demonstrated that h is a bijective function, which implies that the solution of the system exists and is unique.

Now, we discuss the problem of obtaining an approximate solution for the system

$$\mu(r, s) := \psi(r) - \psi(s) = f, \tag{2.97}$$

$$\sigma^2(r, s) := \psi_1(r) + \psi_1(s) = Q. \tag{2.98}$$

The pair (r^o, s^o) is considered an approximate solution to the above system if, for some fixed tolerance $\epsilon > 0$,

$$|\mu(r^o, s^o) - f_t| < \epsilon, \tag{2.99}$$

$$|\sigma^2(r^o, s^o) - Q_t| < \epsilon. \tag{2.100}$$

If $f_t = 0$, then $r^o = s^o$ and the problem reduces to solving

$$\psi_1(s^0) = \frac{Q_t}{2}. \tag{2.101}$$

The above equation can readily be solved numerically. Then consider $f_t \neq 0$. Let S be the set of all values of r and s belonging to the domain of $\mu(.,.)$ and $\sigma^2(.,.)$. Suppose that $(r^{'}, s^{'}) \in S$ is not an approximate solution. In practice, S needs to be just big enough to contain an approximate solution. Therefore, the following algorithm can be stated:

1. Choose $S_0 = (r_{inf}, r_{sup}) \times (s_{inf}, s_{sup})$ so that an approximate solution in S_0 certainly exists.

2. Set
$$r^{'} = \frac{1}{2}(r_{inf} + r_{sup}), \quad s^{'} = \frac{1}{2}(s_{inf} + s_{sup}). \tag{2.102}$$
If this pair is an approximate solution, stop the algorithm (an approximate solution has been found). Otherwise, set $i = 0$ and go to Step 3.

3. While $(r^{'}, s^{'})$ is not an approximate solution:

    (a) Calculate $f = \mu(r^{'}, s^{'})$ and $Q = \sigma^2(r^{'}, s^{'})$.

    i. If $f < f_t$ and $Q < Q_t$, set $s_{sup} = s^{'}$.
    ii. If $f < f_t$ and $Q > Q_t$, set $r_{inf} = r^{'}$.
    iii. If $f > f_t$ and $Q < Q_t$, set $r_{sup} = r^{'}$.
    iv. If $f > f_t$ and $Q > Q_t$, set $s_{inf} = s^{'}$.

    (b) Set $S_{i+1} = (r_{inf}, r_{sup}) \times (s_{inf}, s_{sup})$, $i = i + 1$ and
$$r^{'} = \frac{1}{2}(r_{inf} + r_{sup}), \quad s^{'} = \frac{1}{2}(s_{inf} + s_{sup}). \tag{2.103}$$

    (c) Test if $(r^{'}, s^{'})$ is an approximate solution.

Note that this algorithm creates a sequence of nested rectangles $S_i \supset S_{i+1}$ that converges to the solution. So far, choosing $r_{inf} = s_{inf} = 0$ and $r_{sup} = s_{sup} = 10^{\kappa}$, where κ is a natural number, has been shown to be a good strategy for application to real data. For example, in the article [14] the authors use $r_{sup} = s_{sup} = 10^8$. For this thesis, it was used $r_{sup} = s_{sup} = 10^{30}$.

Chapter 3

Building the model

In this chapter we present the model structure through the conjugate prior distribution to approximate likelihood in Equation 2.1 using the dynamic model and the forecast distribution.

3.1 A conjugate prior distribution to approximate likelihood

For each time t, let $y_{i,t} = 1$ if the i^{th} pool observed at time t generated a positive result and let $y_{i,t} = 0$ otherwise. Let n_t be the number of pools at time t; let $\dot{y}_t = \sum_{i=1}^{n_t} y_{i,t}$ the total number of positive pools at time t.

Here, the notation $L^*(.|k)$ is used to reinforce the fact that k is known. Provided that: (a) the rate is low, that is, the probability $1-(1-\pi_t)^{\bar{k}_t}$ is near to 0 and (b) the variability of k is not very high, in other words, the k have all almost the same size, [15] show that likelihood in 2.1 can be approximated by

$$L^*(\pi_t|\bar{k}_t) = (1-\pi_t)^{n_t\bar{k}_t(1-\bar{y}_t)}\left[1-(1-\pi_t)^{\bar{k}_t}\right]^{n_t\bar{y}_t}. \tag{3.1}$$

After some algebra we can write 3.1 as

$$\begin{aligned} L^*(\pi_t|\bar{k}_t) &= (1-\pi_t)^{n_t\bar{k}_t\left(1-\frac{\dot{y}_t}{n_t}\right)}\left[1-(1-\pi_t)^{\bar{k}_t}\right]^{n_t\frac{\dot{y}_t}{n_t}} \\ &= (1-\pi_t)^{\bar{k}_t(n_t-\dot{y}_t)}\left[1-(1-\pi_t)^{\bar{k}_t}\right]^{n_t\dot{y}_t} \end{aligned} \tag{3.2}$$

It is worth noting that, except for a few constants, the above equation is proportional to the likelihood of the model

$$\dot{y}_t|\bar{k}_t \sim Binomial(n_t, 1-(1-\pi_t)^{\bar{k}_t}). \tag{3.3}$$

It will be said that π_t has $BetaT(r, s|\bar{k}_t)$ distribution if its density is given by

$$p(\pi_t|\bar{k}_t) = \frac{\bar{k}_t}{B(r,s)}(1-\pi_t)^{r\bar{k}_t-1}\left[1-(1-\pi_t)^{\bar{k}_t}\right]^{s-1}, \tag{3.4}$$

where $B(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$, $r, s > 0$, and $\pi_t \in [0,1]$.

The BetaT distribution is a new continuous distribution. It can be proved that its integral $\int_0^1 p(\pi_t|\bar{k}_t)d\pi_t = 1$ and some properties are presented in the Theorem 3.1.1 below.

It is trivial to show that $BetaT(r, s|\bar{k}_t)$ is a conjugate prior for $L^*(\pi_t|\bar{k}_t)$, with the constant being $\frac{\bar{k}_t}{B(r^*,s^*)}$, where r^* and s^* can be found in the Equations 3.6 and 3.7. Considering that $\pi_t|\bar{k}_t$ has the likelihood given for 2.1 and $\pi_t|\bar{k}_t \sim BetaT(r, s|\bar{k}_t)$ we see that the *posterior* distribution is given by

$$\begin{aligned} p(\pi_t|\dot{y}_t, \bar{k}_t) &\propto p(\dot{y}_t|\pi_t, \bar{k}_t)p(\pi_t|\bar{k}_t) \\ &\propto (1-\pi_t)^{n_t\bar{k}_t(1-\bar{y}_t)}[1-(1-\pi_t)^{\bar{k}_t}]^{n_t\bar{y}_t}(1-\pi_t)^{r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1} \\ &\propto (1-\pi_t)^{n_t\bar{k}_t(1-\bar{y}_t)+r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t})]^{n_t\bar{y}_t+s-1} \\ &\propto (1-\pi_t)^{r^*-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s^*-1}, \end{aligned} \tag{3.5}$$

where

$$r^* = \bar{k}_t(n_t(1-\bar{y}_t)+r) \quad \text{and} \tag{3.6}$$

$$s^* = n_t\bar{y}_t + s. \tag{3.7}$$

Some important properties of the $BetaT(r, s|\bar{k}_t)$ distribution are shown below.

Theorem 3.1.1. *Let $\dot{y}_t|\bar{k}_t \sim Binomial(n_t, 1-(1-\pi_t)^{\bar{k}_t})$. Consider the prior distribution $\pi_t \sim BetaT(r, s, |\bar{k}_t)$.*

1. *$1-(1-\pi_t)^{\bar{k}_t} \sim Beta(s, r)$.*

2. *$\dot{y}_t|\bar{k}_t$ belongs to the exponential family with natural parameter*

$$\eta_t = \log\left(\frac{1-(1-\pi_t)^{\bar{k}_t}}{(1-\pi_t)^{\bar{k}_t}}\right). \tag{3.8}$$

3. *$BetaT(r, s|\bar{k}_t)$ belongs to the exponential family.*

4. *Some important moments:*

   (a)

$$E(\pi_t) = 1 - \frac{B(r+\frac{1}{k}, s)}{B(r,s)}; \tag{3.9}$$

(b)

$$Var(\pi_t) = \frac{B(r+\frac{2}{\bar{k}_t}, s)}{B(r,s)} - \frac{B(r+\frac{1}{\bar{k}_t}, s)^2}{B(r,s)^2}; \tag{3.10}$$

(c)

$$E(\eta_t) = \psi(s) - \psi(r); \tag{3.11}$$

(d)

$$Var(\eta_t) = \psi_1(s) + \psi_1(s). \tag{3.12}$$

Proof. 1. Considering that $p(\pi_t)$ has distribution BetaT Equation 3.4. Let's use the following transformation:

$$\underbrace{\mu_t = 1-(1-\pi_t)^{\bar{k}_t}}_{(1)} \Rightarrow (1-\pi_t)^{\bar{k}_t} = 1-\mu_t \Rightarrow \sqrt[\bar{k}_t]{1-\mu_t} = 1-\pi_t \Rightarrow \underbrace{\pi_t = 1-\sqrt[\bar{k}_t]{1-\mu_t}}_{(2)}. \tag{3.13}$$

We want to find that $\mu_t = 1-(1-\pi_t)^{\bar{k}_t} \sim Beta(s,r)$.

The Jacobian of the transformation is given by

$$|J| = \left|\frac{d\mu_t}{d\pi_t}\right| = \left|-\frac{1}{\bar{k}_t}(1-\mu_t)^{\frac{1}{\bar{k}_t}-1}(-1)\right| = \frac{1}{\bar{k}_t}(1-\mu_t)^{\frac{1}{\bar{k}_t}-1}. \tag{3.14}$$

From (1) consider $1-\mu_t = (1-\pi_t)^{\bar{k}_t}$ and from (2) consider $1-\pi_t = (1-\mu_t)^{\frac{1}{\bar{k}_t}}$. These two equations will be used in the following demonstration. Then,

$$\begin{aligned} p_{\mu_t} = p_{\pi_t}(\mu_t)|J| &= (1-\pi_t)^{r\bar{k}_t-1}(1-(1-\pi_t)^{\bar{k}_t})^{s-1}\frac{1}{\bar{k}_t}(1-\mu_t)^{\frac{1}{\bar{k}_t}-1}\frac{\bar{k}_t}{B(r,s)} \\ &= \frac{(1-\mu_t)^r}{(1-\mu_t)^{\frac{1}{\bar{k}_t}}}\mu_t^{s-1}\frac{(1-\mu_t)^{\frac{1}{\bar{k}_t}}}{(1-\mu_t)}\frac{1}{B(r,s)} \\ &= (1-\mu_t)^{r-1}\mu_t^{s-1}B(r,s)^{-1}. \end{aligned} \tag{3.15}$$

Therefore, it can be concluded that $\mu_t \sim Beta(s,r)$.

2. Let $\mu_t = 1-(1-\pi_t)^{\bar{k}_t}$. This distribution can be written in its exponential family form as

$$
\begin{aligned}
p(\dot{y}_t|\pi_t) &= \begin{pmatrix} n_t \\ \dot{y}_t \end{pmatrix} \exp\left\{\dot{y}_t \log(\mu_t) + (n_t - \dot{y}_t)\log(1-\mu_t)\right\} \\
&= \begin{pmatrix} n_t \\ \dot{y}_t \end{pmatrix} \exp\left\{\dot{y}_t \log\left(\frac{\mu_t}{1-\mu_t}\right) + n_t \log(1-\mu_t)\right\} \\
&= \begin{pmatrix} n_t \\ \dot{y}_t \end{pmatrix} \exp\left\{n_t\left(\frac{\dot{y}_t}{n_t}\log\left(\frac{\mu_t}{1-\mu_t}\right) + \log(1-\mu_t)\right)\right\}, \qquad (3.16)
\end{aligned}
$$

where $y_t(Y_t) = \frac{\dot{y}_t}{n_t}$, $\eta_t = \log\left(\frac{\mu_t}{1-\mu_t}\right)$, $V_t^{-1} = \phi_t = n_t$, $a(\eta_t) = -\log(1-\mu_t) = \log\left(\frac{1}{1-\mu_t}\right) = \log(1+\exp\{\eta_t\})$ and $b(Y_t, V_t) = \begin{pmatrix} n_t \\ \dot{y}_t \end{pmatrix}$.

3. Let's see if $p(\pi_t|\bar{k}_t)$ is in the exponential family:

$$
p(Y_t|\eta_t, V_t) = \exp\{V_t^{-1}[y_t(Y_t)\eta - a(\eta_t)]\}b(Y_t, V_t), \qquad (3.17)
$$

$$
\begin{aligned}
p(\pi_t|\bar{k}_t) &= \frac{\bar{k}_t}{B(r,s)}(1-\pi_t)^{r\bar{k}_t - 1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1} \\
&= \exp\left\{log\left(\frac{\bar{k}_t}{B(r,s)}(1-\pi_t)^{r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}\right)\right\} \\
&= \exp\{\log(\bar{k}_t) - \log(B(r,s)) + (r\bar{k}_t - 1)\log(1-\pi_t) + (s-1)\log(1-(1-\pi_t)^{\bar{k}_t})\} \\
&= \exp\{\log(\bar{k}_t) - \log(B(r,s)) + r\bar{k}_t log(1-\pi_t) + s\log(1-(1-\pi_t)^{\bar{k}_t}) - \\
&- (\log(1-\pi_t) + \log(1-(1-\pi_t)^{\bar{k}_t}))\} \\
&= \exp\{\log(\bar{k}_t - \log(B(r,s)) + r\bar{k}_t log(1-\pi_t) + s\log(1-(1-\pi_t)^{\bar{k}_t}) - \\
&- [\log((1-\pi_t)(1-(1-\pi_t)^{\bar{k}_t}))]\} \\
&= \exp\left\{\bar{k}_t\left[\log(1-\pi_t)r + \frac{\log(1-(1-\pi_t)^{\bar{k}_t})}{\bar{k}_t}s - \frac{\log(B(r,s))}{\bar{k}_t}\right]\right\} \\
&\bar{k}_t[(1-\pi_t) - (1-\pi_t)^{\bar{k}_t+1}]^{-1},
\end{aligned}
$$

where $y_t(\pi_t)_1 = log(1-\pi_t)$, $y_t(\pi_t)_2 = \frac{\log(1-(1-\pi_t)^{\bar{k}_t})}{\bar{k}_t}$, $\eta_1(r,s) = r$, $\eta_2(r,s) = s$, $V_t^{-1} = \bar{k}_t$, , $a(\eta_t) = \frac{log(B(r,s))}{\bar{k}_t}$ and $b(\pi_t, V_t) = \frac{\bar{k}_t}{(1-\pi_t)-(1-\pi_t)^{\bar{k}_t+1}}$.

4. (a)

$$
\begin{aligned}
E(1-\pi_t) &= \int (1-\pi_t)\frac{\bar{k}_t}{B(r,s)}(1-\pi_t)^{r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}d\pi_t \\
&= \frac{\bar{k}_t}{B(r,s)}\int (1-\pi_t)^{[r\bar{k}_t+1]-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}d\pi_t \\
&= \frac{\bar{k}_t}{B(r,s)}\frac{B(r+\frac{1}{\bar{k}_t},s)}{\bar{k}_t}\int (1-\pi_t)^{[r\bar{k}_t+1]-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}\frac{\bar{k}_t}{B(r+\frac{1}{\bar{k}_t},s)}d\pi_t \\
&= \frac{B(r+\frac{1}{\bar{k}_t},s)}{B(r,s)} \\
&\Rightarrow 1-E(\pi_t) = \frac{B(r+\frac{1}{\bar{k}_t},s)}{B(r,s)} \Rightarrow E(\pi_t) = 1-\frac{B(r+\frac{1}{\bar{k}_t},s)}{B(r,s)}.
\end{aligned}
\tag{3.18}
$$

(b)

$$
\begin{aligned}
E((1-\pi_t)^2)) &= \int (1-\pi_t)^2\frac{\bar{k}_t}{B(r,s)}(1-\pi_t)^{r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}d\pi_t \\
&= \frac{\bar{k}_t}{B(r,s)}\int (1-\pi_t)^{[r\bar{k}_t+2]-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}d\pi_t \\
&= \frac{\bar{k}_t}{B(r,s)}\frac{B(r+\frac{2}{\bar{k}_t},s)}{\bar{k}_t}\int (1-\pi_t)^{[r\bar{k}_t+2]-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}\frac{\bar{k}_t}{B(r+\frac{2}{\bar{k}_t},s)}d\pi_t \\
&= \frac{B(r+\frac{2}{\bar{k}_t},s)}{B(r,s)}.
\end{aligned}
\tag{3.19}
$$

Now, let's do

$$
E(1-(1-\pi_t)^2) = 1-E((1-\pi_t)^2) = 1-\frac{B(r+\frac{2}{\bar{k}_t},s)}{B(r,s)}. \tag{3.20}
$$

Observe that

$$
\begin{aligned}
E(1-(1-\pi_t)^2) &= 1-E(1-\pi_t-\pi_t+\pi_t^2) = 1-(1-E(\pi_t)-E(\pi_t)+E(\pi_t^2)) \\
&= E(\pi_t)+E(\pi_t)-E(\pi_t^2) = 1-\frac{B(r+\frac{2}{\bar{k}_t},s)}{B(r,s)} \\
&\Rightarrow E(\pi_t^2) = 2\left(1-\frac{B(r+\frac{1}{\bar{k}_t},s)}{B(r,s)}\right)-1+\frac{B(r+\frac{2}{\bar{k}_t},s)}{B(r,s)}.
\end{aligned}
\tag{3.21}
$$

And the variance is

$$
\begin{aligned}
Var(\pi_t) &= E(\pi_t^2) - E(\pi_t)^2 \\
&= 2\left(1 - \frac{B(r+\frac{1}{\bar{k}_t}, s)}{B(r,s)}\right) - 1 + \frac{B(r+\frac{2}{\bar{k}_t}, s)}{B(r,s)} - \left(1 - \frac{B(r+\frac{1}{\bar{k}_t}, s)}{B(r,s)}\right)^2 \\
&= 2 - 2\frac{B(r+\frac{1}{\bar{k}_t}, s)}{B(r,s)} - 1 + \frac{B(r+\frac{2}{\bar{k}_t}, s)}{B(r,s)} - 1 + 2\frac{B(r+\frac{1}{\bar{k}_t}, s)}{B(r,s)} - \frac{B(r+\frac{1}{\bar{k}_t}, s)^2}{B(r,s)^2} \\
&= \frac{B(r+\frac{2}{\bar{k}_t}, s)}{B(r,s)} - \frac{B(r+\frac{1}{\bar{k}_t}, s)^2}{B(r,s)^2}. \qquad (3.22)
\end{aligned}
$$

(c) Let's find its expected value:

$$E(\eta_t) = E\left(\log\left(\frac{1-(1-\pi_t)^{\bar{k}_t}}{(1-\pi_t)^{\bar{k}_t}}\right)\right).$$

Consider the following variable change:

- $x = 1 - (1-\pi_t)^{\bar{k}_t}$;

whose derivative is

- $dx = \bar{k}_t(1-\pi_t)^{\bar{k}_t-1}d\pi_t$

$$
\begin{aligned}
E(\eta_t) = &\int_0^1 \log\left(\frac{1-(1-\pi_t)^{\bar{k}_t}}{(1-\pi_t)^{\bar{k}_t}}\right)(1-\pi_t)^{r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1} \\
&\frac{\bar{k}_t}{B(r,s)}d\pi_t \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)(1-\pi_t)^{r\bar{k}_t-1}x^{s-1}\frac{\frac{\bar{k}_t}{B(r,s)}}{\bar{k}_t(1-\pi_t)^{\bar{k}_t-1}}dx \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)(1-\pi_t)^{r\bar{k}_t-1-\bar{k}_t+1}x^{s-1}\frac{1}{B(r,s)}dx \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)(1-\pi_t)^{\bar{k}_t[r-1]}x^{s-1}\frac{1}{B(r,s)}dx \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)(1-x)^{r-1}x^{s-1}B(r,s)^{-1}dx \\
&= \psi(s) - \psi(r). \qquad (3.23)
\end{aligned}
$$

Take a look in the Appendix A.3 to see the definition and properties of the digamma function ($\psi(.)$) and trigamma function ($\psi_1(.)$) for the Equation 3.25.

(d) And the variance is:

$$Var(\eta_t) = E(\eta_t^2) - E(\eta_t)^2. \tag{3.24}$$

$$\begin{aligned}
E(\eta_t^2) = &\int_0^1 \log\left(\frac{1-(1-\pi_t)^{\bar{k}_t}}{(1-\pi_t)^{\bar{k}_t}}\right)^2 (1-\pi_t)^{r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1} \\
&\frac{\bar{k}_t}{B(r,s)} d\pi_t \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)^2 (1-\pi_t)^{r\bar{k}_t-1} x^{s-1} \frac{\frac{\bar{k}_t}{B(r,s)}}{\bar{k}_t(1-\pi_t)^{\bar{k}_t-1}} dx \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)^2 (1-\pi_t)^{r\bar{k}_t-1-\bar{k}_t+1} x^{s-1} \frac{1}{B(r,s)} dx \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)^2 (1-\pi_t)^{\bar{k}_t[r-1]} x^{s-1} \frac{1}{B(r,s)} dx \\
&= \int_0^1 \log\left(\frac{x}{1-x}\right)^2 (1-x)^{r-1} x^{s-1} B(r,s)^{-1} dx \\
&= \int_0^1 \log(x)^2 (1-x)^{r-1} x^{s-1} B(r,s)^{-1} dx + \\
&+ \int_0^1 \log(1-x)^2 (1-x)^{r-1} x^{s-1} B(r,s)^{-1} dx - \\
&- 2\int_0^1 \log(x)\log(1-x)(1-x)^{r-1} x^{s-1} B(r,s)^{-1} dx \\
&= (\psi(s) - \psi(s+r))^2 + \psi_1(s) - \psi_1(s+r) + \\
&+ (\psi(r) - \psi(s+r))^2 + \psi_1(r) - \psi_1(s+r) - \\
&- 2[(\psi(s) - \psi(s+r))(\psi(r) - \psi(s+r)) - \psi_1(s+r)] \\
&= \psi(s)^2 + \psi_1(s) + \psi(r)^2 + \psi_1(r) - 2\psi(s)\psi(r).
\end{aligned} \tag{3.25}$$

Therefore,

$$Var(\eta_t) = E(\eta_t^2) - E(\eta_t)^2 = \psi_1(r) + \psi_1(s). \tag{3.26}$$

□

3.1.1 The dynamic model

The set D_{t-1} will be defined as the collection of all known information before $\dot{y}_t$ is observed (this includes the vector k_t that contains all pool sizes at time t). In this particular work it will be supposed that $D_{t-1} = \{n_t, \bar{k}_t, D_{t-2}\}$, where t is the actual time, $t-1$ and $t-2$ are one and two times backwards, respectively. The following

dynamic generalized linear model for pooled data can be constructed.
Observation model:

$$\dot{y}_t|\pi_t \approx Binomial(n_t, 1-(1-\pi_t)^{\bar{k}_t}), \tag{3.27}$$

$$\pi_t|D_{t-1} \sim BetaT(r_t, s_t|\bar{k}_t), \tag{3.28}$$

$$\lambda_t = g(\mu_t) = F_t\theta_t, \text{that is the link function.} \tag{3.29}$$

Evolution Equations:

$$\theta_t = G_t\theta_{t-1} + w_t \quad w_t \sim [0, W_t], \tag{3.30}$$

with information *a priori*:

$$\theta_0|D_0 \sim N_p(m_0, C_0). \tag{3.31}$$

Now, considering the distribution $\dot{y}_t|\pi_t \sim Binomial(n_t, 1-(1-\pi_t)^{\bar{k}_t})$.

$$p(\dot{y}_t|\pi_t) = \begin{pmatrix} n_t \\ \dot{y}_t \end{pmatrix} \mu_t^{\dot{y}_t}(1-\mu_t)^{n_t-\dot{y}_t}, \tag{3.32}$$

where $\mu_t = 1-(1-\pi_t)^{\bar{k}_t}$.

This distribution can be written in its exponential family form as

$$\begin{aligned} p(\dot{y}_t|\pi_t) &= \begin{pmatrix} n_t \\ \mu_t \end{pmatrix} \exp\{\dot{y}_t \log(\mu_t) + (n_t - \dot{y}_t)\log(1-\mu_t)\} \\ &= \begin{pmatrix} n_t \\ \mu_t \end{pmatrix} \exp\left\{\dot{y}_t \log\left(\frac{\mu_t}{1-\mu_t}\right) + n_t \log(1-\mu_t)\right\} \\ &= \begin{pmatrix} n_t \\ \mu_t \end{pmatrix} \exp\left\{n_t\left(\frac{\dot{y}_t}{n_t}\log\left(\frac{\mu_t}{1-\mu_t}\right) + \log(1-\mu_t)\right)\right\}, \end{aligned} \tag{3.33}$$

where $y_t(Y_t) = \frac{\dot{y}_t}{n_t}$, $\eta_t = \log\left(\frac{\mu_t}{1-\mu_t}\right)$, $V_t^{-1} = \phi_t = n_t$, $a(\eta_t) = -\log(1-\mu_t) = \log\left(\frac{1}{1-\mu_t}\right) = \log(1+\exp\{\eta_t\})$ and $b(Y_t, V_t) = \begin{pmatrix} n_t \\ \dot{y}_t \end{pmatrix}$.

Its conjugate priori will be

$$p(\pi_t|D_{t-1}) = \frac{\bar{k}_t}{B(r,s)}(1-\pi_t)^{r\bar{k}_t-1}[1-(1-\pi_t)^{\bar{k}_t}]^{s-1}. \tag{3.34}$$

We already know that

$$\begin{aligned} E(\eta_t) &= \psi(s) - \psi(r) \quad \text{and} \\ Var(\eta_t) &= \psi_1(r) + \psi_1(s). \end{aligned}$$

Now we can elicitate r_t and s_t solving the following system

$$\begin{cases} E(\eta_t|D_{t-1}) = f_t \\ Var(\eta_t|D_{t-1}) = q_t, \end{cases} \tag{3.35}$$

using the algorithm described in the section 2.5.1:

$$\begin{cases} \psi(s_t) - \psi(r_t) = f_t \\ \psi_1(r_t) + \psi_1(s_t) = q_t. \end{cases} \tag{3.36}$$

And we can find f_t^* and q_t^* numerically:

$$\begin{cases} \psi(s_t + y_t) - \psi(r_t + (n_t - y_t)) = f_t^* \\ \psi_1(r_t + (n_t - y_t)) + \psi_1(s_t + y_t) = q_t^*. \end{cases} \tag{3.37}$$

Using the relation between m_t and C_t used in the Step 5 we obtain the *posterior* moments of θ_t relating them to the f^* and q^* using the algorithm described in the section 2.5.

3.2 Empirical Bayes estimator for W_t

Empirical Bayes (EB) methods are procedures for statistical inference in which the *prior* distribution is estimated from the data, that is, the hyperparameters of the prior distribution are estimated. So, the EB approach uses the observed data to estimate these final stage parameters and then uses this information in the *prior* distribution.

Note that the EB approach is not fully Bayesian, since we are using the data to determine the value of these final stage parameters and is also not entirely frequentist since it relies on a prior specification of the parameter. The estimation of the hyperparameters of the *prior* distribution may be performed with the moments or likelihood methods, for example. For more details, see Mignon and Dani [9].

The following proposed model requires a *prior* knowledge of W_t. Let assume that $W_t = W = diag(w_1, \ldots, w_p)$ and let estimate these hyperparameters via EB. Consider the following hierarchical structure:

$$\dot{y}_t|\pi_t, D_{t-1} \approx Binomial(n_t, 1-(1-\pi_t)^{\bar{k}_t}) \tag{3.38}$$
$$\pi_t|D_{t-1} \sim BetaT(r_t, s_t|\bar{k}_t), \tag{3.39}$$

that implies in

$$f(\dot{y}_t|r_t, s_t, D_{t-1}) = \binom{n_t}{\dot{y}_t}\frac{B(r_t^*, s_t^*)}{B(r_t, s_t)}, \tag{3.40}$$

where r_t^* and s_t^* are given by the Equations 3.6 and 3.7. The equation 3.40 above, is the predicted function and has distribution known as Beta-Binomial.

Assume the $\Omega_t = \{r_t, s_t\}$ and that $\Omega = \{\Omega_1, \ldots, \Omega_t\}$. Therefore,

$$f(D_t|\Omega) = f(D_0)\prod_{i=1}^{t} f(\dot{y}_i|\Omega_{i-1}). \tag{3.41}$$

Now, through the Equations 3.36 and $(r,s)_t^*$ we know that Ω is a function of W. Thus, without loss of generality, we can write

$$f(D_t|\Omega(W)) \propto \prod_{i=1}^{t} f(\dot{y}_i|\Omega_{i-1}(W)), \tag{3.42}$$

where we assume that D_0 does not carry any information about W. In this way, an estimator using the EB method for W can be obtained by maximizing Equation 3.42.

Chapter 4

Application of the West Nile Virus data from Roden Island

In this chapter, the proposed model was applied in the data [WNV] to evaluate the performance of the proposed methodology. In particular, the main objective is to verify parameter estimates and the ability of the model proposed in Chapter 3 to better fit the data set under analysis.

As it was already said, it was taken from the website of the State of the Roden Island [WNV] a data about the West Nile virus (WNV), the leading cause of mosquito-borne disease in the United States. It is most typically spread to individuals by the bite of an infected mosquito. In the data there are ten variables:

1. Sort: the number of the epidemiological week;
2. Collection week: the days and months of the week;
3. Year: the year of the week;
4. Pools Tested: number of pools at time t;
5. Mosquitoes Tested: number of mosquitoes at the pools at time t;
6. Bristol: location;
7. Kent: location;
8. Newport: location;
9. Providence: location;
10. Washington: location.

The sample is the collection from June 9^{th} 2012, to September 28^{th} 2019, which is $t = 133$ weeks. Since the mosquito season starts in the summer and continues through fall it was taken from June to September and the number of the week varies between 16 and 18 each year.

Year	Pools Tested	Infected Pools
2012	2200	16
2013	2311	17
2014	1727	4
2015	2036	5
2016	1945	4
2017	1533	5
2018	1968	14
2019	2284	12

Table 4.1: Number of pools tested and infected over the years.

The Table 4.1 shows us the total number of the pools tested and infected between 2012 and 2019, that is, the sum of all weeks in 2012, the number of pools tested was 2200. And the sum of all weeks and all locations in 2012, presented only 16 pools infected. It is important to stress that since the number of infected pools is very small, it means that in the data was found a large amount of zero's.

The data was organized as following: n_t as the number of pools at time t, k_t as the mean size of pools at time t, and y_t as the number of positives pools at time t. It was necessary to use the five locations together because the data doesn't show the total per location. It is good to remember that each pool has an amount of mosquitoes to be tested. If the pool is infect, it means that there was at least one mosquito infected. If the pool is not infect, it means that there was any mosquito infected.

To estimate the probability of $\hat{\pi}_t$, the estimator proposed by Burrows [11]. The estimator is

$$\hat{\pi}^*_B = 1 - \left(\frac{2\bar{k}_t(n_t - \dot{y}_t) + \bar{k}_t - 1}{2\bar{k}_t n_t + \bar{k}_t - 1}\right)^{1/\bar{k}_t}, \tag{4.1}$$

and it was applied the following link function

$$\lambda_t = logit(\mu_t) = log\left(\frac{\mu_t}{1-\mu_t}\right) = log\left(\frac{1-(1-\pi_t)^{\bar{k}_t}}{(1-\pi_t)^{\bar{k}_t}}\right), \tag{4.2}$$

where $\mu_t = 1-(1-\pi_t)^{\bar{k}_t}$. In this problem, μ_t is the probability of having an infected pool in the week t, and $1-\mu_t$ is the probability of not having an infected pool. Having an infected pool means that there was a least one mosquito infected in the pool.

Using the WNV data we can finally find the estimators for π_t and λ_t. For each time t will have an estimated $\hat{\pi}_t$ probability. There were 87 epidemiological weeks which were estimated, from the data, the probability equal to zero, as we can see in Figure 4.1. To calculate the λ_t in these times, if we leave the probability zero we will have indeterminacy, since that, in order to be able to calculate λ_t we need to know the value of π_t previously, because λ_t depends on π_t. So that this doesn't happen, when that probability was zero, that probability was replaced by the mean value of $\hat{\pi}_t$.

We propose now a polynomial dglm of first order with

$$F_t = 1 \quad \text{and} \tag{4.3}$$

$$G_t = 1. \tag{4.4}$$

In order to adjust the dynamic model, it was necessary to find the maximum likelihood estimator for W as described in Section 3.2. With the command "optimise" in the software R [12] with the function created with all the steps showed in Section 2.2.

In Figures 4.1 and 4.2 we can compare the fluctuations for $\hat{\pi}_t$ and $\hat{\lambda}_t$ throughout the time t. These figures present the estimated values for $\hat{\pi}_t$ and $\hat{\lambda}_t$ for each time t. The fluctuations of $\hat{\lambda}_t$ seem to occur around a fluctuating level, with a constant variance.

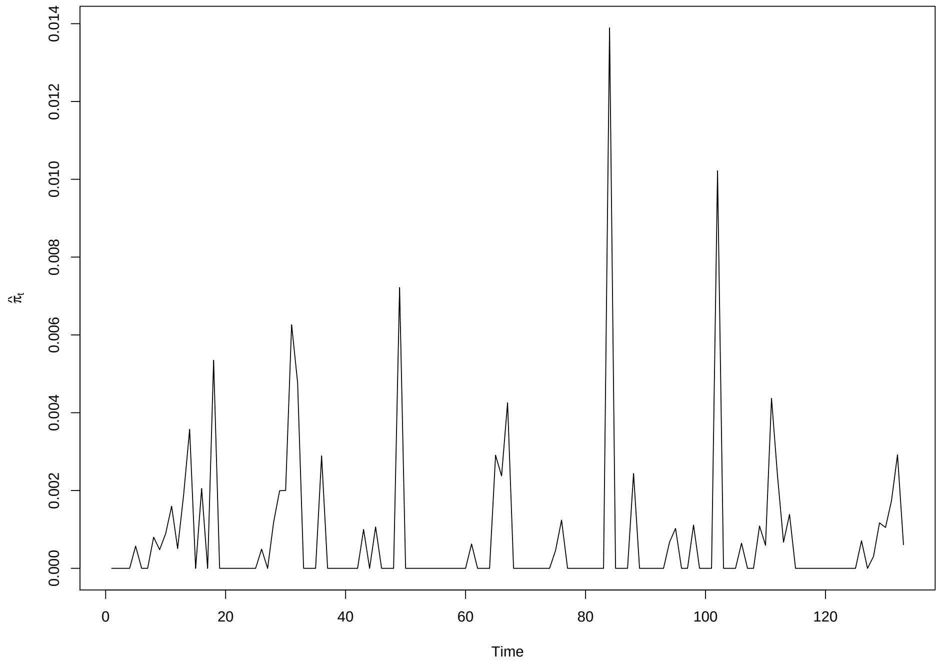

Figure 4.1: Fluctuations for $\hat{\pi}_t$

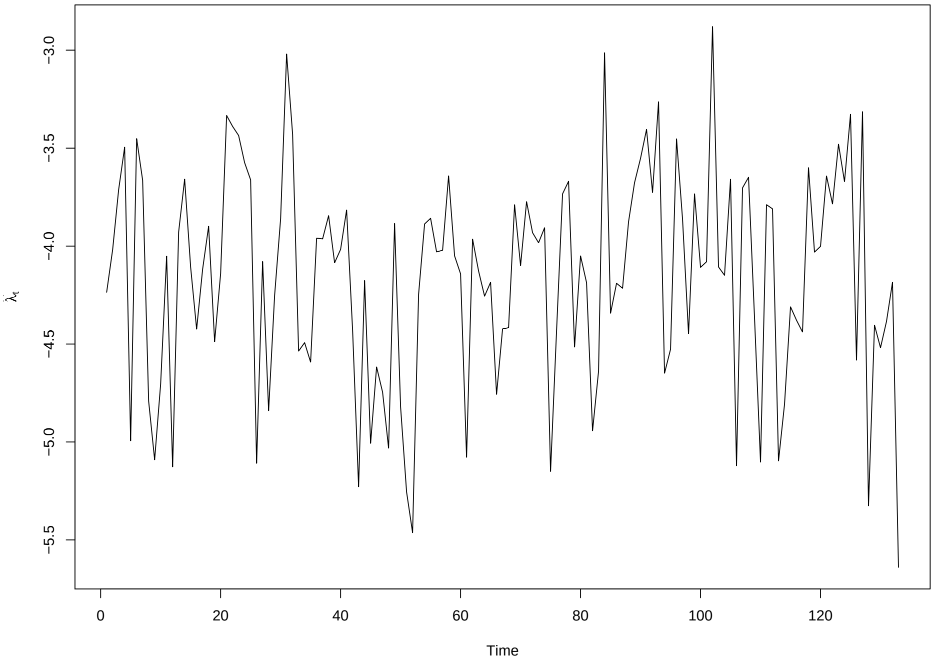

Figure 4.2: Fluctuations for $\hat{\lambda}_t$

If we apply the exponential in the link function we can obtain the odds ratio.

$$\exp\{\lambda_t\} = \exp\left\{log\left(\frac{\mu_t}{1-\mu_t}\right)\right\} = \frac{\mu_t}{1-\mu_t}. \tag{4.5}$$

Odds express the likelihood of an event occurring relative to the likelihood of an event not occurring. In the problem of this work it means the likelihood of the pool contains a mosquito with infection or not. We can see this chart in Figure 4.3. An odds ratio of 1 indicates that the event under study is equally likely to occur in both groups (pool infected or not). An odds ratio greater than 1 indicates that the event is more likely to occur in the infected group. Finally, an odds ratio less than 1 indicates that the probability is lower in the infected group than in the not infected. For $t = 70$, the odd ratio is equal to 0.014, it means that there is a very small chance to the pool in that week be infected, and consequently it's a much lower chance to the mosquito in that week be infected.

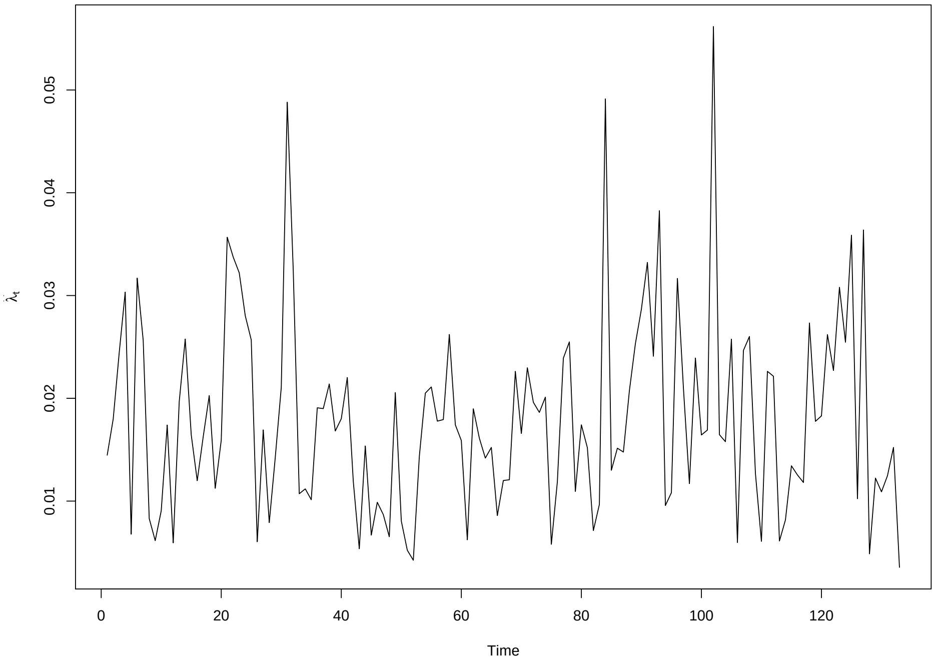

Figure 4.3: Fluctuations for $\exp\{\hat{\lambda}_t\}$

Looking to the Figure 4.3 we can see that the odd ration for all the 133 weeks is very small (less than 0.06). It means that the chance to have an infected pool is low. In practice, it is not necessary to create an alert, in this case presented. Note that the time series in the Figure 4.3 seems to be stationary in the level near to mean 0.018. Observe that the series doesn't even have trends or seasonality.

In Figure 4.4 it shows the observed $\dot{y}_t$ and its predictions, that is, the number of infected pools over the years 2012 to 2019 for each epidemiological week. We find the observed value (solid black line), the predictive mean (dashed line in red) and the predictive median (dashed line in blue). These predictions is an one-step-ahead prediction. The prediction model is a beta-binomial and its mean presented a more realistic forecasts than the median.

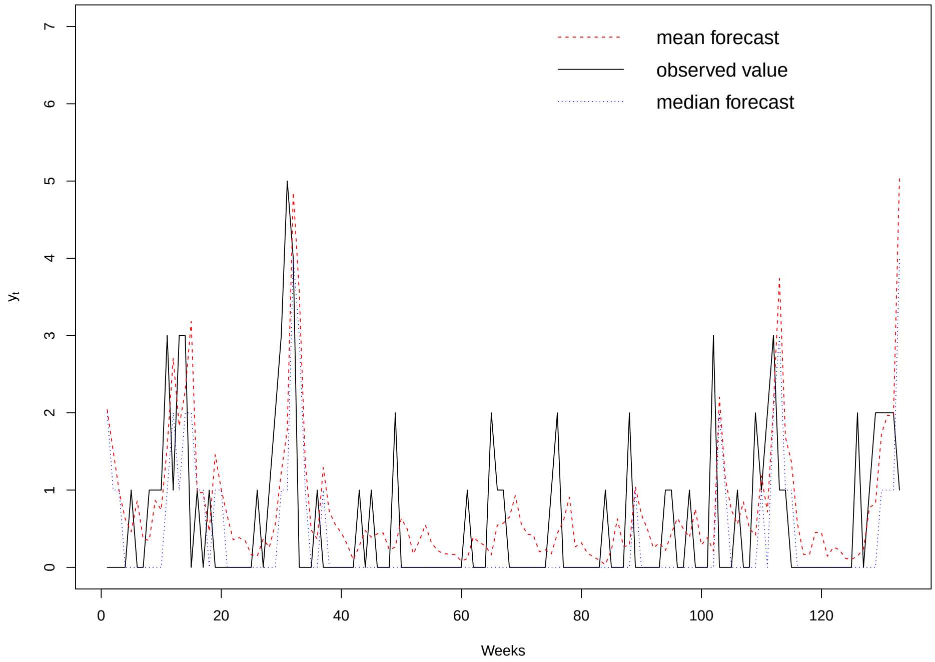

Figure 4.4: Chart of the infected pools and its estimations between 2012 and 2019

Chapter 5

Conclusion

In this thesis we worked on the real time estimation of infection rates in vectors. This work has a practical motivation; a model was found to be used as an aid tool for epidemiological alerts. In this work paper we analysed the data of the West Nile River diseases. West Nile Virus is regarded as one of the most serious mosquito-borne diseases in the United States. About 1 in 5 people who are infected develop a fever and other symptoms. About 1 out of 150 infected people develop a serious, sometimes fatal, illness, as you can see in [WNV].

To be able to help the governments to know when it is time to create an alert to help the population to fight, not only this disease, but any arbovirus diseases, it was created a model that is capable to say when it is the right time.

Using the BetaT distribution the model is defined by the following four components: observation equation, the prior distribution, the link function and the state evolution. Consider the following key components of the analysis for the BetaT dynamic model.

Observation equation:

$$\dot{y}_t|\pi_t \approx Binomial(n_t, 1-(1-\pi_t)^{\bar{k}_t}).$$

Prior:

$$\pi_t|D_{t-1} \sim BetaT(r_t, s_t; \bar{k}_t).$$

Link function: consider, without loss of generality, the logit link

$$\lambda_t = g(\mu_t) = F_t\theta_t = \log\left(\frac{\mu_t}{1-\mu_t}\right).$$

System equation:

$$\theta_t = G_t\theta_{t-1} + w_t \quad w_t \sim (0, W_t).$$

In order to begin the sequential estimation procedure we need to state the initial

information θ_0.
Initial information:

$$\theta_0 | D_0 \sim N_p(m_0, C_0).$$

The model that was proposed is a dglm using the Kalman Filter with a Bayesian approach. The dglm is the logit because it is a model that correlates the values of covariates with the mean of the binomial distribution. The predictive Beta-Binomial model (Equation 5.1) was created to predict the number of infected pools, and it was built from the *prior* and the distribution of the observations.

$$f(\dot{y}_t | r_t, s_t, D_{t-1}) = \binom{n_t}{\dot{y}_t} \frac{B(r_t^*, s_t^*)}{B(r_t, s_t)}, \tag{5.1}$$

where r_t^* and s_t^* are given by the Equations 3.6 and 3.7.

In this thesis we only worked in the fit of the model. Other features of the models are yet to be explored in future works. Besides that, for future works it will be done the monitoring and forecasting. Another propose for a future work is to verify others link functions, as the probit, square root of arcsin, and complimentary log-log, and then do a simulated study comparing these other link functions.

Observing the website [CDC] , we see the table with information on the number of cases reported to the government of the Nile Virus disease. For the State of Rhode Island, from 2012 to 2018 there were only 10 cases, as you can see in the Table 5.1, what goes according to the number of infected pools observed in the sample of the present study. Other states had a much larger number of reported cases.

It is good to remember that we can reduce our risk of WNV by using insect repellent and wearing long-sleeved shirts and long pants to prevent mosquito bites, that is the most effective way to prevent infection. Mosquitoes bite during the day and night, so, take steps to control mosquitoes indoors and outdoors. The government has made some plans to control the proliferation of the mosquitoes, such as the adult mosquito control using pesticides applied from trucks or aircraft.

Year	Reported Cases
2012	4
2013	1
2014	0
2015	0
2016	2
2017	2
2018	1

Table 5.1: Number reported cases over the years.

Appendix A

Appendix

A.1 Theorem 1:

If

$$\begin{aligned} x_1|x_2 &\sim N(\mu_1 + B_1(x_2-\mu_2), B_2) \\ x_2 &\sim N(\mu_2, \Sigma_{22}) \\ \text{Therefore,} & \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &\sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right] \end{aligned}$$

Where $\Sigma_{11} = B_2 + B_1\Sigma_{22}B_1^{'}$ and $\Sigma_{12} = B_1\Sigma_{22}$

A.2 Theorem 2:

If

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]$$

Therefore,

$$x_2|x_1 \sim N\left[\mu_2 + \Sigma_{12}^{'}\Sigma_{11}^{-1}(x_1-\mu_1), \quad \Sigma_{22} - \Sigma_{12}^{'}\Sigma_{11}^{-1}\Sigma_{12}\right]$$

A.3 Moments of logarithmically transformed random variables:

Let $X \sim Beta(\alpha, \beta)$, for $0 \leq x \leq 1$. Its density is

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}. \tag{A.1}$$

The logarithm of the geometric mean G_X of a distribution with random variable X is the arithmetic mean of $\log(X)$, or, equivalently, its expected value: $\log(G_X) = E(\log(X))$.
For a beta distribution, the expected value integral gives:

$$\begin{aligned} E(log(X)) &= \int_0^1 log(x) f(x; \alpha, \beta) dx = \int_0^1 log(x) \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} dx \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 \partial \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\partial \alpha} dx \\ &= \frac{1}{B(\alpha, \beta)} \frac{\partial}{\partial \alpha} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \\ &= \frac{1}{B(\alpha, \beta)} \frac{\partial B(\alpha, \beta)}{\partial \alpha} = \frac{\partial \log(B(\alpha, \beta))}{\partial \alpha} \\ &= \frac{\partial \log(\Gamma(\alpha))}{\partial \alpha} - \frac{\partial \log(\Gamma(\alpha + \beta))}{\partial \alpha} \\ &= \psi(\alpha) - \psi(\alpha + \beta), \end{aligned} \tag{A.2}$$

where ψ is the digamma function, that is defined as the logarithmic derivative of the gamma function:

$$\psi(\alpha) = \frac{\partial \log \Gamma(\alpha)}{\partial \alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}. \tag{A.3}$$

It is the first of the polygamma functions.

The logarithm of the geometric variance, $\log(Var(G_X))$, of a distribution with random variable X is the second moment of the logarithm of X centered on the geometric mean of X, $\log(G_X)$:

$$\begin{aligned} \log Var(G_X) &= E((\log(X) - \log(G_X))^2) \\ &= E((\log(X) - E(\log(X)))^2) \\ &= E((\log(X))^2) - (E(\log(X)))^2 \\ &= Var(\log(X)). \end{aligned} \tag{A.4}$$

For a beta distribution, higher order logarithmic moments can be derived by using the representation of a beta distribution as a proportion of two Gamma distributions and differentiating through the integral. They can be expressed in terms of higher order polygamma functions.

Here, there are some important moments:

- $E\Bigg(\log\left(\frac{X}{1-X}\right)\Bigg) = \psi(\alpha) - \psi(\beta)$;
- $E(\log^2(X)) = (\psi(\alpha) - \psi(\alpha+\beta))^2 + \psi_1(\alpha) - \psi_1(\alpha+\beta)$;
- $E(\log^2(1-X)) = (\psi(\beta) - \psi(\alpha+\beta))^2 + \psi_1(\beta) - \psi_1(\alpha+\beta)$;
- $E(\log(1-X)) = (\psi(\alpha) - \psi(\alpha+\beta))(\psi(\beta) - \psi(\alpha+\beta)) + \psi_1(\alpha+\beta)$,

where the trigamma function, denoted $\psi_1(\alpha)$, is the second of the polygamma functions, and is defined as the derivative of the digamma function:

$$\psi_1(\alpha) = \frac{\partial^2 \log\Gamma(\alpha)}{\partial\alpha^2} = \frac{\partial\psi(\alpha)}{\partial\alpha}. \tag{A.5}$$

An alternative is to use the approximation. For more information see [1].

Bibliography

[CDC] Centers for disease control and prevention. `https://www.cdc.gov/westnile/statsmaps/cumMapsData.html`. Accessed: 2019-09-30.

[Mon] Monitoring and controlling west nile virus: Are your prevention practices in place? `https://www.cdc.gov/nceh/ehs/Docs/JEH/2013/april-wnv.pdf`. Accessed: 2019-09-30.

[WNV] West Nile Virus state of rhode island: Department of health. `https://health.ri.gov/diseases/mosquitoes/?parm=109`. Accessed: 2019-09-30.

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Courier Dover Publications, 9th edition, 1965.

[2] Jeff Harrison (auth.) Andy Pole, Mike West. *Applied Bayesian Forecasting and Time Series Analysis.* Springer US, 1994. ISBN 978-0-412-04401-4,978-1-4899-3432-1. URL `http://gen.lib.rus.ec/book/index.php?md5=5dac26c5619086bb455efc312917241b`.

[3] Luc Bauwens, Michel Lubrano, and Jean-Francois Richard. *Bayesian Inference in Dynamic Econometric Models.* Oxford University Press, 2000. URL `https://EconPapers.repec.org/RePEc:oxp:obooks:9780198773139`.

[4] George Casella. An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[5] C.Q. da Silva, H.S. Migon, and L.T. Correia. Dynamic bayesian beta models. *Computational Statistics Data Analysis*, 55(6):2074–2089, 2011.

[6] Robbins H. An empirical bayes approach to statistics. 1:157–163, 1956.

[7] S. Kullback. *Information Theory and Statistics.* John Wiley Sons. Republished by Dover Publications in 1968, 1959. ISBN 0-8446-5625-9.

[8] James B. McDonald and Yexiao J. Xu. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66:133–152, 1995.

[9] Gamerman D. Louzada F. Migon, H. S. *Statistical Inference An Integrated Approach*. CHAPMAN HALL/CRC Texts in Statistical Science Series, 2nd edition, 2015.

[10] G. Petris, S. Petrone, and P. Campagnoli. *Dynamic Linear Models with R*. Use R. Springer-Verlag New York, 1 edition, 2009. ISBN 0387772375,9780387772370. URL `http://gen.lib.rus.ec/book/index.php?md5=32F8695465A6AA2C74A02D314A3AB59A`.

[11] Burrows P.M. Improved estimation of pathogen transmission rates by group testing. 77(2):363–365, 1987.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

[13] Saerkkae S. *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013. ISBN 978-1-107-03065-7,978-1-107-61928-9. URL `http://gen.lib.rus.ec/book/index.php?md5=2de3979e55774a5f6095ed5e82913d16`.

[14] James D. Santos and José M. J. Costa. An algorithm for prior elicitation in dynamic bayesian models for proportions with the logit link function, 2013.

[15] James D. Santos and Diana Dorgam. An approximate likelihood estimator for the prevalence of infections in vectors using pools of varying sizes. *Biometrical Journal*, 52:95–103, 2010.

[16] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer, 2nd edition, 1997. ISBN 0387947256.