Victor Schmidt Comitti

DYNAMIC BAYESIAN MODELS: EXTENSIONS AND NEW PROPOSALS

Belo Horizonte - MG
05/07/2019

Victor Schmidt Comitti

DYNAMIC BAYESIAN MODELS: EXTENSIONS AND NEW PROPOSALS

Tese apresentada ao Departamento de Estatística da UFMG como requisito parcial para a obtenção do título de Doutor em Estatística.

Orientador: Thiago Rezende dos Santos
Co-orientador: Fábio Nogueira Demarqui

Belo Horizonte - MG
05/07/2019

aos meus pais, Evanice e Guilherme

Agradecimentos

Aos meus pais, à minha irmã e à Nívia pelo suporte, carinho e compreensão nos momentos difíceis. Vocês são a base de tudo.

Aos meus orientadores Thiago e Fábio pela confiança, paciência, dedicação, perseverança e exemplo de profissionalismo.

Ao Fraklin, Domingos, Shikida e Frank. Professores que em algum momento me inspiraram e que fazem parte dessa jornada.

Aos professores Roger, Luiz, Rosângela, Flávio, Vinícius, Marcos e Glaura pelos cursos lecionados com dedicação exemplar.

À Ana Júlia ao Arthur e ao Uriel. Meus companheiros de sala 3041 que tornaram o caminho mais leve.

Aos meus alunos, com quem aprendi muito mais do que ensinei.

Aos meus amigos, felizmente em número grande demais para serem citados nominalmente, por tornarem a vida mais divertida.

À CAPES pelo auxílio financeiro.

Abstract

Even though many time series presents problems such as overdispersion, zero inflation and change-points, these features, usually, are not incorporated into the most common dynamic Bayesian models available in the literature. To address these problems, we worked on two strands in this dissertation. In the first strand, the objective is to introduce new Bayesian dynamic models for time series of counts that allow for observations in distributions that can more adequately adjust to some common features related to the modeling of discrete data. We present a new framework for uniparametric Dynamic Bayesian Models of counts whose particular cases include Bell, Poisson-Lindley, Yule-Simon and Borel models. Furthermore, a biparametric Negative binomial model with unknown shape parameter is provided. The inferential procedure preserves the sequential nature of the Bayesian analysis and is similar to the Dynamic Generalized Linear Models (DGLM) with a novel of incorporating Monte Carlo integration to the recursive algorithm in order to deal with the intractability of the updating distributions and an ARMS step to sample from the posterior distribution of the shape parameter. We also consider a conjugate Beta Prime of the second kind distribution prior for the mean of the process. The simulation results show a good performance of the estimators considered for the static parameter, which can be reasonably estimated. The application results also highlights a better performance of the proposed uni/biparametric models over the Poisson model. In the second strand of this work we incorporate the Product Partition Models class into the DGLM. This new formulation, that we call DGLM-PPM, retains the flexibility and generality of the DGLM class and also provides a framework for Bayesian multiple change-point detection in time series. To sample from the partition and the discount factor we use a Gibbs Sampler with an ARMS step appended. A simulation study is conducted and the results show that the proposed model is able to detect the points of regime switch in the simulated data. The superiority of our proposal over the conventional DGLM is further confirmed in two real data applications in which the DGLM-PPM outperforms the conventional DGLM in-sample and out-of-sample.

Keywords: zero inflated/overdispersed distribution, Negative binomial distribution, Parameter driven model, Dynamic generalized linear model, Bayesian inference, change-point detection, Product Partition Models.

Resumo

Embora muitas séries temporais apresentem problemas como superdispersão, inflação zero e pontos de mudança, essas características, geralmente, não são incorporadas aos modelos Bayesianos dinâmicos mais comuns disponíveis na literatura. Para resolver esses problemas, trabalhamos em duas vertentes nesta tese. Na primeira vertente, o objetivo é introduzir novos modelos dinâmicos Bayesianos para séries temporais de contagem que permitam observações em distribuições que se ajustam melhor a algumas características comuns relacionadas à modelagem de dados discretos. Apresentamos uma nova estrutura para modelos dinâmicos Bayesianos uniparamétricos de contagem cujos casos particulares incluem os modelos Bell, Poisson-Lindley, Yule-Simon e Borel. Além disso, propomos um modelo binomial negativo biparamétrico com parâmetro de forma desconhecido. O procedimento de inferência preserva a natureza seqüencial da análise Bayesiana e é semelhante ao dos Modelos Lineares Generalizados Dinâmicos (DGLM). Nossa proposta incorpora passos de integração Monte Carlo ao algoritmo recursivo para lidar com a intratabilidade das distribuições de atualização e um passo de ARMS para amostrar da distribuição a posteriori do parâmetro de forma. Também consideramos uma distribuição conjugada Beta Prime do segundo tipo para a média do processo. Os resultados de simulação mostram um bom desempenho dos estimadores considerados para o parâmetro estático do modelo mostrando que ele pode ser razoavelmente estimado. Os resultados da aplicação também destacam um melhor desempenho dos modelos uni / biparamétricos propostos sobre o modelo Poisson. Na segunda vertente deste trabalho, incorporamos a classe de Modelos de Partição Produto ao DGLM. Essa nova formulação, aqui chamada de DGLM-PPM, retém a flexibilidade e a generalidade da classe DGLM e também fornece uma estrutura para detecção de múltiplos pontos de mudança em séries temporais. Um estudo de simulação é realizado e os resultados mostram que o modelo proposto é capaz de detectar os pontos de mudança de regime nos dados simulados. A superioridade de nossa proposta em relação ao DGLM convencional é confirmada em duas aplicações a dados reais nas quais o DGLM-PPM supera o DGLM convencional em performance dentro e fora da amostra.

Palavras chave: Distribuições zero-infladas/superdispersas, Distribuição Binomial Negativa, Modelos Parameter Driven, Dynamic Generalized Linear Model, inferência Bayesian, detecção de pontos de mudança, Modelos de Partição Produto.

Contents

List of Figures

List of Tables

List of Abbreviations

ARMS Adaptive Rejection Metropolis Sampling

BE Bayesian Estimator

BF Bayes Factor

BIC Bayesian Information Criterion

CI Credible Interval

CUSUM Cumulative Sum

DBNBM Dynamic Bayesian Negative Binomial Model

DGLM Dynamic Generalized Linear Model

DLM Dynamic Linear Models

EF Exponential Family

ESE Empirical Standard Deviation

GLM Generalized Linear Model

HPD Highest Posterior Density

LBE Linear Bayesian Estimation

LLM Local Level Model

MAE Mean Absolute Error

MAP Maximum a Posteriori

MCI Monte Carlo Integration

MCMC Monte Carlo Markov Chain

MH Metropolis Hastings

MLE Maximum Likelihood Estimator

MSE Mean Squared Error

NB Negative Binomial

PLD Poisson-Lindley Distribution

PLLM Poisson Local Level Model

PPM Product Partition Models

RB Relative Bias

RMSE Root Mean Squared Error

SSM State Space Models

YSD Yule-Simon Distribution

ZII Zero Inflation Index

Chapter 1

Introduction

1.1 Motivation

A time series can be defined as a collection of observations sequentially indexed in time. The decomposition of time series into non-observable, or hidden, and independent components has attracted considerable attention in the last few decades (see Commandeur & Koopman (2007) [1]). An important class of models that allows such decomposition is that of State Space Models (SSM). These models became widespread in the literature, especially after Harvey (1989) [2] textbook treatment on the subject. The success of the SSM class is mainly due to its flexibility, interpretability, and computational efficiency. Within a frequentist framework, State Space Models are also called Structural Models and are, generally, estimated using the Kalman Filter. From a Bayesian perspective these models belong to the class of Dynamic Linear Models (DLM), extensively covered by West & Harrison (1997) [3].

The DLM class is severely limited by the normality hypothesis imposed upon the observations. There are many data sets that, even under some transformation, still cannot be adequately modeled by a Normal distribution. A manner to circumvent this restriction was proposed by West, Harrison, Migon (1985) [4] who extended to time series the regression formalism developed in Nelder & Wedderburn (1972) [5] to model data belonging to the Uniparametric Exponential Family. The authors proposed a class of Dynamic Bayesian Models that combines the sequential nature of the Bayesian inference with conjugate analysis to build closed-form predictive distributions for the observations. The models within this class are called Dynamic Generalized Linear Models (DGLM).

The great applicability of the DGLM class to problems in several areas of knowledge has intensified the research on the subject of Dynamic Bayesian Models. Gamerman & West (1987) [6] showed applications of non-Gaussian dynamic linear Bayesian models to socio-economic data sets in a context of survival analysis. Lindsay & Lambert (1995) [7] extended the class for situations in which observations are available for more than one individual. The authors work out two examples: the first, for count data, is based on an extension of the Poisson DGLM;

the other, for duration data, on an extension of the Gamma DGLM. Chiogna & Gaetan (2002) [8] model the short term relation between exposure to air pollution and general health using a Poisson DGLM with fixed effects. Other relevant and more recent works with applications of DGLM include Das (2013) and Martinez-Bello et al. (2017).

Parameter estimation for the DGLM class can be online (dynamic) or offline (when one assumes the whole sample to be available). The online inference usually follows one of three popular approaches. The first, introduced by West, Harrison, Migon (1985), assumes a conjugate prior distribution for the natural parameter of the observations distribution and rely on linear approximations for the updating steps. The second one, due to Fahrmeir (1992) [9], estimates the dynamic parameters of the model via posterior modes. This method is more adequate for dealing with multivariate time series since it does not require the user to perform numerical integrals of high order. The third approach, treated in Doucet et al. (2000) [10], is based on particle filter techniques. For a comparative analysis among the three approaches we refer the reader to Triantafyllopoulos (2009) [11]. The full offline Bayesian inference is, usually, based on Monte Carlo Markov Chain (MCMC) methods. For more details see, for example, Gammerman (1998) [12].

A somewhat different approach for the problem of inference in Dynamic Bayesian Models was presented by da Silva et al. (2011) [13]. The authors introduce a dynamic model for rates and proportions based on the Beta distribution. A non-conjugate prior is intentionally used for the mean of the process and the inference is analogous to the procedure outlined by West, Harrison, Migon (1985). Aiming to preserve the sequential nature of the Bayesian inference, MCMC methods are avoided in favor of first order approximations whenever necessary. The intractable integrals are solved numerically with the help of Gauss-Legendre quadratures and a marginal predictive distribution is built for forecasting. An important feature that makes this method very appealing is that it allows for a full Bayesian inference including point and interval estimates for the static shape parameter of the observation equation that, in most applications, is considered to be a known constant. The Dynamic Bayesian, Beta Model was later generalized into a Dynamic Bayesian Dirichlet Model by da Silva and Rodrigues (2015) [14].

Some important textbook references for Dynamic Generalized Linear Models other than West & Harrison (1997) include Fahrmeir and Tutz (2001, chapter 8) [15] and Kedem and Fokianos (2002, chapter 6) [16]. Overall, the DGLM class is a very flexible and rich one as it can account for trend, seasonality, cycle and the inclusion of covariates. Research on the field has been intense in the last thirty years and the rapid increase in computational processing capacity opened up many possibilities for the application of techniques like Monte Carlo Markov Chain and Particle Filter in the estimation of the parameters of the DGLM.

Beyond the inference, we are also concerned in this work with the problem of intervention analysis in the class of Dynamic Bayesian Models. The literature on change-point detection for these modelss has, at least, three decades. Raftery and Akman (1986) [17] proposed a Poisson process with heterogeneity to detect single change-points in count data time series . Whittaker

and Fruhwirth-Schnatter (1994) [18] introduced the change-point as a parameter in a structural model and obtained its posterior distribution at each time via a multiprocess Kalman Filter. More recently Silva and da Silva (2017) [19] presented a change-point detection method for count data time series based on an extension of the Chopin filter. The DLM class also allows for intervention analysis since the Bayesian inference permits the user to incorporate subjective information into a model. However, the change-point problem can become a major concern when the uncertainty about external events is too high. In this situations, the dynamic model can take a long time to capture a shock that changes the level of the time series leading to very bad predictions. Textbook treatment on this topic can be found in chapter 11 of West & Harrison (1997).

Another Bayesian approach of great interest to the change-point problem is the Product Partition Models (PPM) class from Hartigan (1990) [20] and Barry & Hartigan (1992) [21]. The basic idea behind those models is to partition the data set into contiguous blocks of observations that are similar to each other. The probability function of the random partitions has a product form and, within each block, the observations are conditionally independent given the parameters. The PPM class allows the user to treat the number of change-points (or blocks) in a time series as a random variable, which is a major advantage over other popular models in the literature. Almeida (2016) [22] suggests the development of a new class of models that combine the DGLM of West, Harrison & Migon (1985) with the PPM. This class permits not only the detection of multiple change-points in time series, but also online inference.

In this work we aim for two main objectives: i) to further expand the class of Dynamic Bayesian Models to new distributions that are unprecedented in the literature. We are specially interested in dynamic models for count data that can account for inflation of zeros and overdisperion. ii) to propose a new method for change-point detection within the DGLM class following the approach of Almeida (2016) [22]. Specific goals will be elaborate in more details in the next section.

1.2 Objectives

In this work we focused on the following specific objectives:

1. To present extensions of the class of Dynamic Bayesian Models for distributions outside of the Exponential Family. We are particularly interested in uniparametric distributions for count data that can serve as alternatives for the Poisson distribution in the presence of overdispersion or zero-inflation;

2. To introduce a novel scheme for a Bayesian Dynamic Negative Binomial Model (DBNBM). The approach proposed in this work is similar to that of da Silva et al. (2011). Since we wish to preserve the sequential nature of the Bayesian inference, Taylor expansions and

numerical integration techniques will be used to obtain approximations for the quantities involved in the updating steps whenever necessary;

3. To Extend the framework developed for the DBNBM to continuous biparametric distributions with support in R^{+};

4. To expand the DGLM class to allow for multiple change-points analysis. For this particular goal we will use the Product Partition Models as proposed by Hartigan (1990) and the approach suggested by Almeida (2016). A Gibbs Sampler with an appended ARMS step will be used to generate samples from the partitions and of the discount factor.

1.3 Structure of this Work

This thesis is organized as follows:

- Chapter 2 presents a brief introduction to some important concepts that will be used along this thesis;

- Chapter 3 is dedicated to the theory of Dynamic Linear Models that will be used as a starting point for the new models developed in this text;

- In Chapter 4 we propose several new dynamic Bayesian models for count data. The new cases introduced here include the Bell, Poisson-Lindley, Borel and Yule-Simon distributions. We also introduce the structure for the DBNBM based on conjugate analysis, linear approximations and Monte Carlo integration. This scheme can be further extended to other biparametric distributions that belong to the Exponential Family. The Gamma and Weibull cases are presented as illustrations in Appendix B;

- Chapter 5 introduces a new model for Bayesian change-point detection. This model, that we call DGLM-PPM, merges the DGLM of West, Harrisson and Migon (1985) to the Product Partition Model of Barry & Hartigan (1992) according to the proposal of Almeida (2016). We show that the discount factor associated with this new class can be estimated using a Metropolis step appended to the Gibb Sampling algorithm used to sample from the partitions. This procedure vastly improves the model making it a viable option for detecting multiple change-points in time series. A simulation study and real life data application with Poisson and Normal responses are also presented;

- Chapter 6 is dedicated to the final remarks and considerations regarding possible future works.

Chapter 2

Preliminary Concepts

In this chapter we present a brief overview of several concepts that are important for a better comprehension of this text.

2.1 Monte Carlo Integration

Monte Carlo methods refer to a set of simulation techniques used to approximate complex results via random sampling. The interest, usually, lies in solving integrals such as:

$$\mathsf{E}_{\mu}[f(X)] = \int_{\Omega} f(x)\mu(dx), \tag{2.1}$$

where Ω is the domain of integration, $f : \Omega \to \mathsf{R}$ is a measurable function and $\mu(dx)$ a probability measure on Ω. Monte Carlo Integration (MCI) consists in approximating expression (2.1) by averaging the function $f(x)$ over random and independent samples $X_1 \cdots X_t$ drawn from a probability density $p(.)$ that is convenient for the problem. That is:

$$\mathsf{E}_{\mu}[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n}\frac{f(X_t)}{p(X_t)}. \tag{2.2}$$

Call the Monte Carlo estimator $\bar{I}$. We can prove that $\mathsf{E}[\bar{I}] = \mathsf{E}[f(X)]$

Proof.

$$\begin{aligned}
\mathsf{E}_{\mu}[\bar{I}] &= \frac{1}{n}\mathsf{E}_{\mu}\left[\sum_{t=1}^{n}\frac{f(X_t)}{p(X_t)}\right] \\
&= \frac{1}{n}\int_{\Omega}\left(\sum_{t=1}^{n}\frac{f(x)}{p(x)}\right)p(x)\mu(dx) \\
&= \frac{1}{n}\sum_{t=1}^{n}\int_{\Omega}\frac{f(x)}{p(x)}p(x)\mu(dx) \\
&= \mathsf{E}(f(x))
\end{aligned}$$

□

This result holds if $f(x) > 0$ and $\frac{f(x)}{p(x)}$ is finite. , Thus, the Monte Carlo estimator is unbiased. In addition, the Law of Large Numbers ensures that the approximation gets more accurate as the number of random draws n increases. The convergence is achieved at the rate of $\mathsf{O}(\sqrt{n})$ for any dimension. This is a major advantage of the MCI over quadrature techniques – specially in high dimensions since the convergence for standard numerical integration becomes exponentially slower as the number of dimensions grow. Monte Carlo Integration also performs significantly better than quadratures when the target function presents singularities.

2.2 Metropolis-Hastings and Gibbs Sampler

The Bayesian inference frequently involves the computation of certain quantities, such as means and quantiles, that can only be attained through indirect sampling methods since posterior distributions are, in most situations, known only up to a normalization constant. The problem of random sampling from a distribution $f(x)$ that is analytically intractable has been extensively covered in the literature and many procedures were suggested – most of them based on simulations from a Markov Chain. These methods, known as Markov Chain Monte Carlo (MCMC), include the Metropolis-Hastings algorithm and the Gibbs Sampler.

Following Gilks, Richardson and Spiegelhalter (1995, chapter 1) [23], consider a sequence of random variables $X_0, X_1, X_2, \cdots$. They form a Markov Chain if the probability distribution that determines the next state, denoted by X_{t+1}, depends only on the current state X_t of the chain. Mathematically, this condition can be expressed as:

$$P(X_{t+1} \mid X_t, X_{t-1}, X_{t-2}, \cdots, X_0) = P(X_{t+1} \mid X_t). \tag{2.3}$$

Now, assume the chain to be time homogeneous – that is, the transition probability $P(. \mid .)$ does not depend on the time t – irreducible and aperiodic. An important property of Markov Chains with such characteristics is that, if the chain is long enough, it will reach equilibrium and the transition probabilities will converge to a stationary distribution that we will denote by $\pi(.)$:

So, as the chain runs, every new sample X_t obtained will look more and more like dependent observations drawn from the stationary distribution $\pi(x)$. Thus, in order to draw samples from an analytically intractable distribution, we just need to build an aperiodic, irreducible Markov Chain whose stationary distribution is exactly the target function desired. Posterior summaries can be easily obtained from simple ergodic averages over the values of the chain. Note, however, that the points sampled before the convergence must be discarded. The iterations needed for the chain to reach its stationary state are usually called the burn-in period.

Such a Markov Chain can be constructed using the generalization of the Metropolis algorithm developed by Hastings (1970) [24]. The method requires a proposal or candidate distribution from which samples are easy to obtain. The procedure here consists in sampling points from this proposal, that we will denote by $g(x \mid X_t)$, and accept them with a given probability. An

important aspect of the algorithm is that the proposal density may depend on the current state of the chain. When a point is accepted it becomes the new current state of the chain, otherwise the chain does not move at all. The acceptance probability is given by:

$$\alpha(X_{cur}, X^*) = min \left[1, \frac{f(X^*)g(X_{cur} \mid X^*)}{f(X_{cur})g(X^* \mid X_{cur})} \right], \tag{2.4}$$

where X^* denotes a sample drawn from the proposal and X_{cur} the current state of the chain. The full Metropolis-Hastings algorithm is described below:

Algorithm 1: Metropolis-Hastings algorithm

```
Initialize X_cur and set t = 0
for t ← 1 to N do
    sample X* from g(X | X_t)
    sample U from an Unif ∼ (0, 1)
    if U ≤ α(X_cur, X*) then
        set X_cur = X*
    else
        X_cur does not change
    end
    Increment t
end
```

Under regularity conditions, the stationary state of the chain will be the target distribution regardless of the proposal distribution chosen. However, the rate of convergence depends heavily on $g(. \mid .)$. A poor choice can lead to a highly inefficient sampling process.

So far no restrictions were imposed on the dimensionality of the target distribution. As the number of dimensions increases convergence of the Metropolis-Hastings algorithm gets progressively harder. A more computationally efficient approach in this situations is to break the parameter space into N blocks of lower dimensions and then update each of these components individually. This method, known as single component Metropolis-Hastings, is a convenient manner to reduce the dimensionality of the problem. Let X_{-i} represent the set of all the components of X except for the i-th component, that is:

$$X_{-i} = \{X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_N\}.$$

The idea of the single component Metropolis-Hastings is to sample from the components separately so that each iteration of the algorithm has N independent updating steps. The proposal distribution for the i-th component in the iteration t is denoted by $g(X_i^* \mid X_{t,i}, X_{t,-i})$, where $X_{t,-i}$ is given by:

$$X_{t,-i} = \{X_{t+1,1}; \cdots; X_{t+1,i-1}; X_{t,i+1}; \cdots .X_{t,N}\}.$$

That is, $X_{t,-i}$ represents the value of X_{-i} after the step $i-1$ of iteration $t+1$. Thus, given the proposal distribution for the i-th block $g_i(|)$, a candidate observation X_i^* of the target distribution will be accepted with probability:

$$\alpha(X_{-i}, X_i, X_i^*) = \min \left[1, \frac{\pi(X_i^* \mid X_{-i}) g(X_i \mid X_i^*, X_{-i})}{\pi(X_i \mid X_{-i}) g(X_i^* \mid X_i, X_{-i})} \right]. \tag{2.5}$$

where $\pi(X_i \mid X_{-i})$ denotes the full conditional distribution for X_i, that is: the distribution of X_i conditional on all the other components of X. Thus we can sample from the target distribution $f(X)$ by taking samples from the blocks X_i as intermediate steps. The Gibbs Sampler is a particular case of single-component Metropolis-Hastings where the proposal distribution for the updating of the i-th component are given by:

$$g(X_i^* \mid X_i, X_{-i}) = \pi(X_i^* \mid X_{-i}), \tag{2.6}$$

that is, the proposals are the full conditionals. Observe that, if we substitute equation (2.6) into (2.5) we will obtain $\alpha = 1$. Thus, in a Gibbs Sampler the candidates are always accepted.

2.3 Adaptive Rejection Metropolis Sampling

A simple method for sampling from a target distribution consists in generating samples from a proposal distribution and accept or reject them based on some condition. Suppose, for example, that samples from an auxiliary distribution $g(x)$ can be readily drawn and that there is a constant m such that $mg(x) \geq f(x)$ for all $x \in F$, where F is the domain of f. Now define $K(x) = \frac{f(x)}{mg(x)}$ and generate independent random observations X and u from g and $U \sim Unif(0,1)$, respectively. The sample X will be accepted if $u < K(X)$ and rejected otherwise; that is: if the condition is met, we accept X as being drawn from the target distribution $f(x)$. Further repetition of these steps will produce as many independent samples from f as desired. This method is known as Rejection sampling.

Note that $K(X)$ represents a probability of acceptance. Ideally one would like to choose a proposal distribution very similar to the target distribution. In this situations m would take values close to one and $K(X)$ would also be close to one, meaning that few iterations are necessary for a random draw from g to be accepted. This rarely happens in practice, though, and the Rejection Sampling algorithm may need a lot of iterations to accept the desired number of samples from the target distribution. One way around this problem is to improve the auxiliary function $g(X)$ each time a sample is rejected. For univariate bounded log-concave densities this can be achieved by the adaptive Rejection Sampling (ARS) algorithm proposed by Gilks (1992) [25]. Assume the log-concavity of $f(x)$ can be written as:

$$log[f(a)] - 2log[f(b)] + log[f(c)] < 0, \quad \text{such that } a < b < c \; \forall \; a, b, c \in \mathsf{D}. \tag{2.7}$$

The idea behind this method is to construct a piecewise linear function $h_n(x)$ that serves as an envelope for the logarithm of the log-concave target distribution. The proposal distribution is,

then, given by: $g_n(x) = \frac{1}{m_n} e^{h_n(x)}$ where m_n is a normalizing constant. The function $g_n(x)$ is piecewise exponential and can be sampled from by usual random number generation methods. So, for every proposal X rejected, the envelope $h_n(x)$ is updated and $g_n(x)$ is improved to resemble more the target function, thereby decreasing the probability of rejections.

The ARS procedure is restricted to log-concave target distributions. For non-log-concave functions Gilks and Best (1995) [26] suggest a new setting where a Metropolis-Hastings step is appended to the ARS scheme. They called this method adaptive Rejection Metropolis Sampling (ARMS). The ARMS envelope function is given by:

$$h_n(x) = max[L_{i,i+1}(x, S_n), min\{L_{i-1}i(x, S_n), L_{i+1,i+2}(x, S_n)\}], \tag{2.8}$$

where $S_n = \{x_i, i = 0, \cdots, n+1\}$ represents a set of abscissae in ascending order. Also, for $1 \leq i \leq j \leq n$, let $L_{ij}(x; S_n)$ be a segment of line going through the pair of coordinates $(x_i, \log[f(x_i)]])$ and $(x_j, \log[f(x_j)]])$. Notice that equation (2.8) does not serve as an envelope for all possible target distributions since the condition that $h_n(x) \geq \log[f(x)]$ everywhere on D will not hold if f is a non-log-concave or multimodal density. The ARMS procedure tackle this problem by allowing the proposal distribution to remain below the target in some regions of D. In these regions, the MH step appended to the algorithm will ensure that the accepted samples are drawn from the desired target distribution. Denote X_{cur} as the value of x in a given iteration, X_A an accepted draw from the proposal in the ARS step, X_M an accepted draw in the MH step and consider $g_n(x) = \frac{1}{m_n} e^{h_n(x)}$ as in ARS; then the ARMS algorithm is given by the pseudo-code below:

Algorithm 2: adaptive Metropolis Rejection Sampling Algorithm

Step 1: Initialize n and S_n independently
Step 2: Sample X from $g(x)$
Step 3: Sample U from $Unif(0, 1)$
if $U > \frac{f(X)}{e^{h_n(X)}}$ **then**
    Reject X, set $S_{n+1} = S_n \cup X$, relabel S_{n+1}, increment n, go back to step 2;
else
    Accept X, set $X_A = X$;
end
Step 4 Sample U from $Unif(0, 1)$
if $U > min\left[1, \frac{f(X_A)min\{f(X_{cur}), e^{h_n(X_{cur})}\}}{f(X_{cur})min\{f(X_A), e^{h_n(X_A)}\}}\right]$ **then**
    Reject X_A, set $X_M = X_{cur}$;
else
    Accept X_A, set $X_M = X_A$;
end
Step 5: Return X_M

When the target function is log-concave, $h_n(x)$ is an envelope for $log[f(x)]$ and the Metropolis-Hastings step is not necessary. In this case ARMS is just the regular ARS. Implementation of the ARMS algorithm is available in the 'dlm' package from R.

2.4 Fubini's Theorem

In many situations it may be desirable to interchange the order of integration in a multiple integral. Fubini's theorem provides a powerful result that assures this is possible under very general conditions. Let X and Y be σ-finite measure spaces and $f(x, y)$ a measurable function in the product space $X \times Y$. Assume, further, that the integral is absolutely convergent:

$$\int_{X\times Y} |f(x,y)|d(x,y) < \infty,$$

then,

$$\int_X \left(\int_Y f(x,y)dy \right) dx = \int_Y \left(\int_X f(x,y)dx \right) dy = \int_{X\times Y} f(x,y)d(x,y),$$

that is, the double integral may be computed via iterated integrals. Demonstration of Fubini's theorem is not in the scope of this work. The interested reader may refer to Resnick (2014, chapter 5) [27] .

A generalization of this theorem was proposed by Leonida Tonelli. The statements made by Tonelli's theorem are the exact same of Fubini's, but under less restrictive assumptions. It does not require $f(x, y)$ to be absolutely convergent, but only that f is a non-negative measurable function.

2.5 Model Comparison

In the following chapters we illustrate the models presented in this thesis with real data sets. To compare between different models we are going to use the following metrics: the Bayes Factor and the Bayesian Information Criterion (BIC) to evaluate in sample performance and the one step ahead prediction accuracy to evaluate out of sample performance.

In a Bayesian context, the Bayes Factor (henceforth BF) is a very popular method for comparing and selecting from different models. To compute the BF, let M represent a particular model. Given the whole data Y, it follows from the Bayes theorem that the posterior probability of the model M is given by:

$$p(M \mid Y) = \frac{p(Y \mid M)p(M)}{p(Y)},$$

where $p(M)$ is the prior distribution for M and $p(Y \mid M)$ the marginal predictive distribution of Y. Now suppose we have to choose from two models – M_1 and M_0, based on the same set of observations Y. The relative plausibility of the models can be measured through the BF defined as:

$$BF(M_0, M_1) = \frac{p(Y \mid M_1)}{p(Y \mid M_0)} = \frac{p(M_1 \mid Y)p(M_0)}{p(M_0 \mid Y)p(M_1)},$$

Notice that if we assign the same prior probability for each model, the Bayes Factor reduces to the ratio of Posterior Probabilities. Values of BF greater than one suggest evidence in favor of model M_1 over M_0. A deeper discussion regarding the interpretation of the Bayes Factor can be found in Kass and Raftery (1995) [28]. The authors stress that the BF summarizes the evidence provided by data in favor (or against) a given model and propose the following interpretation scale.

$BF(M_1, M_0)$	Evidence againsta M_0
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
>100	Decisive

Table 2.1: BF interpretation scale

The Bayesian Information Criterion, proposed by Schwarz (1978) [29], is an alternative to the Akaike Information Criterion (AIC) and is preferred over it for bayesian model selection. We will work with the following definition of the BIC:

$$BIC = \log(n)k - 2\log(p(Y \mid \theta, M)),$$

where k is the number of parameters in the model and n the number of observations.

To evaluate the prediction accuracy two measures were used: the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) given, respectively, by

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{y}_t - y_t\right|,$$

where T denotes the size of the data set and $\hat{y}_t$ represents the one-step-ahead forecast for the instant t.

Chapter 3

Dynamic Bayesian Models

Oksendal (1992) [30] formally defines a stochastic process as:

> A parametrized collection of random variables $\{Y_t\}_{t\in T}$ defined in a probability space $(\Omega, \mathcal{F}, P)$ that takes values in $\mathcal{R}^n$, where P is a probability measure defined in a measurable space $(\Omega, \mathcal{F})$.

If T is defined as the set, or a subset, of the integer numbers, the stochastic process is called discrete and if T is defined as $\mathcal{R}^+$ the process is called continuous. A time series is a realization of a stochastic process $\{Y_t\}$ in discrete time where each value y_t, $t \in \{0, 1, 2, 3, \cdots, T\}$, is generated from a probability distribution function F_Y that defines the process. This chapter presents an overview of a particular class of Bayesian models designed for modeling and forecasting time series: the Dynamic Linear Models (DLM) class and its extension, the Dynamic Generalized Linear Models (DGLM). The DLM class is adequate for modeling time series whose underlying process F_Y is Gaussian, whilst the DGLM class assumes that F_Y belongs to the Exponential Family.

3.1 Dynamic Linear Models

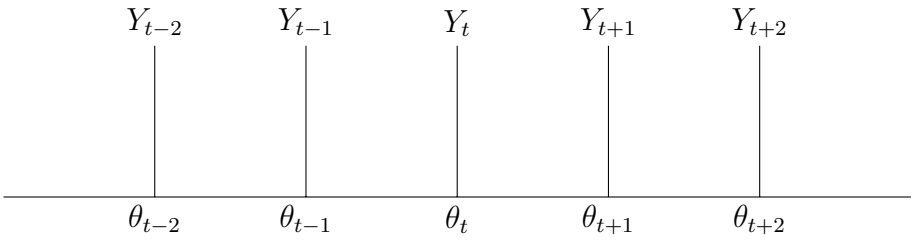
A DLM is defined, for each time t, by a quadruple $\{\mathbf{F}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{V}_t\}$. Let Y_t stand for an observation at time t. The representation of the model is, usually, given in state space form with an observation and a system (or evolution) equation as follows:

$$
\begin{aligned}
Y_t &= \mathbf{F}_\mathbf{t}'\theta_\mathbf{t} + \epsilon_t, \quad \epsilon_t \sim N[0, V_t] && \text{(Observation equation)},\\
\theta_\mathbf{t} &= \mathbf{G_t}\boldsymbol{\theta}_{t-1} + \omega_\mathbf{t}, \quad \omega_\mathbf{t} \sim T_n[0, \mathbf{W_t}] && \text{(System equation)},
\end{aligned}
$$

where T_n denotes the t-Student distribution with n degrees of freedom and $\mathbf{F}_t$, $\mathbf{G}_t$ $\boldsymbol{\theta}_t$ and $\boldsymbol{\omega}_t$, are, respectively, a known $|d \times 1|$ design vector; a $|d \times d|$ known matrix describing the parameter

evolution; the $|d \times 1|$ state vector and the evolution errors vector with a $|d \times d|$ covariance matrix $\mathbf{W_t}$, at time t.

The sequences ϵ_t and $\boldsymbol{\omega}_t$ are assumed to be internally independent and independent of each other. In addition, they are also assumed independent of the initial state $(\boldsymbol{theta}_0 \mid D_0)$. More general structures that allow for cross-dependence between the two sources of errors are possible but, since a correlated model can always be formulated as an independent one, there is no loss of generality by assuming that ϵ_t and $\boldsymbol{\omega}_t$ are mutually independent. Another key aspect of a DLM is the structure of conditional independence. That is, given $\boldsymbol{\theta_t}$, the current, future and past observations of the time series are independent of each other. This condition implies that knowledge about the future of the process is sufficiently described by the probability density functions of the current state $(\boldsymbol{\theta}_{t-1} \mid D_{t-1})$, of the future state $(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$ and of the observations $(Y_t \mid \boldsymbol{\theta}_t)$. The conditional independence assumption is illustrated in the diagram below:



The inference of a DLM is Bayesian and follows a sequential structure of evolution and updating steps. It is typically assumed that the quadruple defining the model does not depend on the time t, that is: $\{\mathbf{F}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{V}_t\} = \{\mathbf{F}, \mathbf{G}, \mathbf{W}, \mathbf{V}\}$. Further, the complete specification of the model requires an initial condition that represents the beliefs of the user about the probability distribution of the state vector at the beginning of the process and a prior distribution for the unknown variance V. From an operational point of view, it is preferable to work with the precision $\phi = \frac{1}{V}$ since it allows a full Bayesian conjugate analysis if an Inverse Gamma prior is assigned to the parameter. This initial state can be expressed as:

$$(\theta_{\mathbf{0}} \mid D_0) \sim T_{n_0}[\mathbf{m_0}, \mathbf{C_0}] \qquad \text{(Initial Condition)}$$

$$(\phi | D_0) \sim Ga\left(\frac{n_0}{2}, \frac{n_0 S_0}{2}\right) \qquad \text{(Precision Prior distribution)}$$

where T_n stands for a Student's t distribution with n degrees of freedom, $\mathbf{m}_0$ and $\mathbf{C}_0$ are prior moments and the pair $\{n_0, S_0\}$ has to be specified. The specification of the variance matrix $\mathbf{W}$ will be addressed later in this chapter. The inference procedure can be summarized by the following set of equations divided into three categories: the first for the evolution equations; the second for the forecasting distributions and the last one for the updating steps:

1. **Evolution**

$$
\begin{aligned}
\mathbf{a_t} &= \mathbf{G}_t \mathbf{m}_{t-1}, \\
\mathbf{R}_t &= \mathbf{G}_t \mathbf{C}_{t-1} \mathbf{G}_t' + \mathbf{W}_t, \\
Q_t &= \mathbf{F}_t' \mathbf{R}_t \mathbf{F}_t + S_{t-1}, \\
f_t &= \mathbf{F}_t' \mathbf{a}_t
\end{aligned}
$$

2. **Forecasting**:

$$
\begin{aligned}
(Y_t \mid D_{t-1}) &\sim T_{n-1}[f_t, Q_t], \\
(\theta_\mathbf{t} | D_{t-1}) &\sim T_{n-1}[\mathbf{a}_t, \mathbf{R}_t].
\end{aligned}
$$

3. **Updating:**

$$
\begin{aligned}
n_t &= n_{t-1} + 1, \\
S_t &= S_{t-1} + \frac{S_{t-1}}{n_t} \left(\frac{(Y_t - f_t)^2}{Q_t} - 1 \right), \\
\mathbf{m}_t &= \mathbf{a}_t + \mathbf{A}_t (Y_t - f_t), \\
\mathbf{C}_t &= \frac{S_t}{S_{t-1}} \left(\mathbf{R}_t - \mathbf{A}_t \mathbf{A}_t' Q_t \right), \\
(\phi \mid D_t) &\sim Ga \left(\frac{n_t}{2}, \frac{n_t S_t}{2} \right), \\
(\boldsymbol{\theta}_t \mid D_t) &\sim T_{n_t}[\mathbf{m}_t, \mathbf{C}_t];
\end{aligned}
$$

with $\mathbf{A}_t = \frac{\mathbf{R}_t \mathbf{F}_t}{Q_t}$.

The full inference of a DLM can be obtained by cycling through the steps above until the whole sample has been used. Naturally, the first iteration of the inference procedure requires the use of the initial state previously defined. The scheme below illustrates this process.

$$
\begin{array}{ccccc}
(\boldsymbol{\theta}_{t-1} \mid D_{t-1}) & \xrightarrow{\text{Evolution}} & (\boldsymbol{\theta}_t \mid D_{t-1}) & \xrightarrow{\text{Updating}} & (\boldsymbol{\theta}_t \mid D_t) \\
 & & \| & & \\
 & & (Y_t \mid D_{t-1}) & &
\end{array}
$$

The procedure described is usually referred as filtering. When the variance V_t is known the equations above reduce to those of the Kalman Filter. Most of the results described in this section are immediate consequences of the normality assumption imposed over the models. Complete proofs can be found in West & Harrison (1997, chapter 4) [3].

3.2 Dynamic Generalized Linear Models

In many situations the normality assumption is not realistic for many time series. In this cases, models that belong to the DLM class may not be an optimal forecasting choice. Thus, a new framework was developed by West, Harrison, Migon (1985) [4] to deal with this problem. The authors extended the DLM to allow for observations within the Exponential Family of distributions. These new models form the class that became known as Dynamic Generalized Linear Models (DGLM).

Let $\{y_t\}$ be a time series whose observations are drawn from a distribution in the Exponential Family (EF) defined as follows:

$$p(y_t \mid \eta_t, \tau_t) = \exp[\tau_t(y_t\eta_t - a(\eta_t))]c(y_t, \tau_t), \tag{3.1}$$

where the quantities η_t and τ_t are, respectively, the natural and scale parameter of the distribution. The functions $a(.)$ and $c(.)$ are assumed known with $a(\eta_t)$ being twice differentiable in η. It can be shown that

$$\mathsf{E}[y_t \mid \eta_t, \tau_t] = \frac{da(\eta_t)}{d\eta_t} \quad \text{and} \quad \mathsf{VAR}[y_t \mid \eta_t, \tau_t] = \frac{1}{\phi_t}\frac{d^2a(\eta_t)}{d\eta_t^2}. \tag{3.2}$$

The DGLM class has the same basic structure of the Generalized Linear Models (GLM) of Nelder and Wedderburn [5]. A GLM applies to conditionally independent observations that belong to the EF. The main idea behind those models is to relate the mean of the observations $\mathsf{E}[y_t \mid \eta_t, \tau_t] = \mu_t$ to a linear predictor of covariates using a suitable link function $g(.)$. In a DGLM, the linear predictor is defined as a SSM in order to incorporate the auto-correlation structure necessary to model a time series. Specifically, we consider that:

$$g(\mu_t) = \lambda_t = \mathbf{F}'\boldsymbol{\theta}_t, \tag{3.3}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim [0, \mathbf{W}_t]. \tag{3.4}$$

As before, $\boldsymbol{\theta}_t$ represents a $|d \times 1|$ state vector, $\mathbf{G}_t$ is a $|d \times d|$ evolution matrix and $\mathbf{F}_t$ is a $|d \times 1|$ design vector. In most situations, $\mathbf{G}_t$ and $\mathbf{F}_t$ are considered to be static, that is: $\mathbf{G}_t = \mathbf{G}$ and $\mathbf{F}_t = \mathbf{F}$. The link function $g(.)$ is continuous, monotonic and strictly positive. The vector $\boldsymbol{\omega}_t$ denotes a sequence of independent errors with zero mean and covariance $\mathbf{W}_t$. Unlike the DLM class, full distributional knowledge of $\boldsymbol{\omega}_t$ is, usually, not required for reasons that will become clear later. In most situations, the distribution of the errors will be defined only by its first two moments. The problem of specifying the variance matrix $\mathbf{W}_t$ is, usually, handled with the help of discount factors. We will deal with this problem in a separate section. Equation (3.3) is called

the link equation and (3.4) is the evolution or state equation. Observe that, if $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ in (3.4), i.e. there is no evolution, the DGLM reduces to the GLM class.

The inference of the DGLM class is similar to that of the DLM as it is also based on evolution and updating steps. However, since a lot of important results derived from standard Normal theory are no longer available, the whole procedure depends on several suppositions and relies heavily on linear approximations. The next section presents a detailed approach of this process.

3.2.1 Inference

The inference procedure of the DGLM class is Bayesian and, as such, it requires a prior distribution for the natural parameter (or, equivalently, to the mean) of the process. In the following sections we will assume that τ_t in equation (3.1) is a known quantity. The conjugate prior for η_t, conditional on the past information of the process D_{t-1}, can be written as:

$$p(\eta_t \mid D_{t-1}) = b(\alpha_t, \beta_t) \exp[\alpha_t \eta_t - \beta_t a(\eta_t)], \tag{3.5}$$

where α_t e β_t are hyperparameters. There are no restrictions on the form of the prior distribution, but the conjugate prior is very convenient since it allows for analytical results for the relevant quantities involved in the Bayesian analysis. The normalizing constant in (3.5) can be calculated using properties of the EF and is given by,

$$b(\alpha_t, \beta_t) = \frac{1}{\int \exp[\alpha_t \eta_t - \beta_t a(\eta_t)] d\eta_t}.$$

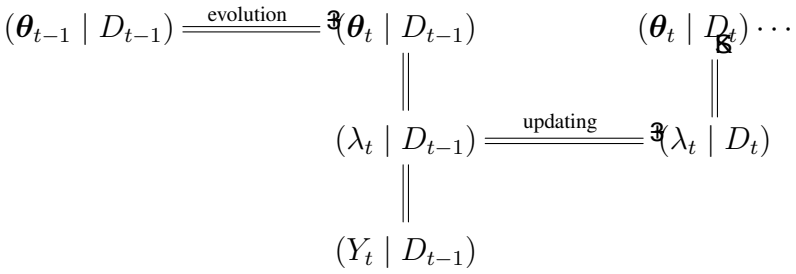
It also follows from Bayesian conjugacy that the predictive distribution of Y_t given the information up to $t-1$ can be written as:

$$p(Y_t \mid D_{t-1}) = \frac{b(\alpha_t, \beta_t) c(Y_t, \frac{1}{\tau_t})}{b(\alpha_t + \tau_t Y_t, \beta_t + \tau_t)}, \tag{3.6}$$

and the posterior distribution of η_t given the set $D_t = D_{t-1} \cup Y_t$ is a direct consequence of the previous results and the Bayes's theorem, that is:

$$p(\eta_t \mid D_t) = b(\alpha_t + \tau_t Y_t, \beta_t + \tau_t) \exp[(\alpha_t + \tau_t y_t)\eta_t - (\beta_t + \tau_t) a(\eta_t)] \tag{3.7}$$

The DGLM inference combines evolution and updating operations. In the evolution step the prior distribution of the state $\boldsymbol{\theta}_t$ is built based on the set D_{t-1} that encompasses information up to the time $t-1$. When a new observation arrives at time t, the state vector is updated in order to incorporate the new information available. This scheme is very similar to that of the DLM. The main difference being the requirement to also update the linear predictor in the DGLM inference. The diagram below illustrates the process:



The first step of the inference is to find a prior distribution for the state vector at time t conditional on the past information of the process, that is: $p(\boldsymbol{\theta}_t \mid D_{t-1})$. Knowledge of this distribution, however, requires full access to the posterior $p(\boldsymbol{\theta}_{t-1} \mid D_{t-1})$ and to the distribution of the error ω_t. It is preferable, then, to consider that $(\boldsymbol{\theta}_{t-1} \mid D_{t-1})$ and ω_t are specified only in terms of their first two moments without assuming any distribution in particular. With this hypothesis, distributional information regarding the model is lost but there is a gain in flexibility and computational efficiency since the inferential cycle can proceed sequentially and without resorting to lengthy MCMC methods. Thus, a DGLM is completely specified by the pair (3.3), (3.4) and by:

$$(\boldsymbol{\theta}_{t-1} \mid D_{t-1}) \sim [\mathbf{m}_{t-1}.\mathbf{C}_{t-1}]. \tag{3.8}$$

A direct consequence of this structure is that the prior distribution of the state vector is also specified in terms of its first moments, that is:

$$(\boldsymbol{\theta}_t \mid D_{t-1}) \sim [\mathbf{a}_t, \mathbf{R}_t], \tag{3.9}$$

where $\mathbf{a}_t = \mathbf{G_t}\mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t$. All these relations follow directly from the evolution equation (3.4).

The next step is to determine the prior distribution of the linear predictor. Observe that $\lambda_t = \mathbf{F}_t'\boldsymbol{\theta}_t = g(\eta_t)$ is a function of $\boldsymbol{\theta}_t$. Thus, the linear predictor and the state vector can be described by a joint prior distribution, also partially specified only in terms of their first and second order moments, and given by:

$$\begin{bmatrix} \lambda_t \\ \boldsymbol{\theta}_t \end{bmatrix} \Bigg| D_{t-1} \sim \left[\begin{bmatrix} f_t \\ \mathbf{a}_t \end{bmatrix}, \begin{bmatrix} q_t & \mathbf{F}_t'\mathbf{R}_t \\ \mathbf{R}_t\mathbf{F}_t' & \mathbf{R}_t \end{bmatrix} \right], \tag{3.10}$$

where $f_t = \mathbf{F}_t'\mathbf{a}_t$ and $q_t = \mathbf{F_t}'\mathbf{R}_t\mathbf{F}_t$. The hyperparameters α_t and β_t can be elicited from the moments of the linear predictor λ_t using the following equations:

$$f_t = \mathsf{E}[\lambda_t \mid D_{t-1}], \qquad q_t = \mathsf{VAR}[\lambda_t \mid D_{t-1}], \tag{3.11}$$

Both α_t and β_t must satisfy the pair (3.11), thus they are the solution of a non-linear system of equations. Under the conjugacy assumption, the posterior distribution of the natural

parameter is given by (3.7) and the posterior moments of the linear predictor can be evaluated as $f_t^* = \mathsf{E}[g(\eta_t) \mid D_t)]$ and $q_t^* = \mathsf{VAR}[g(\eta_t) \mid D_t)]$. The predictive distribution $p(Y_t \mid D_{t-1})$, in this case, can be calculated via equation (3.6). Under quadratic loss, the Bayes estimator for the quantity $Y_t \mid D_{t-1}$ is the expected value, that is: $\hat{Y}_t = \mathsf{E}(Y_t \mid D_{t-1})$, where $\hat{Y}_t$ denotes the desired point prediction.

The posterior distribution for the state vector can be derived from (3.10) using the Bayes's theorem following the sequence below:

$$\begin{aligned} p(\lambda_t, \boldsymbol{\theta}_t \mid D_t) &\propto p(\lambda_t, \boldsymbol{\theta}_t \mid D_{t-1}) p(Y_t \mid \lambda_t) \\ &\propto [p(\boldsymbol{\theta}_t \mid \lambda_t, D_{t-1}) p(\lambda_t \mid D_{t-1})] p(Y_t \mid \lambda_t) \\ &\propto p(\boldsymbol{\theta}_t \mid \lambda_t, D_{t-1}) p(y_t \mid \lambda_t). \end{aligned}$$

Then, marginally we obtain:

$$p(\boldsymbol{\theta}_t \mid D_t) = \int p(\boldsymbol{\theta}_t \mid \lambda_t, D_{t-1}) p(y_t \mid \lambda_t) d\lambda_t. \tag{3.12}$$

Observe that, by the construction of the model, the first component on the right-hand side of the equation (3.12) is not completely specified. To complete the cycle of inference, however, full distributional knowledge of $(\boldsymbol{\theta}_t \mid D_t)$ is not necessary, only its first two moments. Even so, these moments cannot be calculated from the joint prior distribution of λ_t and $\boldsymbol{\theta}_t$ without making additional hypothesis about the model. West, Harrison & Migon (1984) [4] suggest to deal with this problem using Linear Bayesian Estimation (LBE). LBE is a procedure that allows for the estimation of unknown non-linear functions through linear approximations (for more details we refer to West & Harrison 1997, section 4.9). It follows that the optimal estimators for the expected value and variance of $(\boldsymbol{\theta}_t \mid \lambda_t, D_{t-1})$ are given by:

$$\hat{\mathsf{E}}[\boldsymbol{\theta}_t \mid \lambda_t, D_{t-1}] = \mathbf{a}_t + \frac{1}{q_t} \mathbf{R}_t \mathbf{F}_t (\lambda_t - f_t),$$

$$\hat{\mathbf{VAR}}[\boldsymbol{\theta}_t \mid \lambda_t, D_{t-1}] = \mathbf{R_t} - \frac{1}{q_t} \mathbf{R_t} \mathbf{F_t} \mathbf{F_t}' \mathbf{R_t}.$$

Now, assume that $(\boldsymbol{\theta}_t \mid D_t) \sim [\mathbf{m}_t, \mathbf{C_t}]$. From the Law of Total Expectation we can approximate the first posterior moments as:

$$\begin{aligned} \mathbf{m}_t = \mathsf{E}[\boldsymbol{\theta}_t \mid D_t] &= \mathsf{E}[\hat{\mathsf{E}}[\boldsymbol{\theta}_t \mid \lambda_t, D_{t-1}] \mid D_t] \\ &= \mathsf{E}[\mathbf{a}_t + \frac{1}{q_t} \mathbf{R}_t \mathbf{F}_t (\mu_t - f_t) \mid D_t] \\ &= \mathbf{a_t} + \frac{1}{q_t} \mathbf{R}_t \mathbf{F}_t (\mathsf{E}[\mu_t \mid D_t] - f_t) \\ &= \mathbf{a}_t + \frac{1}{q_t} \mathbf{R}_t \mathbf{F}_t (f_t{}^* - f_t), \end{aligned} \tag{3.13}$$

and from the Law of Total Variance the second posterior moment will be given by,

$$\begin{aligned}
\mathbf{C}_t = \mathsf{VAR}[\boldsymbol{\theta}_t \mid D_t] &= \mathsf{VAR}[\mathsf{E}[\boldsymbol{\theta}_t \mid \mu_t, D_{t-1} \mid D_t]] + \mathsf{E}[\mathsf{VAR}[\boldsymbol{\theta}_t \mid \mu_t, D_{t-1} \mid D_t]] \\
&= \mathsf{VAR}\left[\mathbf{a}_t + \frac{1}{q_t}\mathbf{R}_t\mathbf{F}_t(\mu_t - f_t) \mid D_t\right] + \mathsf{E}\left[\mathbf{R}_t - \frac{1}{q_t}\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t \mid D_t\right] \\
&= \frac{1}{q_t^2}\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t\mathsf{VAR}[\mu_t \mid D_t] + \mathbf{R}_t - \frac{1}{q_t}\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t \\
&= \mathbf{R}_t - \frac{1}{q_t}\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t\left(1 - \frac{q_t^*}{q_t}\right). \qquad (3.14)
\end{aligned}$$

Equations (3.13) and (3.14) complete the DGLM sequential inference cycle.

3.2.2 Discount Factors

The specification of the sequence of matrices $\mathbf{W}_t$ in the system equation, of great importance for the proper adjustment of a DGLM, has been left aside so far. As already seen, at the instant $t-1$ the posterior variance of the state vector is given by $\mathsf{VAR}[\boldsymbol{\theta}_{t-1} \mid D_{t-1}] = \mathbf{C}_{t-1}$. Direct application of the system equation leads to the prior variance at t, that is – $\mathsf{VAR}[\boldsymbol{\theta}_t \mid D_{t-1}] = \mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t = \mathbf{P}_t + \mathbf{W}_t$, where $\mathbf{P}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t'$. Observe that, if $\mathbf{W}_t = 0$, then $\mathsf{VAR}[\boldsymbol{\theta}_t \mid D_{t-1}] = \mathbf{P}_t$ – that is: the model conveys no loss of information in the passage from $t-1$ to t. On the other hand, if $\mathbf{W}_t >> \mathbf{P}_t$, the stochastic error dominates the evolution implying big loss of information in the state transition. One can interpret the values of the matrix $\mathbf{W_t}$ as a measure of information loss in the passage from one state to the next. West & Harrison (1997) [3] suggest to think of $\mathbf{W}_t$ as fraction of $\mathbf{P}_t$. Under this assumption, $\mathbf{R}_t$ can be rewritten as:

$$\mathbf{R}_t = \frac{\mathbf{P}_t}{\delta},$$

which means that

$$\mathbf{W}_t = \frac{1-\delta}{\delta}\mathbf{P}_t,$$

where $\delta \in (0, 1]$ is called the discount factor.

3.2.3 Smoothing

A very important problem in time series analysis is that of estimating the state vector retrospectively based on the whole set of information available. This procedure is called smoothing. Following the notation of West & Harrison (1997) [3] the smoothing equations for the DGLM class are given by:

$$(\boldsymbol{\theta}_{\mathbf{t-h}} \mid D_t) \sim [\mathbf{a}_t(-h), \mathbf{R}_t(-h)], \qquad \text{for all t, such that } 1 \leq h \leq t;$$

where

$$\mathbf{a}_t(-h) = \mathbf{m}_{t-h} - \mathbf{B}_{t-h}[\mathbf{a}_{t-h+1} - \mathbf{a}_t(-h+1)],$$

$$\mathbf{R}_t(-h) = \mathbf{C}_{t-h} - \mathbf{B}_{t-h}[\mathbf{R}_{t-h+1} - \mathbf{R}_t(-h+1)]\mathbf{B}'_{t-h},$$

$$\mathbf{B}_t = \mathbf{C}_t\mathbf{G}'_{t+1}\mathbf{R}^{-1}_{t+1},$$

$$\mathbf{a}_t(0) = \mathbf{m}_t \qquad \text{and} \qquad \mathbf{R}_t(0) = \mathbf{C}_t.$$

3.2.4 Forecasting

The estimation of the state vector k-steps-ahead can be obtained by applying the evolution equation repeatedly. As already shown, at the time t the posterior moments of $\boldsymbol{\theta}_t$ are given by:

$$(\boldsymbol{\theta}_t, D_t) \sim [\mathbf{m}_t, \mathbf{C}_t].$$

Successive application of the evolution equation leads to:

$$\boldsymbol{\theta}_{t+k} \sim [\mathbf{a}_t(k), \mathbf{R}_t(k)],$$

where

$$\mathbf{a}_t(k) = \mathbf{G}_{t+k}\mathbf{a}_t(k-1) \qquad \text{and} \qquad \mathbf{R}_t(k) = \mathbf{G}_{t+k}\mathbf{C}_{t-k}\mathbf{G}'_{t+k} + \mathbf{W}_{t+k}$$

with $\mathbf{a}_t(0) = \mathbf{m}_t$ e $\mathbf{R}_t(0) = \mathbf{C}_t$.

The equations above also apply to DLM. For the DGLM class there is an additional step: updating the linear predictor. That is,

$$(\lambda_{t+k} \mid D_t) \sim (f_t(k), q_t(k)),$$

where $f_t(k) = \mathbf{F}'_{t+k}\mathbf{a}_t(k)$ and $q_t(k) = \mathbf{F}'_{t+k}\mathbf{R}_t(k)\mathbf{F}_{t+k}$.

Under the conjugacy assumption and using the results derived from the EF, it is possible to obtain a predictive probability density $p(y_{t+k}|D_T)$ by applying the one-step- ahead equation repeatedly. This density will be given by:

$$p(y_{t+k} \mid D_t) = \frac{b(\alpha_t(k), \beta_t(k))c(y_{t+k}, \frac{1}{\tau_{t+k}})}{b(\alpha_t(k) + \tau_{t+k}Y_{t+k}, \beta_t(k) + \tau_{t+k})}. \tag{3.15}$$

Example: Poisson DGLM

Suppose $\{y_t\}$ is a time series of counts whose observations are drawn from a Poisson distribution with parameter $\mu_t > 0$, that is:

$$p(y_t \mid \mu_t) = \frac{\exp(-\mu_t)\mu_t^{y_t}}{y_t!}, y_t = 0, 1, 2, 3, \cdots .$$

The Poisson distribution belongs to the EF (3.1) with $\tau = 1$, $a(\eta_t) = \exp(\eta_t)$, $c(y_t, \tau) = \frac{1}{y_t!}$ and $\eta_t = \log(\mu_t)$. From the pair (3.2) it is easy to see that $\mathsf{E}[y_t|\mu_t] = \mathsf{VAR}[y_t|\mu_t] = \mu_t$.

According to properties of the Exponential Family, the canonical link function is given by $\lambda_t = g(\mu_t) = F_t'\theta_t = \log(\mu_t)$ *and the conjugate prior for* μ_t *is a Gamma with parameters* α_t *and* β_t*, that is:*

$$p(\mu_t \mid D_{t-1}) = \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)}\mu_t^{\alpha_t - 1} exp(-\beta_t \mu_t). \tag{3.16}$$

The hyperparameters α_t *and* β_t *can be elicited in terms of the pair (*f_t*,* q_t*) using the following relations:*

$$f_t = \mathsf{E}[\log(\mu_t \mid D_{t-1})], \qquad q_t = \mathsf{VAR}[\log(\mu_t \mid D_{t-1})].$$

The expressions for the moments of the log of a Gamma random variable are well known and given by: $f_t = \psi(\alpha_t) - \log(\beta_t)$ *and* $q_t = \psi'(\alpha_t)$*. West & Harrison (1997) argue that the digamma (*$\psi(x)$*) and trigamma (* $\psi'(x)$*) functions can be approximated by* $log(x)$ *and* $\frac{1}{x}$ *respectively. Solving the resulting system of equations it is easy to find that* $\alpha_t = \frac{1}{q_t}$ *and* $\beta_t = \frac{\exp(-f_t)}{q_t}$*.*

From Bayes Theorem it follows that:

$$p(\mu_t \mid D_t) \propto p(y_t \mid \mu_t, D_{t-1}) p(\mu_t \mid D_{t-1}).$$

Thus, $(\mu_t \mid D_t) \sim$ *Ga*$(\alpha_t + y_t, \beta_t + 1)$*. The updating equations for the linear predictor are given by the pair below:*

$$f_t^* = \psi(y_t + \alpha_t) - \log(\beta_t + 1) \approx log(y_t + \alpha_t) - \log(\beta_t + 1),$$
$$q_t^* = \psi'(y_t + \alpha_t) \approx \frac{1}{y_t + \alpha_t}.$$

The predictive distribution can be obtained directly through equation (3.6) or marginally by integrating the expression below over μ_t*:*

$$p(y_t, \mu_t \mid D_{t-1}) = p(y_t \mid \mu_t, D_{t-1}) p(\mu_t).$$

The resulting distribution is a Negative Binomial that can be written as:

$$p(y_t \mid D_{t-1}) = \frac{\Gamma(y_t + \alpha_t)}{y_t! \Gamma(\alpha_t)} \frac{\beta_t^{\alpha_t}}{(1 + \beta_t)^{y_t + \alpha_t}}, \qquad y_t = 0, 1, 2 \cdots .$$

Under quadratic loss, the Bayesian estimator for the one-step-ahead prediction $\hat{y}_t$ *is the expected value, that is*:

$$\mathsf{E}[y_t \mid D_{t-1}] = \hat{y}_t = \frac{\alpha_t}{\beta_t}.$$

The inferential cycle ends with the updating of the state vector using equations (3.13) and (3.14).

Chapter 4

Dynamic Bayesian Linear Models for Count Data

4.1 Introduction

Modeling count data typically involves a choice between two distributions: Poisson and Negative Binomial. The Poisson is well suited for equidispersed datasets since its mean equals the variance. Most datasets, however, are not equidisperse. In this case, the Negative Binomial (NB) distribution is often a better alternative since it allows the variance of the process to be larger than its mean. The NB is a biparametric distribution whose extra parameter can be used to adjust the data variance, hence, making it a naturally overdispersed distribution.

When dealing with cross sectional count data, one can work within the well established General Linear Models (GLM) framework developed by Nelder and Wedderburn (1972) [5]. For instance, in most applications concerning non dependent count data the Poisson and Negative Binomial Regression are, usually, the first approach to the problem. For dependent count data, on the other hand, there is no consensual method. Since discrete response time series arises frequently in many fields such as Finance, Hydrology, Economics, Epidemiology and others, this has become a research topic of great interest over the last years.

Cox (1981) [31] made a distinction between two kinds of models with time varying parameters classified as Parameter Driven (PD) or Observation Driven (OD). Let Y_t represent an observation from a time series in time t. An OD model can be written as:

$$Y_t = f(D_{t-1}, \epsilon_t),$$

where D_{t-1} denotes the history of the process up to the time $t-1$ and ϵ_t is a white noise. In OD models the parameters are dynamic but their evolution is deterministic and invariant over time given past information. Parameter-Driven models have the following structure:

$$Y_t = f(\theta_{\mathbf{t}}, \nu_t)$$

$$\theta_t = g(\theta_{\mathbf{t-1}}, \eta_t),$$

again ν_t and η_t are white noise processes and θ represents a state vector that evolves as a Markovian process independently of past observations. In a PD model, the parameter evolution has an idiosyncratic error component making it non predictable even with full information of the underlying process. A modeler having to choose between an observation driven and a parameter driven model will often face a trade-off: the OD model is, usually, easy to estimate and difficult to interpret whereas the PD models have a direct interpretation but are hard to estimate because the likelihood function may be analytically intractable.

Examples of OD models include the family of volatility models derived from the Generalized Autoregressive Conditional Heterokedastic (GARCH) models of Bollerslev (1986) [32]. As for the PD models, they encompass all models that can be expressed in a state space representation where the dynamic parameter is the unobserved state variable.

Within the class of OD models, several approaches for modeling discrete valued time series came up. Many of these methods rely on the GLM theory to extend the ARMA models to this kind of data. Some important works in this line of research are Kedem and Fokianos (2002) [33], Davis, Dunsmuir and Streett (2003 [34]) and Heinen (2003) [35]. For a more extensive review on the subject the reader may refer to Fokianos (2011) [36]. Computer implementations of these methods has recently become available for the open source software R (Liboschik, Fokianos and Fried, 2017) [37].

Another important class of OD models are those based on the binomial thinning operator such as the integer-valued autoregressive (INAR) process introduced independently by McKenzie (1985) [38] and Al-Osh and Alzaid (1987) [39] that try to mimic the autocorrelation structure of an ARMA model for count data. The INAR class has been extended since then to include covariates (Brannas, 95) [40], geometric innovations (Jazi et al, 2012) [41], conditional heterocedasticity (Ferland et al, 2006) [42] and inflation of zeros (Souza, 2015) [43].

The PD models for count data belong to the wider class of state space models (SSM) (Durbin and Koopman, 2012) [44] under the Classical and Bayesian perspectives. The auto-correlation and overdispersion are, generally, introduced into the model through a hidden process. Zeger (1988) [45] proposes a log linear model for counts in which a multiplicative stationary latent process ϵ_t captures the heterogeneity of the process. Conditionally on ϵ_t the sequence of observations Y_t have the same mean and variance but, marginally, the model is overdispersed. Zeger Method does not require full distributional knowledge of the observations since the estimation is done via a quasi-likelihood scheme; only the first two moments are necessary for consistent estimation of the parameters. The idea of using a latent process to introduce overdispersion and autocorrelation into a model has been further developed in several works like Brännäs and Johansson (1994) [46], Campbell (1994) [47], Chan and Ledolter (1995) [48] and Davis, Dunsmuir and Wang (2000) [49].

From the classical perspective Durbin and Koopman, (1997) [50] extend the SSM class to non-Gaussian distributions. The likelihood of the model is approximated using Importance

Sampling techniques and the hyperparameters are obtained through numerical optimization. Gamerman, Santos and Franco (2013) [51] proposed yet another family of Non-Gaussian State Space Models (NGSSM) with exact marginal likelihood in which the previous dynamic Poisson model can be derived as a particular case.

From a Bayesian point of view Harvey and Fernandes (1989) [52] introduce a model that assumes the counts to be drawn from a Poisson distribution whose mean μ_t has prior distribution Gamma. The hyperparameters of the Gamma prior evolve according to a stochastic process aimed to inflate the prior variance over time. The likelihood of the model is constructed through a recursive algorithm. Aktekin, Soyer and Xu (2013) [53] present a Bayesian dynamic Poisson model to assess mortgage default risk in the NGSSM family. After that, Aktekin, Polson, and Soyer (2018) [54] provide a Bayesian Poisson dynamic model for modeling multivariate count data, and its sequential procedure. For further information, see Gamerman, Santos and Franco (2013) [51] and, in the count data context, a good review about these models may be found in Soyer (2018) [55].

West, Harrison and Migon(1985) [4] use an approximated Bayesian approach to propose a much more general class of non-Gaussian and non-linear models. This wider class, named as Dynamic Generalized Linear Models (DGLM), extends the GLM formalism to time series in a Bayesian framework. The analytical tractability of theses models, however, is easily lost. For a more complete overview of these methods the reader may refer to West and Harrison (1997) [3] (chapter 13). Recent developments on the subject can be found in the works of da Silva, Migon and Correia (2011) [13] and da Silva and Rodrigues (2015) [14].

Triantafyllopoulos (2009) [11] proposes a unified treatment for the DGLM class from which Poisson and Negative Binomial models are derived as particular cases. In this formulation, however, the precision parameter of the NB distribution is taken as known and has to be specified using prior knowledge of the process which is not a realistic assumption in most cases. Souza, Migon and Pereira (2018) [56] extend the DGLM class to include observation from distributions in the biparametric Exponential Family. In this framework both the mean and precision of the distributions are dynamic and different link-functions are introduced to model each parameter. This class includes discrete distributions such as the Binomial and Poisson.

The main goal of this chapter is to introduce new models for discrete response time series under a Bayesian perspective. The methods presented here are related to the DGLM family of West, Harrison and Migon, but are not restricted to the EF. In what follows, two lines of work are presented. In the first, we introduce a new general framework for uniparametric Bayesian dynamic count models. The proposed estimation procedure for these models includes the evolution and updating steps of the DGLM class but, since the use of conjugate priors is not possible most of the times, the analytic tractability of the results is lost. For this reason we incorporate Monte Carlo integration to the recursive inference algorithm in order to deal with the updating steps when no closed-form expressions are available for them. Under this new framework we are particularly interested in the following overdispersed/zero-inflated distributions: Bell, Poisson

Lindley, Yule Simon and Borel – each of them has unique features that can make them more appealing than the Poisson in some particular situations.

In the second line of work we introduce a biparametric Dynamic Bayesian Negative Binomial Model (DBNBM) for modeling overdispersed time series. Two formulations of the model are presented: in the first, the static shape parameter of the NB is taken as known within the model and, in the other, as unknown with prior distribution Gamma. We consider that the main contributions of our proposal to the literature are: i) The observation equation of the DBNBM is parametrized in terms of the mean of the process (denoted by μ_t), in the same way as in the Negative Binomial GLM, thus improving the interpretability of the model; ii) a conjugate Beta Prime of the Second Kind prior distribution for μ_t is used, which allows the model to retain analytic tractability throughout the inference cycle; iii) an efficient scheme for estimating the static shape parameter of the model using ARMS is introduced. As far as the authors are aware, this method was not used in a similar context before.

This chapter has the following structure: in Section 2 we introduce a general method for Bayesian Dynamic Models of uniparametric distributions. Several count data distribution are presented as particular cases. In Section 3 the new Dynamic Negative Binomial is introduced. Two different approaches for the model are presented along with their respective proposed inference procedures. The uniparametric models and the DNBM are applied to two real count data time series presenting overdipersion and inflation of zeros in Section 4. Finally, Section 5 is left for the final remarks.

4.2 Dynamic Bayesian Models for Uniparametric distributions

In many real life situations one may have to deal with data sets that are highly overdispersed, with an abnormal count of zeros or posses any other unique feature that the Poisson distribution may not adequately capture. In this section we present several alternatives to the Poisson DGLM. The structure of the models presented here follows closely that of the DGLM class presented in chapter 3. We focus on four distributions: Bell, Borel, Poisson-Lindley and Yule-Simon; but the method described here can be easily extended to other count uniparametric distributions such as the logarithm or Zeta distributions. The Geometric distribution was purposely omitted since it is a special case of the Negative Binomial distribution that will be addressed in the next section.

4.2.1 Some Uniparametric Distributions for Count Data

Bell distribution

Castellares et al. (2018) [57] introduce the Bell distribution and an associated regression model. The Bell probability function for $\theta > 0$ is given by

$$p(y_t \mid \theta_t) = \frac{B_{y_t}}{y_t!}\theta_t^{y_t}\exp(-\exp(\theta_t)+1), \qquad y_t = 0,1,2,\cdots,T. \tag{4.1}$$

where B_{y_t} are the Bell number defined as $B_n = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}$ and T denotes the last observation. The distribution given by expression (4.1) belongs to the EF with $\eta_t = \log(\theta_t)$, $\tau_t = 1$, $c(y_t) = \frac{B_{y_t}}{y_t!}$ and $a(\eta_t) = \exp(\exp(\eta_t)) = \exp(\theta_t)$. Using properties of the EF it is easy to prove that:

$$\mathsf{E}[Y_t \mid \theta_t] = \theta_t e^{\theta_t},$$
$$\mathsf{VAR}[Y_t \mid \theta_t] = (1+\theta_t)\theta_t e^{\theta_t}.$$

Since θ_t is strictly positive, then $\mathsf{VAR}[Y_t \mid \theta_t] > \mathsf{E}[Y_t \mid \theta_t]$, that is, Bell is an uniparametric naturally overdispersed distribution. The conjugate prior of the Bell can be found according to the properties of the EF but the normalizing constant of the resulting distribution does not have a closed-form making it hard to work with.

On the context of regression models, usually the interest lies on the mean μ_t of the process. So, letting $\mu_t = \theta_t e^{\theta_t}$ it is possible to rewrite (4.1) as:

$$p(y_t \mid \mu_t) = \frac{B_{y_t}}{y_t!}\exp(1-e^{W_0(\mu_t)})W_0(\mu_t)^{y_t}, \qquad y_t = 0,1,2,\cdots. \tag{4.2}$$

where $W_0(.)$ represents the Lambert function [1]. In this new parametrization we have that $\mathsf{E}[y_t \mid \mu_t] = \mu_t$ and $\mathsf{VAR}[y_t \mid \mu_t] = \mu_t(1+W_0(\mu_t))$.

Borel Shifted distribution

The Borel uniparametric distribution for a discrete random variable was proposed by Borel(1942) [58] in the context of branching process and queuing theory. A discrete random variable U_t follows a Borel distribution if its probability function is given by:

$$p(u_t \mid \lambda_t) = \frac{exp(-\lambda_t u_t)(\lambda_t u_t)^{u_t-1}}{u_t!}, \qquad u_t = 1,2,3,\cdots,T. \tag{4.3}$$

where $\lambda_t \in [0,1]$. The expected value and variance of (4.3) are, respectively:

$$\mathsf{E} \mid u_t \mid \lambda_t] = \frac{1}{1-\lambda_t},$$
$$\mathsf{VAR}[u_t \mid \theta_t] = \frac{\lambda_t}{(1-\lambda_t)^3}.$$

We are interested, however, in stochastic processes that can take zero values. So, defining $Y_t = U_t - 1$ and $\lambda_t = \frac{\mu_t}{(1+\mu_t)}$, it is easy to prove that the distribution for Y_t has the form:

[1] The Lambert function is the transcendental function that solves the equation $\theta\exp(\theta) = \mu$ for θ. That is, $\theta = W_0(\mu)$.

$$P(y_t \mid \mu_t) = \frac{\exp\left(\frac{\mu_t}{\mu_t+1}(y_t+1)\right)\left(\frac{\mu_t}{1+\mu_t}(y_t+1)\right)^{y_t}}{(y_t+1)!}, \qquad y_t = 0,1,2,3,... \tag{4.4}$$

where $\mu_t > 0$ represents the mean and $\frac{\mu_t}{(1+\mu_t)^2}$ the variance. Since $\frac{\mathsf{VAR}[Y_t|\mu_t]}{\mathsf{E}[Y_t|\mu_t]} = (1+\mu_t)^2 > 1$, the Borel distribution is also overdispersed. Equation (4.4) is similar to the Modified Borel Tanner distribution proposed by Gómez-Déniz et al. (2017) [59].

It is also possible to show that the Borel distribution is zero inflated, that is: the proportion of zeros drawn from a Borel is larger than the same proportion drawn from a Poisson distribution. To demonstrate that, one can use the zero-inflated index (ZII) ($z_i = 1 + \frac{\log(p_0)}{\mu_t}$, where p_0 is the probability of getting a zero value) proposed by Puig and Valero (2006) [60]. Direct application of the Index formula to the Borel distributions yields

$$z_i = 1 + \frac{1}{1+\mu_t} > 2.$$

Any $z_i > 0$ indicates a distribution that produces more zeros than the Poisson. The Borel distribution has a ZII at least greater than two which is a distinctive feature among count data distributions and makes it a candidate for modeling datasets with high frequency of zeros.

Poisson-Lindley distribution

Another way to introduce overdispersion into a model is by using mixture Poisson distributions as the observation equation. These distributions arise when we allow the rate parameter μ_t of the Poisson to be a random variable following some distribution $F(\mu_t \mid \theta)$. If $F(\mu_t \mid \theta)$ is a Lindley distribution, the resulting mixture is the Poisson-Lindley (PLD) first presented by Sankaran (1970) [61]. Its corresponding probability function for $\theta_t > 0$ is given by:

$$p(y_t \mid \theta_t) = \frac{\theta_t^2(\theta_t + y_t + 2)}{(\theta_t+1)^{y_t+3}}, y_t = 0,1,2,\cdots; \theta_t > 0, \tag{4.5}$$

with mean and variance, given respectively by,

$$\mathsf{E}[Y_t \mid \theta_t] = \frac{\theta_t+2}{\theta_t(\theta_t+1)} = \mu_t, \tag{4.6}$$

$$\mathsf{VAR}[Y_t \mid \theta_t] = \frac{\theta_t^3 + 4\theta_t^2 + 6\theta_t + 2}{\theta_t^2(\theta_t+1)}. \tag{4.7}$$

To verify that (4.5) is overdispersed notice that:

$$\frac{\mathsf{VAR}[Y_t \mid \theta_t]}{\mathsf{E}[Y_t \mid \theta_t]} = \frac{\theta_t^3 + 4\theta_t^2 + 6\theta_t + 2}{\theta_t(\theta_t+2)}. \tag{4.8}$$

It is easy to see that, since θ_t is positive, the numerator in (4.8) is always greater than the denominator, thus making the Poisson Lindley distribution overdispersed. In fact, it is a simple calculus exercise to show that the Equation (4.8) has a global minimum of, approximately, 4.33.

To parametrize the PLD in terms of the mean we solve (4.6) for θ_t to obtain $\theta_t = \frac{\sqrt{\mu_t^2+6\mu_t+1}-\mu_t+1}{2\mu_t}$. Then:

$$p(y_t|\mu_t) = \frac{\left(\frac{\sqrt{\mu_t^2+6\mu_t+1}-\mu_t+1}{2\mu_t}\right)^{y_t+2}}{\left(\frac{\sqrt{\mu_t^2+6\mu_t+1}-\mu_t+1}{2\mu_t}\right)^{y_t+3}+1}, y_t = 0,1,2,\cdots; \mu_t > 0,$$

with $\mathsf{E}[Y_t \mid \mu_t] = \mu_t$.

Yule-Simon Shifted

Yule-Simon distribution (YSD) belongs to the broader class of Power Law distributions (see [62]). It is also, like the Poisson-Lindley, a mixture distribution. Suppose υ is a random variable following an exponential distribution of rate ρ_t, then a random variable U_t following a $Geometric(e^{\upsilon})$ is Yule-Simon distributed with probability function

$$p(U_t \mid \rho_t) = \rho_t B(U_t, \rho_t + 1), \qquad X_t = 1,2,3,\cdots, \rho_t > 0, \tag{4.9}$$

where $B(.)$ represents the Beta function. Now, defining $Y_t = U_t - 1$ we can shift (4.9) to obtain a modified YSD with support in the positive integers including zero, that is:

$$p(y_t \mid \rho_t) = \rho_t B(y_t + 1, \rho_t + 1), \qquad y_t = 0,1,2,3,\cdots T, \rho_t > 0. \tag{4.10}$$

The first two central moments of (4.10) are given by:

$$\mathsf{E}[Y_t \mid \rho_t] = \frac{1}{\rho_t - 1}, \qquad \rho_t > 1,$$

$$\mathsf{VAR}[Y_t \mid \rho_t] = \frac{\rho_t^2}{(\rho_t - 1)^2(\rho_t - 2)}, \qquad \rho_t > 2.$$

The ration between variance and mean can be easily calculated as:

$$\frac{\mathsf{VAR}[Y_t \mid \rho_t]}{\mathsf{E}[Y_t \mid \rho_t]} = \frac{\rho_t^2}{(\rho_t - 1)(\rho_t - 2)}, \qquad \rho_t > 2.$$

Note that, for $\rho_t > 2$, the above expression is always positive and decreases monotonically. Using L'Hopital theorem it is easy to demonstrate that $\lim_{\rho_t \to \infty} \frac{\mathsf{VAR}[Y_t|\rho_t]}{\mathsf{E}[Y_t|\rho_t]} = 1$. Therefore, the Yule-Simon Shifted distribution is overdispersed. Direct application of the ZII also shows inflation of zeros. Reparametrizing (4.10) in terms of $\mu_t = \frac{1}{\rho_t - 1}$ gives us:

$$p(y_t \mid \mu_t) = \frac{\mu_t + 1}{\mu_t} B\left(y_t + 1, \frac{2\mu_t + 1}{\mu_t}\right), y_t = 0,1,2,\cdots. \tag{4.11}$$

where, naturally, $E[Y_t \mid \mu_t] = \mu_t$.

4.2.2 A General Framework for Dynamic Bayesian Uniparametric Models

Many uniparametric discrete data distributions are not in the EF. Even for those that belong to the EF, like the Bell distribution, very often, the conjugate prior distribution is analytically intractable. Bayesian inference for these distributions are, usually, not exact and, to obtain samples from the posterior distributions, the user has to rely on MCMC methods. In the context of the DGLM class it is usually preferred, whenever possible, not to use intensive computational methods such as the MCMC in order to preserve the sequential nature of the Bayesian inference and to avoid lengthy computational times (see, for example, [13]). We propose here a general and simple framework based on the DGLM class for dealing with Dynamic Bayesian models when the observation equation is an uniparametric distribution. This setting preserves the sequence of evolution and updating steps of the DGLM inference but does require Bayesian conjugacy since the intractable integrals involved in the updating steps can be handled via Monte Carlo Integration (MCI). It is also not necessary that the observation equation belongs to the EF. The proposed approach has the following basic structure:

- **Observation equation:**

  Any uniparametric count distribution with probability function $p(y_t \mid \mu_t)$ where $\mathsf{E}[y_t \mid \mu_t] = \mu_t$ and $\mu_t > 0$.

- **Prior distribution for μ_t:**

  Since the parameter μ_t is strictly positive any distribution with support in the interval $(0, \infty)$ may be appropriate. The Gamma distribution (3.16) arises as natural choice due to its flexibility and ease in the elicitation of its moments. Another distribution with those characteristics is the Beta Prime (BP) distribution with density:

$$p(\mu_t \mid \alpha_t, \beta_t) = \frac{\Gamma(\alpha_t + \beta_t)}{\Gamma(\alpha_t)\Gamma(\beta_t)} \frac{\mu_t^{\alpha_t - 1}}{(1 + \mu_t)^{\alpha_t + \beta_t}}. \tag{4.12}$$

- **Link Function:**

  There are several possible link functions. For the rest of this section we will assume a logarithmic one. The log-link is very suitable since it maps the strictly positive mean μ_t to the real line. It also provides closed-form expressions for some of the quantities that will be needed in the updating steps of the model. Thus, we can define:

$$g(\mu_t) = \log(\mu_t) = \mathbf{F}_t' \boldsymbol{\theta}_t. \tag{4.13}$$

  Where $\boldsymbol{\theta}_t$ represents the state vector containing non-observables components such as level, trend and seasonality. Other link functions can be viable options, but the evaluation of the parameters of the prior distribution can become cumbersome depending on the choice.

- **Evolution equation:**

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \omega_t \sim [0, \mathbf{W}_t]. \tag{4.14}$$

- **Initial Information:**

$$(\boldsymbol{\theta}_0 \mid D_0) \sim [\mathbf{m}_0, \mathbf{C}_0] \tag{4.15}$$

The inference cycle follows the steps of evolution and updating presented in Chapter 3. Applying the evolution equation we obtain the priors for the state vector $\boldsymbol{\theta}_t$ and the linear predictor λ_t. That is: $(\boldsymbol{\theta}_t \mid D_{t-1}) \sim (\mathbf{a}_t, \mathbf{R}_t)$ and $(\lambda_t \mid D_{t-1}) \sim (f_t, q_t)$. Again, we call attention to the fact that the distributions are only partially specified.

The hyperparameters α_t and β_t of the prior distribution can be elicited by matching the moments of the linear predictor with f_t and q_t as described in the DGLM Poisson example of Chapter 3. Assuming a Gamma prior distribution and a logarithm link function, we have that $\alpha_t = \frac{1}{q_t}$ and $\beta_t = \frac{\exp(-f_t)}{q_t}$, as already shown. If one chooses the Beta Prime distribution the results would be:

$$f_t = \mathsf{E}[\log(\mu_t) \mid D_{t-1}] = \psi(\alpha_t) - \psi(\beta_t),$$

$$q_t = \mathsf{VAR}[\log(\mu_t)|D_{t-1}] = \psi'(\alpha_t) + \psi'(\beta_t).$$

Proof of the results above is presented in the Appendix A1. Using the first order Taylor approximations for the digamma and trigamma functions and with some algebra we can show that $\alpha_t = \frac{\exp(-f_t)+1}{q_t}$ and $\beta_t = \frac{\exp(f_t)+1}{q_t}$.

The predictive distribution can be obtained, marginally, from the joint distribution given by $p(y_t, \mu_t \mid D_{t-1}) = p(y_t \mid \mu_t, D_{t-1})p(\mu_t \mid D_{t-1})$. That is

$$p(y_t \mid D_{t-1}) = \int_0^\infty p(y_t \mid \mu_t, D_{t-1})p(\mu_t \mid D_{t-1})d\mu_t. \tag{4.16}$$

Now, we can write $p(y_t \mid \mu_t, D_{t-1})$ as product of two functions: the first entirely dependent on the observation y_t and the other dependent on y_t and μ_t, therefore:

$$p(y_t \mid \mu_t, D_{t-1}) = Z(y_t)f(\mu_t, y_t)$$

Then, rewriting (4.16) we have:

$$\begin{aligned} p(y_t \mid D_{t-1}) &= Z(y_t)\int_0^\infty f(\mu_t, y_t)p(\mu_t \mid D_{t-1})d\mu_t \\ &= Z(y_t)\mathsf{E}[f(\mu_t)]. \end{aligned} \tag{4.17}$$

Since y_t is a known quantity we are left with the problem of evaluating the expected value of a function $f(\mu_t)$. In most cases, there is no closed form solution to $\mathsf{E}[f(\mu_t)]$. A simple approach to handle the problem is to draw samples from the prior $p(\mu_t \mid D_{t-1})$ and approximate Expression (4.17) using MCI.

The posterior distribution for the dynamic mean μ_t can be found by direct application of the Bayes' theorem and is given by:

$$p(\mu_t \mid D_t) = \frac{p(y_t \mid \mu_t)p(\mu_t \mid D_{t-1})}{p(y_t \mid D_{t-1})}.$$

Following the approach described in Chapter 3, the posterior moments of the linear predictor λ_t can be calculated from the pair: $f_t^* = \mathsf{E}[g(\mu_t) \mid D_t)]$, $q_t^* = \mathsf{VAR}[g(\mu_t) \mid D_t)]$. So, we are interested in the following quantities:

$$\begin{aligned}
\mathsf{E}[(g(\mu_t)^n \mid D_t] &= \int_0^\infty g(\mu_t)^n p(\mu_t \mid D_t) d\mu_t \\
&= \frac{Z(y_t)}{p(y_t \mid D_{t-1})} \int_0^\infty g(\mu_t)^n f(\mu_t) p(\mu_t \mid D_{t-1}) d\mu_t \\
&= \frac{Z(y_t)}{p(y_t \mid D_{t-1})} \int_0^\infty h(\mu_t) p(\mu_t \mid D_{t-1}) d\mu_t, \qquad (4.18)
\end{aligned}$$

where $n = 1, 2, \cdots$ and $h(\mu_t) = [g(\mu_t)]^n f(\mu_t)$. As before, it is possible to approximate (4.18) via MCI. Updating of the state vector $\boldsymbol{\theta}_t$ can be achieved through Linear Bayes Estimation following the procedure described in previous chapters and with the help of results (3.13) and (3.14) . The box below shows the expressions of $Z(y_t)$ and $f(\mu_t)$ for the distributions described in the Subsection 4.2.1.

Distribution	$Z(y_t)$	$f(\mu_t)$
Bell	$\frac{B_{y_t}}{y_t!}$	$exp(1 - e^{W_0(\mu_t)})W_0(\mu_t)^{y_t}$
Borel shifted	$\frac{1}{(y_t+1)!}$	$exp\ \frac{\mu_t}{\mu_t+1}(y_t+1)^{y_t}$
Poisson-Lindley	1	$\frac{\mu_t}{1+\mu_t}(y_t+1)^{y_t}$
Yule-Simon shifted	1	$\frac{\mu_t+1}{\mu_t} B\ y_t+1, \frac{2\mu_t+1}{\mu_t}$

Table 4.1: Corresponding $f(\cdot)$ and $Z(\cdot)$ functions of the predictive distribution for each uniparametric distribution.

4.3 A Biparametric Dynamic Bayesian Negative Binomial Model

In this section we present an approach for a Dynamic Bayesian Negative Binomial Model (DBNBM). The observation equation of this model is parametrized as in the Negative Binomial

regression model from GLM and a prior conjugate distribution is assigned for the mean μ_t of the process. Let y_t, $t = 1, 2, \cdots, T$ be a time series of counts, the DBNBM can be defined:

- **Observation equation:**

$$p(y_t \mid \mu_t, \kappa) = \frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t + 1)\Gamma(\kappa^{-1})} \frac{(\kappa\mu_t)^{y_t}}{(1 + \kappa\mu_t)^{y_t + \kappa^{-1}}}. \tag{4.19}$$

  The Binomial distribution has two parameters: the dynamic mean μ_t and a static parameter $\kappa > 0$ associated with the variance of the process. If the parameter κ is known, equation (4.19) belongs to the Exponential family with $\phi_t = 1$, $\eta_t = log\left(\frac{\kappa}{1+\kappa\mu_t}\right)$, $c(y_t, \phi_t) = \frac{\Gamma\left(y_t + \kappa^{-1}\right)}{\Gamma(y_t+1)\Gamma(\kappa^{-1})}$. Using the Equation (3.2) we can prove that $\mathsf{E}[y_t \mid \mu_t] = \mu_t$ and $\mathsf{VAR} = \mu_t + \kappa\mu_t$. Observe that the geometric distribution is a particular case of the Negative Binomial when $\kappa = 1$.

- **Prior distribution for μ_t:**

$$p(\mu_t \mid \kappa, \alpha_t, \beta_t D_{t-1}) = \frac{\kappa^{\alpha_t}}{B(\alpha_t, \beta_t)} \frac{\mu_t^{\alpha_t - 1}}{(1 + \kappa\mu_t)^{\alpha_t + \beta_t}}. \tag{4.20}$$

  where $\alpha_t > 0$ and $\beta_t > 0$. Equation (4.20) is a particular case of the Generalized Beta of the second kind distribution. This distribution was chosen because it is the conjugate prior of the NB in the parametrization defined in (4.19).

- **Link Function:**

$$g(\mu_t) = \log(\mu_t) = \mathbf{F}_t'\boldsymbol{\theta}_t. \tag{4.21}$$

- **Evolution equation:**

$$\theta_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \omega_t \sim [0, \mathbf{W_t}]. \tag{4.22}$$

- **Initial Information:**

$$(\boldsymbol{\theta}_0 \mid D_0) \sim [\mathbf{m}_0, \mathbf{C}_0]. \tag{4.23}$$

4.3.1 Inference when the shape parameter is known

In this subsection we consider the parameter κ to be static and known. Under those two assumptions, Equation (4.19) belongs to the EF and the inference cycle for the DBNBM may follow the DGLM scheme shown in the previous chapter. The prior distributions for the state

vector θ_t and the linear predictor λ_t can be obtained by direct application of the system equation as described in section 3.2. To equate the hyperparameters α_t and β_t notice that:

$$f_t = \mathsf{E}[\log(\mu_t) \mid \kappa, \alpha_t, \beta_t D_{t-1}] = \psi(\alpha_t) - \psi(\beta_t) - log(\kappa), \tag{4.24}$$

$$q_t = \mathsf{VAR}[\log(\mu_t) \mid \alpha_t, \beta_t, \kappa, D_{t-1}] = \psi'(\alpha_t) + \psi'(\beta_t). \tag{4.25}$$

The proof of the two expression above can be found in appendix B. Again, using first order approximations for the Digamma and Trigamma functions we can solve (4.24) and (4.25) for α_t and β_t to obtain:

$$\alpha_t \approx \frac{1 + \kappa \exp(f_t)}{q_t},$$
$$\beta_t \approx \frac{\exp(-f_t) + \kappa}{q_t \kappa}.$$

Since κ is considered to be known we can compute the predictive distribution marginally from the joint distribution $p(y_t, \mu_t \mid D_{t-1}) = p(y_t \mid \mu_t, \kappa, D_{t-1}) p(\mu_t \mid D_{t-1})$ in the following way:

$$\begin{aligned}
p(y_t \mid D_{t-1}, \kappa) &= \int_0^\infty \frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t + 1)\Gamma(\kappa^{-1})} \frac{(\kappa\mu_t)^{y_t}}{(1 + \kappa\mu_t)^{y_t + \kappa^{-1}}} \frac{\kappa^{\alpha_t}}{B(\alpha_t, \beta_t)} \frac{\mu_t^{\alpha_t - 1}}{(1 + \kappa\mu_t)^{\alpha_t + \beta_t}} d\mu_t \\
&= \frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t + 1)\Gamma(\kappa^{-1})} \frac{\kappa^{y_t}\kappa^{\alpha_t}}{B(\alpha_t, \beta_t)} \int_0^\infty \frac{\mu_t^{y_t + \alpha_t - 1}}{(1 + \kappa\mu_t)^{\alpha_t + \beta_t}} d\mu_t \\
&= \frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t + 1)\Gamma(\kappa^{-1})} \kappa^{y_t + \alpha_t} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)\kappa^{\alpha_t + \beta_t}} \\
&= \frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t + 1)\Gamma(\kappa^{-1})} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)}.
\end{aligned} \tag{4.26}$$

Now, observe that, even though, the closed-form of (4.26) is exact, it may not be convenient from a computational point of view. This is because in most software there is an upper limit for the Gamma function. In R, for instance, this limit is 171 so that any observation with value equal or larger than this threshold would cause a numerical error in the estimation algorithm. To overcome this problem we chose to work with the Beta function instead of the Gamma function. The Beta function is defined in the following manner: $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ for $a, b \in \mathsf{R}^+$. Thus, it is easy to see that:

$$\frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t + 1)\Gamma(\kappa^{-1})} = \frac{1}{y_t B(y_t, \kappa^{-1})}$$

Notice that this expression still presents some complications since it is not well defined when $y_t = 0$. The solution is to use the recurrence relation $B(y_t, \kappa^{-1}) = B(y_t + 1, \kappa^{-1}) \frac{y_t + \kappa^{-1}}{y_t}$ and plug this result into (4.26). That yields:

$$p(y_t \mid D_{t-1}, \kappa) = \frac{1}{B(\alpha_t, \beta_t)} \frac{1}{y_t + \kappa^{-1}} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(y_t + 1, \kappa^{-1})}. \tag{4.27}$$

A point prediction $\hat{y}_t = \mathsf{E}[y_t \mid D_{t-1}, \kappa]$ can be obtained by direct application of the law of total expectations. That is:

$$\mathsf{E}[y_t \mid D_{t-1}] = \mathsf{E}[\mathsf{E}[y_t \mid \mu_t] \mid D_{t-1}] = \frac{\alpha_t}{\kappa(\beta_t - 1)}.$$

The variance of the one-step- ahead forecast distribution can also be calculated from properties of conditional expectations through the expression below:

$$\begin{aligned} \mathsf{VAR}[y_t \mid D_{t-1}] &= \mathsf{E}[\mathsf{VAR}[y_t \mid \mu_t] \mid D_{t-1}] + \mathsf{VAR}[\mathsf{E}[y_{t-1} \mid D_{t-1}]] \\ &= \frac{\alpha_t}{\kappa(\beta_t - 1)} + \frac{\alpha_t}{(\beta_t - 1)} \end{aligned}$$

since $\mathsf{VAR}[\mathsf{E}[y_{t-1} \mid D_{t-1}]] = \mathsf{VAR}\left[\frac{\alpha_t}{\kappa(\beta_t - 1)}\right] = 0$.

Given a new observation y_t, we can update the dynamic parameter μ_t using the Bayes's Theorem, as usual. The distribution (4.20) is the conjugate prior for the mean of the Negative Binomial distribution with the parametrization used in this work. Then, it follows that:

$$\begin{aligned} p(\mu_t \mid \kappa, D_t) &\propto \frac{\mu_t^{y_t}}{(1 + \kappa\mu_t)^{y_t + \kappa^{-1}}} \frac{\mu_t^{\alpha_t - 1}}{(1 + \kappa\mu_t)^{\alpha_t + \beta_t}} \\ &\propto \frac{\mu_t^{y_t + \alpha_t - 1}}{(1 + \kappa\mu_t)^{\alpha_t + \beta_t} + y_t + \kappa^{-1}}. \end{aligned}$$

Thus, the posterior distribution for μ_t is a Beta of the second kind with parameters $\alpha^* = \alpha_t + y_t$ and $\beta^* = \beta_t + \kappa^{-1}$. Updating of the linear predictor λ_t, now, can be easily achieved through Expressions (4.24) and (4.25), which gives

$$f_t^* = \mathsf{E}[log(\mu_t) \mid D_t, \kappa] = \psi(\alpha_t + \beta_t) - \psi(\beta_t + \kappa^{-1}) - log(\kappa),$$

$$q_t^* = \mathsf{VAR}[log(\mu_t) \mid D_t, \kappa] = \psi'(\alpha_t + y_t) + \psi'(\beta_t + \kappa^{-1}).$$

Finally, updating of the partially specified state vector $\theta_\mathbf{t}$ can be done using the equations already developed in Section 3.2

$$\mathbf{m_t} = \mathbf{a_t} + \frac{1}{q_t}\mathbf{R_t}\mathbf{F_t}(f_t^* - f_t),$$

$$\mathbf{C_t} = \mathbf{R_t} - \frac{1}{q_t}\mathbf{R_t}\mathbf{F_t}\mathbf{F_t'}\mathbf{R_t}\left(1 - \frac{q_t^*}{q_t}\right).$$

In most real life situations, however, the static parameters κ and $\mathbf{W}$ will not be known. An elegant solution for specifying $\mathbf{W}$ is to use a discount factor, as already shown in Chapter 3.

The parameter κ can be set via Maximum a Posteriori Estimation (MAP) as follows. Let Θ be a vector of static parameters, and define the likelihood function as the product of the predictives, then the observed log-likelihood function is given by:

$$p(y_1, \cdots, y_T; \Theta) = \log\left(\prod_{t=1}^{T} p(y_t \mid \Theta, D_{t-1})\right) = \sum_{t=1}^{T} \log\left[p(y_t \mid \Theta, D_{t-1})\right],$$

where T represents the last observation. The MAP estimator can be written as

$$\hat{\Theta}_{MAP} = \arg\max_{\theta}\left(\sum_{t=1}^{T} p(y_t, \mid \Theta, D_{t-1})\right) p(\Theta). \tag{4.28}$$

See that, if $p(\Theta)$ is a non-informative uniform prior, (4.28) reduces to:

$$\hat{\Theta}_{MAP} = \arg\max_{\theta}\left(\sum_{t=1}^{T} p(y_t, \mid \Theta, D_{t-1})\right), \tag{4.29}$$

which is the usual Maximum Likelihood Estimator (MLE). The procedure outlined above may be seen as an Empirical Bayes method and has been mentioned in the literature by Lindsey & Lambert (1995)[7] and da Silva & Rodrigues (2015)[14] Since there is no exact closed-form solution to equation (4.29) for the DBNBM, the optimization must be done numerically. The R software [63] offers several options for this task. In this work we used the Nelder-Mead algorithm [64].

4.3.2 Inference when the shape parameter is unknown

The DBNBM when κ is unknown has its basic structure determined by Equations (4.19)-(4.23). Since κ is a strictly positive continuous quantity, we can assign a Gamma prior to the static parameter of the observation equation. Then, the model is fully specified with:

$$p(\kappa) = \frac{s^r}{\Gamma(r)}\kappa^{r-1}\exp(-s\kappa), \tag{4.30}$$

where $r, s > 0$. The Gamma distribution is very flexible and this is the main reason why it was chosen as a prior for κ. The specification of the hyperparameters r and s, however, is tricky since, in most times, we will not have much information about the shape parameter. A common approach is to assign very small values to both parameters in order to make the distribution as vague as possible. We opted, instead, to give more weight to smaller values of κ. The reasoning for this choice is that κ accounts for the dispersion of the model and very high values might not be realistic since it would imply an extremely overdispersed data set. Thus, from hereon, we are going to set $r = s = 1$ in all the following analysis.

The prior distributions for the state vector and the linear predictor are the same as when κ is known and results (4.24) and (4.25) also hold. Observe, however, that α_t and β_t are functions of the quantity κ that is now considered unknown. The predictive distribution can be obtained

marginally, as before, from the joint distribution $p(y_t, \mu_t, \kappa \mid D_{t-1}) = p(y_t \mid \mu_t, \kappa, D_{t-1})p(\mu \mid \kappa, D_{t-1})p(\kappa \mid D_{t-1})$ according to the following steps:

$$\begin{aligned}
p(y_t \mid D_{t-1}) &= \int_0^\infty \int_0^\infty p(y_t \mid \mu_t, \kappa, D_{t-1}) p(\mu_t \mid \kappa, D_{t-1}) p(\kappa) d\mu_t d\kappa \\
&= \int_0^\infty \int_0^\infty \frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t+1)\Gamma(\kappa^{-1})} \frac{s^r (\kappa\mu_t)^{y_t} \mu_t^{\alpha_t - 1} \kappa^{\alpha_t} \kappa^{r-1} \exp(-s\kappa)}{\Gamma(r) B(\alpha_t, \beta_t)(1+\kappa\mu_t)^{y_t + \kappa^{-1}} (1+\kappa\mu_t)^{\alpha_t + \beta_t}} d\mu_t d\kappa \\
&= \int_0^\infty \int_0^\infty \frac{s^r}{\Gamma(r) B(\alpha_t, \beta_t)} \left[\frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t+1)\Gamma(\kappa^{-1})} \frac{\kappa^{\alpha_t + y_t} \mu_t^{y_t + \alpha_t - 1}}{(1+\kappa\mu_t)^{y_t + \alpha_t + \beta_t + \kappa^{-1}}} d\mu_t \right] \\
&\times \kappa^{r-1} \exp(-s\kappa) d\kappa \\
&= \int_0^\infty \frac{1}{B(\alpha_t, \beta_t)} \frac{\Gamma\left(y_t + k^{-1}\right)}{\Gamma(y_t+1)\Gamma(\kappa^{-1})} B(\alpha_t + y_t, \beta_t + \kappa^{-1}) \frac{s^r}{\Gamma(r)} \kappa^{r-1} exp(-s\kappa) d\kappa \\
&= \int_0^\infty \frac{1}{B(\alpha_t, \beta_t)} \frac{1}{y_t + \kappa^{-1}} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(y_t + 1, \kappa^{-1})} \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa. \qquad (4.31)
\end{aligned}$$

Integral (4.31) does not have an analytic solution. Da Silva, Migon & Correia (2011) [13] use numerical integration to handle a similar problem when working on a Dynamic Bayesian Beta Model. We found, however, that this approach is very prone to numerical errors when we are dealing with count data. Since it is very easy to sample from a Gamma distribution with parameters r and s, we chose, instead, to approximate equation (4.31) using Monte Carlo integration, like in Section 4.2. MCI, in this situation, can work as an efficient tool at a lower computational cost in relation to MCMC methods. To obtain point predictions, again, we use iterated expectations following the sequence below:

$$\begin{aligned}
\mathsf{E}[y_t \mid D_{t-1}] &= \mathsf{E}[\mathsf{E}[\mathsf{E}[y_t \mid \mu_t, \kappa]] \mid D_{t-1}] \\
&= \mathsf{E}[\mathsf{E}[\mu_t \mid \kappa] \mid D_{t-1}] \\
&= \mathsf{E}\left[\frac{\alpha_t(\kappa)}{\kappa(\beta_t(\kappa) - 1)} \mid D_{t-1} \right]. \qquad (4.32)
\end{aligned}$$

Since the hyperparameters α_t and β_t are both functions of κ, there is no exact solution to Equation (4.32), although we can approximate the expression by means of the usual Monte Carlo methods.

The posterior distribution for μ_t can be obtained from:

$$p(\mu_t, \kappa \mid D_t) = \frac{p(y_t \mid \mu_t, \kappa, D_{t-1}) p(\mu_t \mid \kappa, D_{t-1}) p(\kappa)}{p(y_t \mid D_{t-1})},$$

then, integrating over κ we have:

$$\begin{aligned}
p(\mu_t \mid D_t) &= \frac{\int_0^\infty p(y_t \mid \mu_t, \kappa, D_{t-1}) p(\mu_t \mid \kappa, D_{t-1}) p(\kappa) d\kappa}{p(y_t \mid D_{t-1})} \\
&= \frac{1}{p(y_t \mid D_{t-1}) B(\alpha_t, \beta_t)} \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{\kappa^{\alpha_t + y_t} \mu_t^{\alpha_t + y_t - 1}}{(1+\kappa\mu_t)^{y_t + \kappa^{-1} + \alpha_t + \beta_t}} \frac{s^r \kappa^{r-1} \exp(-s\kappa)}{\Gamma(r)} d\kappa. \qquad (4.33)
\end{aligned}$$

Based on Equation (4.33) the n-th moment of the linear predictor can be found from

$$\mathsf{E}[log(\mu_t^n \mid D_t)] = \int_0^\infty log^n(\mu_t) p(\mu_t \mid D_t) d\mu_t. \tag{4.34}$$

Then, for $n = 1$ and $n = 2$ we have, respectively

$$\begin{aligned}\mathsf{E}[log(\mu_t) \mid D_t) = &\frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{1}{y_t + \kappa^{-1}} \frac{1}{B(y_t + 1, \kappa^{-1})} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)} \\ &\times \left[\psi(\alpha_t + y_t) - \psi(\beta_t + \kappa^{-1}) - log(\kappa)\right] \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa,\end{aligned} \tag{4.35}$$

$$\begin{aligned}\mathsf{E}[log^2(\mu_t) \mid D_t] = &\frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{1}{y_t + \kappa^{-1}} \frac{1}{B(y_t + 1, \kappa^{-1})} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)} \\ &\times \left[(\psi(\alpha_t + y_t) - \psi(\beta_t + \kappa^{-1}) - log(\kappa))^2 + \psi'(\alpha_t + y_t) + \psi'(\beta_t + \kappa^{-1})\right] \\ &\times \frac{s^r}{\Gamma(r)} \kappa^{r-1} exp(-s\kappa) d\kappa.\end{aligned} \tag{4.36}$$

Demonstration of Expressions (4.35) and (4.36) is left for Appendix A3. Notice that both equations can be decomposed into a function of κ times the kernel of a Gamma distribution. Thus, it is easy to approximate the integrals via Monte Carlo integration techniques. Therefore, using the results above, $f_t^* = \mathsf{E}(\lambda_t \mid D_t) = \mathsf{E}[log(\mu_t) \mid D_t)$ and $q_t^* = \mathsf{VAR}(\lambda_t \mid D_t) = \mathsf{E}[log^2(\mu_t) \mid D_t] - f_t^{*^2}$. As usual, we can employ Equations (3.13) and (3.14) to update the state vector via Linear Bayes Estimation. Finally, the posterior distribution for the static parameter κ considering the whole information set D_T is given by:

$$p(\kappa \mid D_T) = \frac{\left[\prod_{t=1}^T p(y_t \mid D_{t-1}, \kappa)\right] p(\kappa \mid D_{t-1})}{\int_0^\infty \left[\prod_{t=1}^T p(y_t \mid D_{t-1}, \kappa)\right] p(\kappa \mid D_{t-1}) d\kappa}, \tag{4.37}$$

where $p(y_t \mid D_{t-1}, \kappa) = \int_0^\infty p(y_t \mid \mu_t, D_{t-1}) p(\mu_t \mid D_{t-1}) d\mu_t = \frac{1}{B(\alpha_t, \beta_t)} \frac{1}{y_t + \kappa^{-1}} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(y_t + 1, \kappa^{-1})}$.

Under quadratic loss the optimal predictor for κ is the posterior expected value, that is, $\hat{\kappa} = \mathsf{E}[\kappa \mid D_T] = \int_0^\infty \kappa p(\kappa \mid D_T) d\kappa$. Direct integration is not possible, though, because there is no closed-form for (4.37). In this work, we chose to sample from the posterior distribution using the adaptive Rejection Metropolis Sampling (ARMS) technique proposed by Gilks and Best (1995) [26]. The ARMS algorithm is a powerful and computationally efficient method to sample from uniparametric distributions. The samples obtained can be used straightforwardly for the computation of point estimates and credible intervals for the parameter κ.

4.3.3 Simulation experiment

To evaluate both the MLE and Bayesian estimators for κ we carried out a Monte Carlo experiment. Four scenarios were designed with true parameter values $\kappa = 0.3$, $\kappa = 0.5$, $\kappa = 0.8$ and $\kappa = 1$. For each of these scenarios $L = 1000$ Monte Carlo with samples with of $n = 50$, $n = 100$, $n = 200$ and $n = 300$ were generated from a Negative Binomial Local Level Model (LLM) [2]

[2] A basic random walk plus noise model.

with static variance $W = 0.02$. Due to machine limitations imposed by R, we filtered samples containing values equal or larger than 200.

The ARMS algorithm was implemented using the function of the same name contained in the 'dlm' package from R [65]. The application of the algorithm requires a bounded convex set for the target density. This is not the case here since equation (4.37) has support on R^+. In these situations, the authors of the package suggest to restrict the support to a bounded set of probability close to one. In all scenarios studied in this work the true value of the parameter κ does not exceed one, so we chose to restrict the support for the target density in the interval $[0, 5]$. For the Bayesian inference chains of size 2000 were generated from which the first 1000 samples were discarded. Preliminary tests using diagnostic tools available on the 'coda' package [66] from R showed that a burn in of 1000 is enough to obtain stationary chains. We also could not detect any signs of auto-correlation in the chains, thus there was no need to use a lag. The prior distribution for κ was assumed to be a Gamma$(1, 1)$ for reasons already discussed. Two Bayesian Estimators were considered: the posterior median (BE-Median) and the posterior mean (BE-mean). We observed superior performance for the latter estimator and for this reason only Be-mean will be reported along with a 95% credible interval and an empirical coverage probability. To compute the Maximum Likelihood (or MAP) estimator the Nelder-Mead algorithm was used with initial value for the parameter of interest arbitrarily set at $\kappa_0 = 0.8$. No asymptotic theory is available for this estimator in the context of Bayesian Dynamic Linear Models, so we will report in this text only a point estimate.

The Bayesian and Maximum Likelihood estimators were evaluated according to their relative bias (RB) and Root of the Mean Squared Error (RMSE). A summary of the Monte Carlo study is presented on table (4.2). The BE-mean column reports the sample average over the 1000 estimates obtained for κ along with the mean RB defined, in percentage, as: $RB = \sum_{i=1}^{L} \frac{100(\hat{\kappa}_i - \kappa)}{|\kappa|}$, where $\hat{\kappa}_i$ denotes the estimate for κ obtained in the i-th Monte Carlo sample. The CI column indicates the 95% credibility interval (calculated as the mean lower and upper limits across the 1000 Monte Carlo samples generated) associated with the point estimate and also the corresponding coverage probability (CP). In the last column the RMSE, defined as $RMSE = \sqrt{\frac{\sum_{i=1}^{L}(\hat{\kappa}_i - \kappa)^2}{L}}$ is presented.

Notice that the results point to a better performance of the MLE over the BE since the RB and RMSE reported are significantly smaller for the first estimator in all scenarios. The Bayesian estimator seems to be heavily influenced by the prior distribution chosen for the parameter. As expected, larger biases are observed for small values of true κ since the probability mass of the prior is concentrated around one. It is noteworthy, however, that as the sample size increases, the biases get progressive smaller, suggesting good behaviour of the Bayesian estimator for large data sets even when the prior distribution for the static parameter is distant from the target. When we have true $\kappa = 0.8$ or $\kappa = 1.0$, the biases are smaller than 5% regardless of samples size indicating that, if the choice of the prior distribution is close enough to the real value of the parameter, the Bayesian Estimator is well behaved even for small samples. As expected, the

True	Size	MLE (Bias%)	MLE RMSE	BE Mean (Bias%)	CI (CP)	BE Mean RMSE
$\kappa = 0.3$	50	0.295 (-1.695)	0.178	0.373 (24.196)	[0.074, 1.112] (0.968)	0.288
	100	0.300 (-0.024)	0.123	0.346 (15.185)	[0.102, 0.890 (0.952)	0.245
	200	0.297 (-1.085)	0.083	0.326 (8.709)	[0.128, 0.732] (0.928)	0.206
	300	0.297 (-0.859)	0.068	0.310 (3.377)	[0.134, 0.657] (0.928)	0.191
$\kappa = 0.5$	50	0.496 (-0.791)	0.229	0.555 (10.912)	[0.16, 1.432] (0.957)	0.228
	100	0.500 (-0.048)	0.152	0.521 (4.178)	[0.218, 1.117] (0.949)	0.235
	200	0.497 (-0.638)	0.111	0.523 (4.530)	[0.258, 0.989] (0.946)	0.220
	300	0.501 (0.218)	0.088	0.519 (3.828)	[0.278, 0.945] (0.959)	0.212
$\kappa = 0.8$	50	0.770 (-3.717)	0.288	0.828 (3.527)	[0.311, 1.868] (0.96)	0.337
	100	0.798 (-0.212)	0.213	0.821 (2.652)	[0.393, 1.585] (0.958)	0.292
	200	0.798 (-0.208)	0.150	0.822 (2.698)	[0.458, 1.434] (0.941)	0.257
	300	0.801 (0.105)	0.109	0.833 (4.146)	[0.493, 1.386] (0.945)	0.245
$\kappa = 1.0$	50	0.973 (-2.708)	0.333	0.994 (-0.646)	[0.409, 2.112] (0.969)	0.379
	100	0.982 (-1.784)	0.237	1.017 (1.696)	[0.525, 1.870] (0.954)	0.337
	200	0.999 (-0.103)	0.157	1.022 (2.162)	[0.612, 1.666] (0.943)	0.281
	300	0.996 (-0.422)	0.130	1.013 (1.267)	[0.639, 1.598] (0.937)	0.266

Table 4.2: Summary of the Monte Carlo study

variance decreases for larger samples, as can be observed by the credible intervals. The coverage probabilities are all very close to the fixed nominal level of 95%, further confirming the good performance of the Bayesian estimator.

4.4 Real Data Application

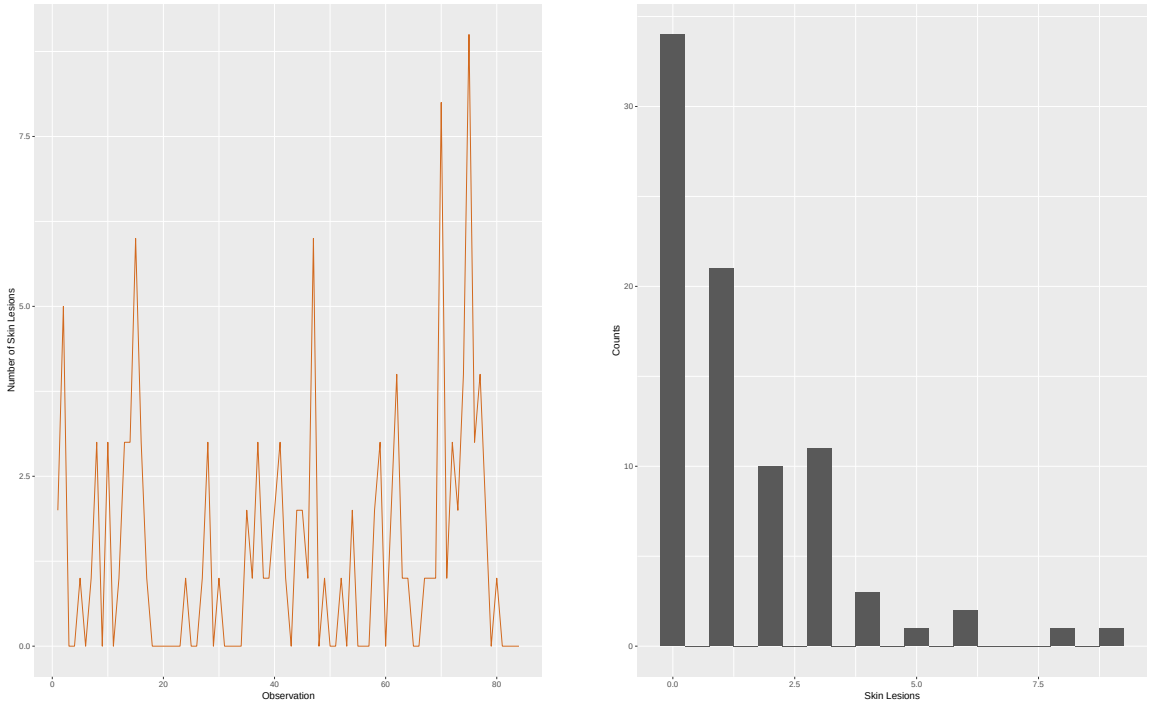
In this Section, we present two applications of the models developed in this work to real count time series. The data sets were chosen because of their idiosyncratic features: they both display overdispersion and inflation of zeros. The objective here is to show scenarios where the traditional Poisson DGLM may not be the most appropriate modeling choice. To compare between different models we are going to use the following metrics: the Bayes Factor and the Bayesian Information Criterion (BIC) to evaluate in sample performance and the one-step-ahead prediction accuracy to evaluate out of sample performance. In both applications, the MCI steps were carried out with samples of size 5000. Specifically for the DBNBM, the estimation of the static shape parameter was performed via ARMS as already shown in the previous section. A Gamma(1,1) prior was chosen for κ and single chains of sizes 5000 with a burn-in of 1.000 were generated from the posterior distribution (4.37). Point and interval estimates were calculated straightforwardly from these chains.

4.4.1 Skin Lesions Data

The Skin Lesions time series consists of the total number of skin lesions on bovines reported monthly to animal health laboratories in New Zealand from January 2003 until December 2009. This data set was first analyzed by Jazi et al (2011) [67] in the context of zero-inflated count data models. Figures (4.1a) and (4.1b), below, shows the behaviour of the time series and its corresponding barplot.

The plot shows no discernible trend or seasonality components, therefore, a LLM would fit the observations well. There is also no sign of non stationarity in the time series as the KPSS (Kwiatkowski–Phillips–Schmidt–Shin) test reports a p-value greater than 0.1 The sample mean and variance of the data are, respectively, 1.4286 and 3.3563. Since the variance is more than twice as big as the mean, there is evidence that the skin lesion time series is overdispersed.

We fitted the uniparametric models of section 4.2.1 and both versions of the DBNBM to the skin lesions data set. According to the dynamic models literature, values for the discount factor δ, typically, range from 0.8 to 1 (West $ Harrison (1997) [3]) since, in structural models, the information loss is expected to increase smoothly from one state to the next. Notice that a very small value for δ would cause the stochastic term to dominate the system equation meaning that there is almost no flow of information between states. For this reason we used $\delta = \{0.80, 0.85, 0.90, 0.95, 0.99\}$ in all the following analysis. The values of m_0 and C_0 were set so that the first predicted values coincides with the real value of the time series. Table 4.3



(a) Skin lesions time series

(b) Barplot of the skin lesions time series

Figure 4.1: Skin Lesions data set.

presents a comparison of in-sample and out-of-sample performance for the uniparametric and Negative Binomial dynamic models. The second column indicates the values of δ, the third column shows the estimates for the shape parameter (only available for the Negative Binomial Models) and the others display the values of the goodness of fit criteria used for comparison.

Values in bold indicate the model with best performance according to each of the metrics evaluated. First our focus is on the uniparametric models. Looking at the table we notice that the Poisson DGLM is superior to its competitors in forecast accuracy when we consider the MSE. The good performance of the Poisson DGLM is not all unexpected, though. Remember that, as shown in chapter 3, conditionally on the mean of the process, we assume the observations to be drawn from a Poisson distribution but, marginally, they follow a Poisson-Gamma mixture – that is: the predictive distribution of the model is overdispersed. As for the in sample performance, the BIC values suggest that, in this exercise, the Bell Dynamic Model specified with high discount factors is superior to the other uniparametric models. The Bayes factor also shows strong evidence in favor of the Bell and Poisson Lindley models against the traditional Poisson DGLM for $\delta = 0.99$ and $\delta = 0.95$. Compared to the other models, the Yule Simon and Borel Dynamic Models have poor in sample performance according to the Bayesian Information Criterion, but similar out of sample performance when we look to the MAE statistic. Overall, the new uniparametric count models presented in this text and the Poisson DGLM had similar performances for the skin lesions data set.

Table 4.3 also summarizes the results for the DBNBM. We provide both the MLE (or MAP) and Bayesian Estimator for κ. The BE reported is the posterior mean. For this example a chain

Model	δ	Estimated κ	MSE	MAE	BIC	BF
Poisson	0.99	-	3.763	1.333	343.580	1
	0.95	-	3.750	1.345	340.591	1
	0.90	-	3.715	1.346	337.693	1
	0.85	-	3.697	1.341	342.266	1
	0.80	-	3.822	1.359	412.868	1
Borel	0.99	-	4.333	1.379	351.533	0.0187
	0.95	-	4.306	1.378	354.963	8.00e-04
	0.90	-	4.287	1.372	368.220	2.35e-07
	0.85	-	4.306	1.364	413.794	2.94e-16
	0.80	-	4.479	1.373	620.345	8.85e-46
Bell	0.99	-	3.977	1.349	326.679	4.68e+03
	0.95	-	3.969	1.354	329.002	3.29e+02
	0.90	-	3.916	1.345	333.868	6.771e+00
	0.85	-	3.919	1.340	359.881	1.00e-04
	0.80	-	3.989	1.332	462.387	1.77e-11
Poisson-Lindley	0.99	-	4.054	1.353	328.018	2.39e+03
	0.95	-	4.033	1.353	330.469	1.58e+02
	0.90	-	3.979	1.347	334.233	5.64e+00
	0.85	-	3.961	1.335	361.075	8.24e-05
	0.80	-	4.096	**1.331**	501.690	5.16e-20
Yule-Simon	0.99	-	4.048	1.332	334.845	78.825
	0.95	-	4.183	1.349	345.285	0.096
	0.90	-	4.264	1.356	357.057	6.24e-05
	0.85	-	4.300	1.354	366.244	6.21e-06
	0.80	-	4.286	1.341	373.307	3.896e+08
Negative Binomial MLE	0.99	0.9573	3.509	1.41	**289.7**	7.655e+11
	0.95	0.8933	3.506	1.407	**289.7**	1.034e+12
	0.90	0.7992	3.561	1.432	290.4	1.736e+11
	0.85	0.6942	3.63	1.465	291.0	1.225e+12
	0.80	0.5836	3.699	1.493	291.5	2.051e+27
Negative Binomial BE	0.99	0.9701 (0.455, 1.694)	**3.488**	**1.394**	290	7.655e+11
	0.95	0.9119 (0.3975, 1.6313)	3.502	1.407	290.3	4.015e+12
	0.90	0.8277 (0.3066, 1.5685)	3.544	1.43	290.9	1.302e+11
	0.85	0.7211 (0.2006, 1.4601)	3.633	1.474	291.5	9.467e+11
	0.80	0.5821 (0.08824, 1.3326)	3.777	1.53	292.4	1.291e+27

Table 4.3: Performance summary of the proposed models applied to the skin Lesions data.

of size 5000 with a burn in of 1000 was generated us from the posterior distribution of κ using the ARMS algorithm and the 95% credible interval was calculated straightforwardly from the 2.5% and 97.5% quantiles of the distribution. To calculate the MLE, the Nelder-Mead algorithm was used with initial value set at 0.8. For this estimator only the point estimate is provided since, as already discussed, there is no asymptotic theory for the MLE in the context of Bayesian Dynamic Models.

The results show that both estimators behave similarly for the skin lesions data set. In a GLM context it is a well known fact that the presence of a dispersion parameter adds flexibility to the Negative Binomial Regression Model in relation to its Poisson counterpart. The same effect can be observed here: the DBNBM has better in sample performance in relation to the Poisson DGLM according to the BIC and BF measures and better out of sample performance according to the MSE. The comparison against the other uniparametric dynamic models also favors the DBNBM according to all criteria evaluated except for the MAE.

Figure 4.2 displays the one-step-ahead prediction for the skin lesions data set using the DBNBM with the Maximum Likelihood Estimator. The shaded area represents the 95% approximate credible interval limits. To compute the CI we generated 2000 random values from the predictive distribution and took the 2,5% and 97,5% quantiles. The first 10 observations can be considered a sort of learning period for the model and were excluded from the plot. For this illustration we used $\delta = 0.95$.

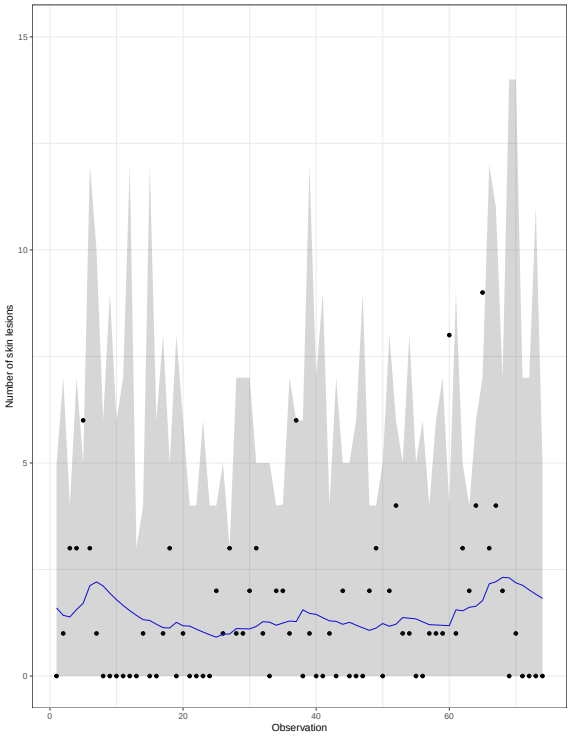
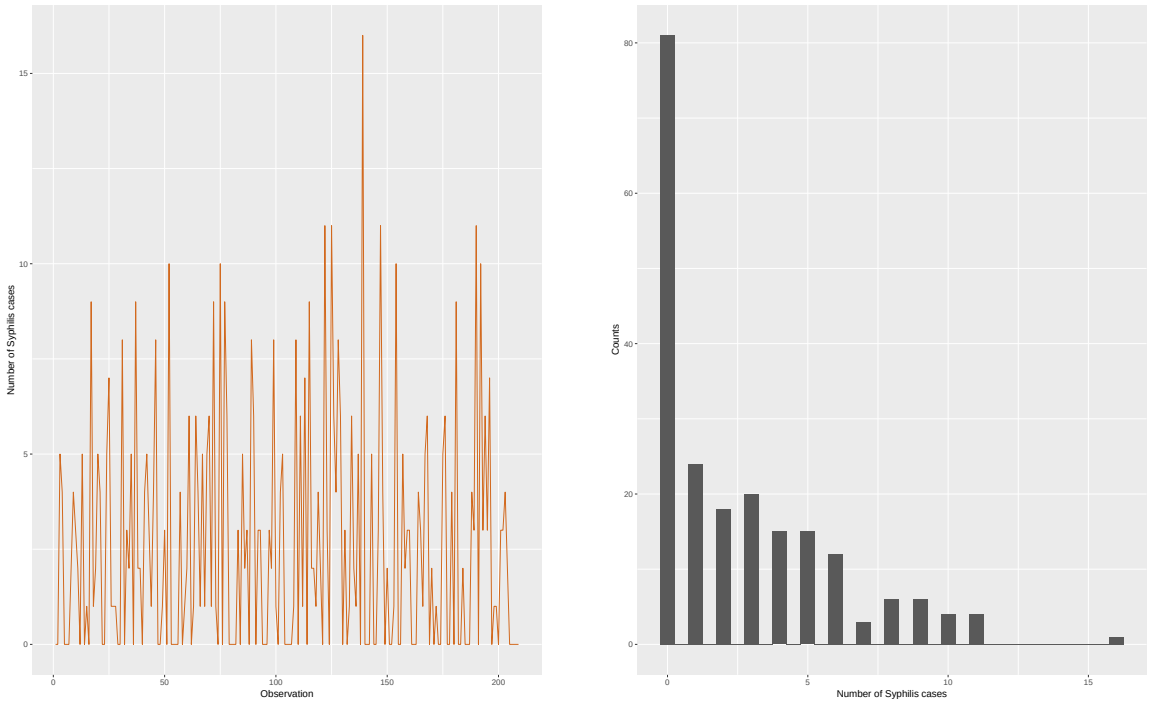


Figure 4.2: One step ahead prediction for the skin lesions data set. Solid line: predictions from the DBNBM using $\delta = 0.95$; Shaded Area: approximate 95% credible interval

4.4.2 Syphilis Data

The Syphilis Data data set consists of the weekly number of Syphilis cases reported in Porto Rico from 2007 until 2010, totalling 209 observations. This data is available from the ZIM package in R and its primary source is the Center of Disease Control of the United States (CDC). Figures (4.3a) and (4.3b), below, display the plot and histogram of the data, respectively.



(a) Syphilis time series

(b) Histogram of the Syphilis time series

Figure 4.3: Syphilis in Porto Rico time series.

Likewise the skin lesions data set, there is no visual indication of trend or seasonality. The sample mean and variance are 2.632 and 9.772, respectively. The KPSS test also shows evidence in favor of the stationarity null hypothesis (p-value>0.1). An inspection of the histogram reveals a very high frequency of zeros suggesting that zero-inflated models, such as the Borel or Yule-Simon, could be adequate to model the data. As before, we fitted the uniparametric dynamic models and the DBNBM to the Syphilis time series using discount factors ranging from 0.80 to 0.99. In all cases a LLM specification was used. To make the first predicted value coincide with the observed value we would have to set $m_0 = log(y_1)$, which is impossible because $y_1 = 0$. So, the initial conditions for m_t and C_t were set as $m_0 = C_0 = 1$. A summary of the results obtained for the uniparametric models can be seen on Table (4.4). Again, values in bold emphasize the model with the best performance in that particular metric.

Similarly to the skin lesions data set, the Poisson DGLM exhibits good performance in forecasting accuracy according to the MSE criterion. Considering the remaining criteria, however, the Poisson DGLM is outperformed by all other models fitted to the data. In particular, the Bayes Factor shows strong evidence in favor of the alternative uniparametric dynamic models

Model	δ	Estimated κ	**MSE**	**MAE**	**BIC**	**BF**
Poisson	0.99	-	10.31	2.467	1235.195	1
	0.95	-	10.42	2.499	1222.145	1
	0.90	-	10.69	2.532	1214.512	1
	0.85	-	11.00	2.567	1208.582	1
	0.80	-	11.34	2.607	1204.152	1
Borel	0.99	-	12.85	2.455	1024.745	4.995e+45
	0.95	-	12.80	2.457	1027.090	2.269e+42
	0.90	-	12.83	2.466	1036.586	4.328e+38
	0.85	-	12.86	2.476	1048.949	4.612e+34
	0.80	-	12.95	2.485	1071.177	7.504e+28
Bell	0.99	-	10.87	2.439	1016.298	3.411e+47
	0.95	-	10.89	2.460	1009.269	1.681e+46
	0.90	-	11.10	2.482	1011.627	1.137e+44
	0.85	-	11.36	2.511	1017.533	3.061e+41
	0.80	-	11.64	2.535	1027.079	2.824e+38
Poisson-Lindley	0.99	-	11.25	**2.436**	**988.090**	4.552e+53
	0.95	-	11.25	2.448	984.8453	3.38e+51
	0.90	-	11.38	2.470	987.3465	2.13e+49
	0.85	-	11.58	2.485	997.654	6.346e+45
	0.80	-	11.85	2.506	1013.0343	3.168e+41
Yule-Simon	0.99	-	12.02	2.449	1035.641	2.151e+43
	0.95	-	13.05	2.472	1075.841	5.883e+31
	0.90	-	13.35	2.490	1101.480	3.503e+24
	0.85	-	13.49	2.498	1121.530	8.004e+18
	0.80	-	13.62	2.506	1144.144	1.073e+13
Negative Binomial MLE	0.99	1.561 (1.179, 2.000)	**9.947**	2.539	**904.4**	9.751e+72
	0.95	1.524 (1.126, 1.980)	10.26	2.622	933.5	1.348e+69
	0.90	1.467 (1.107, 1.945)	10.97	2.764	916.5	7.437e+65
	0.85	1.401 (1.060, 1.842)	12.19	2.947	924.7	6.381e+62
	0.80	1.325 (0.967, 1.719)	14.74	3.266	933.5	9.002e+59
Negative Binomial Bayesian	0.99	1.58 (1.165, 2.082)	10.03	2.574	927.2	1.045e+68
	0.95	1.603 (1.175, 2.118)	10.40	2.653	930.3	3.478e+64
	0.90	1.63 (1.181, 2.165)	10.93	2.742	935.2	6.746e+61
	0.85	1.651 (1.185, 2.222)	11.66	2.855	940.3	2.511e+59
	0.80	1.661 (1.177, 2.256)	12.77	3.007	945.9	1.700e+57

Table 4.4: Performance summary of the proposed models applied to the Syphilis data.

and the DBNBM. By the MAE and BIC criteria, the Poisson-Lindley Model with $\delta = 0.99$ is the uniparametric model that best fits the Syphilis time series.

Looking specifically to the results concerning the Negative Binomial Models we observe that, unlike what was verified for the skin lesions data set, now the point estimates for the Bayesian and Maximum Likelihood estimators only coincide for very high values of discount factors. As δ decreases the in sample and out of sample performance of the DBNBM gets considerably worse regardless of the estimator used for κ. If $\delta = 0.99$, however, the DBNBM, in both specifications, still outperforms the uniparametric models both in sample and out of sample. Figure 4.4 shows the one step ahead prediction along with 95% confidence intervals for the DBNBM-MLE fitted to the syphilis time series with $\delta = 0.99$.

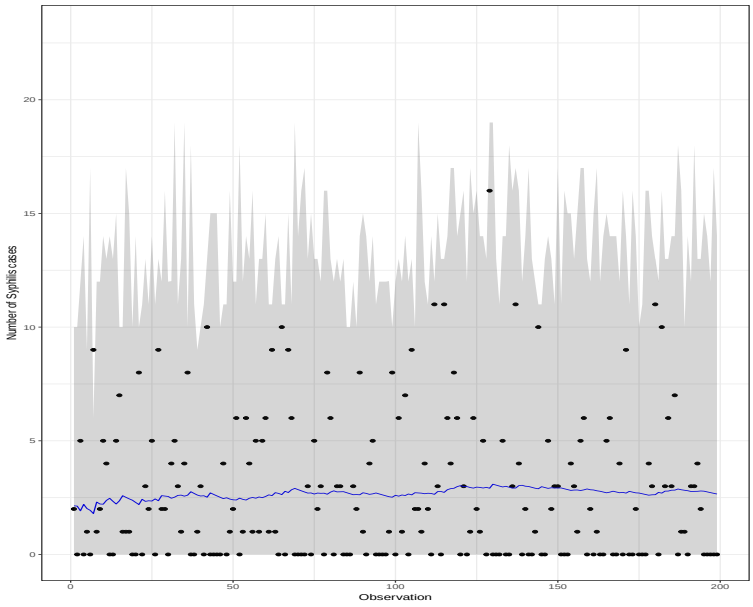


Figure 4.4: One step ahead prediction for the syphilis data set. Solid line: predictions from the DBNBM using $\delta = 0.99$; Shaded Area: approximate 95% credible interval.

4.5 Final Remarks

Time series of count data can often exhibit characteristics such as overdispersion or inflation of zeros. In these situations, dynamic models based on the Poisson distribution may not be appropriate because they require the mean and the variance of the data to be equal. In this chapter we presented several alternatives to the Poisson DGLM for dealing with single time series of counts based on distributions that can more adequately capture these aspects in a data set. The main contributions of this work are the development of: 1) a general method of inference for Bayesian Dynamic Models that is suited for uniparametric distributions; 2) the introduction of a framework for a Dynamic Bayesian Negative Binomial Model. The inference procedures of the models presented here follows a Bayesian structure that resembles the DGLM framework, but is not restricted to distributions within the EF. To keep the sequential nature of the Bayesian

inference, some approximations, such as Linear Bayes Estimation, first order Taylor expansions and Monte Carlo Integration, were used whenever the exact results were not available.

An important aspect concerning the DBNBM is the estimation of the static shape parameter κ. In the context of DGLM, it is well known that the estimation of the shape parameter when the observation equation is biparametric usually presents complications. Here, we propose two methods for estimating κ: first we consider the parameter to be known in the structure of the model and estimate it via Maximum Likekihood; in the other approach κ is considered an unknown quantity following a Gamma prior distribution – the estimation is carried out via adaptive Rejection Metropolis Sampling. An extensive simulation study was conducted and showed a good performance for both estimators. The models proposed were also applied to two real data sets and compared to the Poisson DGLM in both in-sample and out-of-sample performance. Results showed superiority of the DBNBM over the Poisson Model in both cases. The uniparametric models, on the other hand, displayed good in-sample performance but were outperformed by the Poisson DGLM in out-of-sample performance according to the Mean Squared Error of prediction.

Overall, the class of Bayesian Dynamic Models is very flexible since its structure allow for trend and seasonality components and intervention analysis. Covariates can also be included in the model by means of the F_t design matrix in the observation equation. The methods presented here are very general and can be extended, with a few modifications, to positive asymmetric data. A procedure similar to the one described for the DBNBM could be easily applied to the Gamma and Weibull distributions for example. We work out this cases in Appendix B. Another aspect of these dynamic models that could be further explored is the maximum likelihood estimation. The simulations showed good performance of the MLE, but an asymptotic theory assuring the existence and well behaviour of the estimator is still an open problem in the literature according to the best of author's knowledge.

Chapter 5

Dynamic Generalized Linear Models via Product Partition Models

5.1 Introduction

In the modeling of time series, it is very common the occurrence of change points or, in the context of econometrics, the so-called structural breaks. This type of phenomenon is the result of some external intervention that changes the structure of the analyzed process. These interventions may manifest themselves as the lack of stability either in the mean or in the variance of the parameters driving the time series. Also, in many cases, the change-points appear multiple times making it very important to correctly model the underlying process generating these breaks. For these reasons, the analysis of change-points plays a key role in fields such as economics, finance, engineering, climatology, hydrology among others. Failing to model structural breaks correctly can lead to large forecasting errors.

The intervention analysis literature is vast and covers both Bayesian and classical approaches. Within the classical framework, most of the single change-point detection models are variations of the well known Cumulative Sum (CUSUM) process. These models are, usually, formulated in terms of hypothesis tests in which the null hypothesis is the stability of the parameters. An extensive overview of this methods can be found in Csorgo & Horváth (1997) [68], Perron (2006) [69] and Aue (2013) [70]. Tests that allow for the detection of multiple unknown change-points are, in most cases, least squares type tests. Important works on this setting include Bai & Perron (1998) [71], Bai (1999) [72], Qu & Perron (2007), Kurozumi & Tuvaandorj [73] and Preuss, Puchstein and Dette (2015) [74]. For an application of these methods to finance problems, we refer the reader to Andreou & Ghysels (2002).

In this chapter we are concerned with identifying and modeling multiple change-points in a Bayesian framework. The Bayesian modeling of breaks in time series generally considers that the underlying process is governed by a latent discrete state vector following a Markovian evolution. This state variable is, usually, taken as piecewise constant and undergoes changes at the points of

regime switch. In this kind of setting, the posterior means for the state vector can be estimated via computational methods such as MCMC. Relevant papers following this kind of approach include Chib (1998) [75], Lai (2005) [76], Lai & Xing (2011) [77] and Martínez & Mena (2014) [78]. Of particular interest to the work we present here is the Product Partition Model (PPM) proposed by Hartigan (1990) [20] and expanded by Barry & Hartigan (1992, 1993) [21] [79]. PPM induces a block structure that divides the data set into blocks of observations governed by different regimes. Within each of these blocks, it is assumed that the observations are conditionally independent. Posterior probabilities for the number of change-points can be obtained exactly but, usually, at a high computational cost. For these reason, in most situations, the inference for the PPM class is done via Gibbs Sampling. Many works related to the Barry & Hartigan proposal can be found in the literature. Loschi & Cruz (2002) [80], for example, study the influence that different prior specifications for the cohesion have in the PPM product estimates; Fearnhead (2006) [81] and Loschi & Cruz (2005) [82] provide a method for sampling direct from the posterior distribution of the number of change-points; Loschi, Pontel & Cruz (2010) [83] extend the PPM to detect multiple change-points in linear regressions.

In this work we address the change-point problem by incorporating the PPM into the DGLM class of West, Harrison & Migon (1985) [4]. This approach was first presented by Almeida (2016) [22]. The inference of this class of models that, from here on, we shall call DGLM-PPM, follows the sequential nature of Bayesian statistics with evolution and updating steps. The conditional independence assumption of the PPM is preserved in DGLM-PPM and, between blocks, the parameters are dependent. The partitions are estimated from the data available using the Gibbs Sampler algorithm proposed by Barry & Hartigan (1993) [79]. In this work we propose to append an ARMS step within the Gibbs Sampling scheme to estimate the optimal discount factor associated with the block structure. Under this framework, posterior estimates for the change-point probabilities are readily available. This new formulation permits not only retrospective analysis, as in most structural breaks models, but also online inference. We also provide a simulation study and two examples with real life data.

This chapter is organized as follows: in Section 2 we present the PPM as described by Barry & Hartigan (1992) [21] along with the Gibbs sampling scheme used in this work to obtain inference about the partition. This formalism is extended in Section 3 where we introduce the DGLM-PPM class. To have a better understanding of the DGLM-PPM we conduct a simulation experiment that is presented in Section 4. Finally, in Section 5 we apply the conventional DGLM and the DGLM-PPM to two well known real-life time series and compare their performances.

5.2 Product Partition Models

Product Partition Models, as proposed by Hartigan (1990) [20], associate to each random partition induced in the data by the change-points a prior product distribution. It is also assumed that, given the partition, observations in different components are independent of each other.

Barry & Hartigan (1992) define partitions as blocks contiguous to each other. They further assume that, for each block, the set of observations obeys a different probability model. More formally, the authors define a block as follows: let $Y = \{Y_1, \cdots, Y_n\}$ be a sequence of consecutive observations, $I = \{0, 1, 2, \cdots, n\}$ a set of indexes, $\rho = \{i_0, i_1, \cdots, i_b\}$ a random partition of the set I and B a random variable that denotes the number of blocks (or components) in a given partition. It must also be true that $0 = i_0 < i_1 < \cdots < i_b = n$. In the case where $B = b$, the partition can be written as

$$[Y_1, \cdots, Y_{i_1}],[Y_{i_1+1}, \cdots, Y_{i_2}], \cdots,[Y_{i_{b-1}+1}, \cdots, Y_{i_b}].$$

In the expression above each block is denoted by $Y_{[i_{j-1}i_j]} = (Y_{i_{j-1}+1} \cdots, Y_{i_j})^\intercal$ with $j = 1, 2, \cdots, b$. Identify the j-th block as the set of observations $i + 1, \cdots, j$, where $i, j \in \rho$ and $i < j$. Barry & Hartigan (1992) define $c_\rho^{(j)}$ as the prior cohesion associated with the block j. Cohesion, in this context, must be understood as the degree of similarity between observations belonging to the same block. Details on how to calculate it will be presented later. The endpoints of each block $(i_0, i_1, \cdots, i_b)$ in the partition can also be seem as part of a Markov Chain that satisfies the following conditions:

$$0 \leq i_r < i_{r+1} \leq n \qquad \text{for} \qquad i_r < n,$$

$$i_r = i_{r+1} = n \qquad \text{for} \qquad i_r = n.$$

One can think, then, on the cohesion as a transition probability of the Markov Chain defined as above. According to Loschi & Cruz (2002) [80] the random set $(Y_1, \cdots, Y_n; \rho)$ follows a PPM if the two conditions below are verified:

1. The prior distribution that describes the probability that a partition ρ have endpoints $\{i_0, i_1, \cdots, i_b\}$ has a product form given by:

$$P(\rho = \{i_0, i_1, \cdots, i_b\}) = \frac{1}{K} \prod_{j=1}^{b} c_\rho^{(j)}, \tag{5.1}$$

   where $K = \sum_{\mathcal{C}} \prod_{j=1}^{b} c_\rho^{(j)}$ is a normalizing factor with $\mathcal{C}$ representing all possible partitions of the set I into b contiguous blocks with endpoints satisfying $0 = i_0 < i_1 < \cdots < i_b = n$, $\forall\, b \in I$.

2. Conditionally on $\rho = \{i_0, i_1, \cdots, i_b\}$ the observations $Y_1, \cdots, Y_n$ have the following joint distribution:

$$p(Y_1, \cdots, Y_n \mid \rho = \{i_0, i_1, \cdots, i_b\}) = \prod_{j=1}^{b} p_j(Y_\rho^{(j)}), \tag{5.2}$$

   where $p_j(Y_\rho^{(j)})$ is the density of the random vector $Y_\rho^{(j)}$.

Under those two assumptions, according to Barry & Hartigan (1992), the posterior distribution for the partition ρ will follow the same product form of the Equation (5.1) with the posterior cohesion for the j-th block defined as

$$c_{\rho}^{*(j)} = c_{\rho}^{(j)} p_j(Y_{\rho}^{(j)}). \tag{5.3}$$

Observe that the whole construction presented here, so far, does not assume any parametric form for the PPM. In the parametric approach one considers that each observation Y_k, $k \in \{1, 2, \cdots, n\}$ can be described by a marginal density conditioned on an unknown parameter θ_k that we will denote by $p_k(Y_k \mid \theta_k)$. Given $\theta_1, \cdots, \theta_n$, the $Y_1, \cdots, Y_n$ are conditionally independent with joint density $\prod_{j=1}^{b} p_j(Y_{\rho}^{(j)} \mid \theta_{\rho}^{(j)})$. It is also assumed that, within each block j, the corresponding observations are identically distributed. That is, given the partition $\rho = \{i_0, i_1, \cdots, i_b\}$, $b \in I$ and a block $[i_{r-1} i_r]$ we have $\theta_k = \theta_{i_r} = \theta_{[i_{r-1} i_r]}$, $\forall\ k$ such that $i_{r-1} + 1 < k < i_r$ and $r \in I$; that is, within, say, the j-th block, it must hold that: $\theta_{\rho}^{(j)} = \theta_{i+1} = \theta_{i+2} = \cdots = \theta_{j-1} = \theta_j$. To every $\theta_{\rho}^{(j)}$ we assign a block prior density $p_j(\theta_{\rho}^{(j)})$ with respect to some measure $\Theta_{\rho}^{(j)}$. Thus, the block predictive function (or data factor) $p_j(Y_{\rho}^{(j)})$ can be calculated as

$$p_j(Y_{\rho}^{(j)}) = \int_{\Theta_{\rho}^{(j)}} p_j(Y_{\rho}^{(j)} \mid \theta_{\rho}^{(j)}) p_j(\theta_{\rho}^{(j)}) d\theta_{\rho}^{(j)}, \tag{5.4}$$

and the block posterior density is given by,

$$p_j(\theta_{\rho}^{(j)} \mid Y_{\rho}^{(j)}) = \frac{p_j(\theta_{\rho}^{(j)}) \prod_{i_{j-1}+1}^{i_j} p_j(Y_k \mid \theta_{\rho}^{(j)})}{p_j(Y_{\rho}^{(j)})} \text{.for} \qquad j = 1, \cdots, b.$$

An important aspect concerning the PPM class is that the parameters $(\theta_{\rho}^{(1)} \cdots \theta_{\rho}^{(b)})$ with $b \in I$ are allowed to be time varying as long as the product form of the joint distribution for the observations and partitions is preserved. Thus the PPM can be used to model change-points in time series.

To obtain inference about the parameters, Barry & Hartigan (1992) use the posterior relevance defined as

$$r_{\rho}^{*(j)} = P([i_{j-1} i_j] \in \rho \mid Y_1, \cdots, Y_n). \tag{5.5}$$

This quantity denotes the probability that the j-th block belongs to the partition ρ given the data Y. The posterior distribution of θ_k conditional on the observations can be written as

$$p(\theta_k \mid Y_1, \cdots, Y_n) = \sum_{i_{j-1} < k \leq i_j} r_{\rho}^{*(j)} p_j(\theta_k \mid Y_{\rho}^{(j)}), \tag{5.6}$$

The posterior expected value of θ_k, or product estimates, can be computed from

$$\mathsf{E}(\theta_k \mid Y_1, \cdots, Y_n) = \sum_{i_{j-1} < k \leq i_j} r_\rho^{*(j)} E(\theta_k \mid Y_\rho^{*(j)}), \tag{5.7}$$

where $r_\rho^{*(j)} = \frac{\lambda_{i_0 i_{j-1}} c_\rho^{*(j)} \lambda_{i_0 i_{j-1}}}{\lambda_{i_j i_b}}$ with $\lambda_{i_{j-1} i_j} = \sum \prod_{k=1}^{b} c_\rho^{*(j)}$, the summation being over all sets $i = i_0 < i_1 < \cdots < i_{b-1} < i_b = j$.

Thus, the PPM provides a framework in which inference about clustered parameters can be obtained for each block j using standard Bayesian tools and the observations within the blocks. The posterior distributions for every θ_k, $k = 1, \cdots, n$ are computed, according to Equation (5.7), as weighted averages over the densities associated with the blocks containing θ_k where the weights are taken as the posterior relevances $r_\rho^{*(j)}$ defined in Equation (5.5).

An exact calculation of the posterior relevance in Equation (5.7) is feasible but, since the possible number of partitions grows exponentially as the number of observations increases, it can be too expensive from a computational point of view. For this reason, the product estimates are, frequently, obtained using MCMC techniques. In this work we used the Gibbs Sampling setting similar to the one described in Loschi & Cruz (2002) [80].

5.3 DGLM-PPM

5.3.1 The Model

The DGLM class, presented in Chapter 3, can be extended to allow random partitions on the observations using the PPM of Barry & Hartigan. As previously shown, a PPM is sufficiently defined by the two following conditions:

1. Given $\theta_1, \cdots, \theta_n$, the observations $(Y_1, \cdots, Y_n)$ must be conditionally independent.

2. The joint distribution of the observations and partitions must have a product form.

Almeida (2016) [22] introduces a new class of dynamic models that incorporates the PPM into the DGLM of West, Harrison & Migon (1985). In this new class, that we shall call DGLM-PPM, the two conditions that define a PPM are preserved and the parameters $\theta_\rho^{(1)}, \cdots, \theta_\rho^{(j)}, \cdots, \theta_\rho^{b}$ associated to each block j in the partition ρ are now allowed to be dependent. This new formulation is very rich since it retains the flexibility of the DGLM while allowing for the detection of multiple change-points using the block structure of the PPM. For a different approach that allows for across cluster correlation we refer the reader to Ferreira, Loschi & Costa (2014) [84].

The construction of the DGLM-PPM class will be done analogously to the traditional DGLM as described in Chapter 3, but in a block structure. Let $Y = (y_1, \cdots, y_n)$ be a time series with observations drawn from a distribution in the uniparametric EF. Using the definitions of

the previous section, $I = \{0, 1, 2, \cdots, n\}$ is a set of indexes; $\rho = \{i_0, i_1, \cdots, i_b\}$ is a random partition of the set I and B is a random variable that denotes the number of components in the partition. If $B = b$, the partition is comprised by b blocks, each of them represented by: $Y_\rho^{(j)} = (y_{i_{j-1}+1} \cdots, y_{i_j})^\intercal \equiv (y_{j1}, \cdots, y_{j_{n_j}})^\intercal$, $j = 1, 2, \cdots, b$, where n_j represents the number of elements of the j-th block and y_{jk} is the k-th element in the block. For the purposes of this work, the uniparametric EF is defined as

$$p(y_{jk} \mid \eta_{jk}, \tau_t) = \exp[\tau_{jk}(y_{jk}\eta_t - a(\eta_{jk}))]c(y_{jk}, \tau_{jk}). \tag{5.8}$$

Equation (5.8) represents the observation equation in a DGLM-PPM. Under the condition that $0 = i_0 < i_1 < \cdots < i_b = n$, the probability that a partition ρ has blocks with endpoints in $i_0, i_1, \cdots, i_b$ is given by equation (5.1). The prior distribution for the number $B = b$ of blocks comprising the partition can be obtained from:

$$p(B = b) \propto \sum_{\mathcal{C}_1} \prod_{j=1}^{b} c_\rho^{(j)}, \tag{5.9}$$

where $\mathcal{C}_1$ also denotes a set of all possible partitions of I into exactly b contiguous blocks.

Suppose now, that there is a known vector $\mathbf{F}_{jk}$ of explanatory variables where $j = 1, 2, \cdots, b$, $k = 1, 2, \cdots, n_j$ with n_j denoting the last observation of the j-th block. We can relate the observation y_{jk} to the vector $\mathbf{F_{jk}}$ through a non linear and monotonic link function $g(.)$ that maps the natural parameter η_{jk} of the observation distribution to the real line. That is:

$$\lambda_{jk} = g(\eta_{jk}) = \mathbf{F}_{jk}^{'}\boldsymbol{\theta}_\rho^{(j)}. \tag{5.10}$$

From properties of the EF, the conjugate prior distribution for the natural parameter η_{jk} of the observation equation will be given by:

$$p(\eta_{jk}) = b(\alpha_{jk}, \beta_{jk}) \exp[\alpha_{jk}\eta_{jk} - \beta_{jk}a(\eta_{jk})], \tag{5.11}$$

where α_{jk} e β_{jk} are hyperparameters that can be estimated from the moments of the linear predictor. The normalizing constant of (5.11) is given by the integral,

$$b(\alpha_{jk}, \beta_{jk}) = \frac{1}{\int \exp[\alpha_{jk}\eta_{jk} - \beta_{jk}a(\eta_{jk})]d\eta_{jk}}. \tag{5.12}$$

To compute the predictive distribution of the observations Y given the partition observe that, conditionally on $\rho = \{i_0, i_1, \cdots, i_b\}$, the joint density of the observations can be written as a product of block predictives, that is:

$$p(Y \mid \rho) = \prod_{j=1}^{b} p(Y_\rho^{(j)}), \tag{5.13}$$

where, by the properties of the EF, the predictive distribution $p(Y_\rho^{(j)})$ associated to the observations in the j-th block can be calculated from:

$$\begin{aligned} p(Y_\rho^{(j)}) &= \prod_{k=1}^{nj} \int p(y_{jk} \mid \eta_{jk}) p(\eta_{jk}) d\eta_{jk} \\ &= \prod_{k=1}^{nj} \frac{b(\alpha_{jk}, \beta_{jk}) c(y_{jk}, \frac{1}{\phi_{jk}})}{b(\alpha_{jk} + \phi_{jk} Y_{jk}, \beta_{jk} + \phi_{jk})}. \end{aligned} \tag{5.14}$$

According to Barry & Hartigan (1992), the posterior distribution of the partition $\rho = \{i_0, i_1, \cdots, i_n\}$ can be obtained from the expression:

$$P(\rho = \{i_0, i_1, \cdots, i_b\}) = \frac{\prod_{j=1}^{b} c_\rho^{*(j)}}{\sum_{\mathcal{C}} \prod_{j=1}^{b} c_\rho^{*(j)}}, \tag{5.15}$$

where $c_\rho^{*(j)} = c_\rho^{(j)} p_j(Y_\rho^j)$ represents the posterior cohesion of the block $[i_{j-1} i_j]$.

The posterior distribution for the random variable B can also be computed as the product of posterior cohesions over all possible partitions, that is:

$$p(B = b) \propto \sum_{\mathcal{C}_1} \prod_{j=1}^{b} c_\rho^{*(j)}. \tag{5.16}$$

Barry & Hartigan (1992) consider that the joint density of parameters, observations and partitions define a parametric PPM if the following two criteria are met: i) the probability of a partition ρ can be written as a product of non-negative cohesions; ii) given the θ_k, the observations are conditionally independent. The conditional independence is an assumption of the DGLM class and the product form for the distribution of ρ is given by (5.15). Thus, DGLM-PPM satisfies both requirements.

The evolution and updating steps of the DGLM-PPM class are very reminiscent of this same process as described by West, Harrison & Migon (1985) for the DGLM class, so we will follow the same structure used in Chapter 3 here. The state parameters of each block evolve according to the following relation:

$$\boldsymbol{\theta}_\rho^{(j)} = \mathbf{G}_j \boldsymbol{\theta}_\rho^{(j-1)} + \boldsymbol{\omega}_t, \qquad \omega \sim [0, \mathbf{W}_j], \tag{5.17}$$

where $\mathbf{G}_j$ represents an evolution matrix, and $\boldsymbol{\omega}_j$ is a random evolution error following a non-specified distribution of zero mean and covariance matrix denoted by $\boldsymbol{W}_j$. Denote by $D_\rho^{(j)}$ the information set up to the j-th block. The inference of the DGLM-PPM requires the following specification for the posterior distribution of the state vector $\boldsymbol{\theta}_\rho^{(j-1)}$:

$$(\boldsymbol{\theta}_\rho^{(j-1)} \mid D_\rho^{(j-1)}) \sim [\mathbf{m}_{j-1}, \mathbf{C}_{j-1}]. \tag{5.18}$$

Observe that the inference of the DGLM-PPM follows the same structure of the conventional DGLM including the partial specification of the distributions only in terms of their first two moments. The block state prior $(\boldsymbol{\theta}_\rho^{(j)} \mid D_\rho^{(j-1)})$ can be obtained directly from the evolution equation (5.17) and is given by:

$$(\boldsymbol{\theta}_\rho^{(j)} \mid D_\rho^{(j-1)}) \sim [\mathbf{a}_\rho^{(j)}, \mathbf{R}_\rho^{(j)}], \tag{5.19}$$

where $\mathbf{a}_\rho^{(j)} = \mathbf{G}_j \mathbf{m}_\rho^{(j)}$ and $\mathbf{R}_\rho^{(j)} = \mathbf{G_j} \mathbf{C}_\rho^{(j-1)} \mathbf{G_j}' + \mathbf{W_j}$.

Let $D_\rho^{(j-1,k-1)}$ be the information set available up to the block $j-1$ and the observation $k-1$ of the j-th block that is being processed. The linear predictor λ_{jk} and the state vector $\boldsymbol{\theta}_\rho^{(j)}$ are assumed to have the following joint distribution:

$$\begin{bmatrix} \lambda_{jk} \\ \boldsymbol{\theta}_\rho^{(j)} \end{bmatrix} \Bigg| D_{(j-1,k-1)} \sim \left[\begin{bmatrix} f_{jk} \\ \mathbf{a_{jk}} \end{bmatrix}, \begin{bmatrix} q_{jk} & \mathbf{F}'_{jk}\mathbf{R}_{jk} \\ \mathbf{R}_{jk}\mathbf{F}'_j & \mathbf{R}_{jk} \end{bmatrix} \right] \tag{5.20}$$

with $f_{jk} = \mathbf{F}'_j \mathbf{R}_{jk}$ and $q_{jk} = \mathbf{F}'_k \mathbf{R}_{jk} \mathbf{F}_j$. The one-step-ahead distribution can be derived from the properties of the EF and were already shown in Chapter 3. Then, using equation (3.6), we get:

$$p(Y_{jk} \mid D_\rho^{(j-1,k-1)}) = \frac{b(\alpha_{jk}, \beta_{jk}) c(y_{jk}, \frac{1}{\phi_{jk}})}{b(\alpha_{jk} + \phi_{jk} Y_{jk}, \beta_{jk} + \phi_{jk})} \tag{5.21}$$

From Equation (3.7) the posterior distribution for the natural parameter η_{jk} can be computed as

$$p(\eta_{jk} \mid D_\rho^{(j-1,k-1)}) = b(\alpha_{jk} + \phi_{jk} Y_{jk}, \beta_{jk} + \phi_{jk}) exp[(\alpha_{jk} + \phi_{jk} y_{jk}) \eta_{jk} - (\beta_{jk} + \phi_{jk}) a(\eta_{jk})] \tag{5.22}$$

Updating of the linear predictor can proceed in the same way as in the conventional DGLM through the equations below:

$$\mathsf{E}[g(\eta_{jk}) \mid D_\rho^{(j-1,k-1)}] = f_{jk}^*, \qquad \text{and} \qquad \mathsf{VAR}[g(\eta_{jk}) \mid D_\rho^{(j-1,k-1)}] = q_{jk}^*.$$

Finally, the block state vector can be updated in terms of its first two moments using the LBE procedure shown in Chapter 3. That is,

$$(\boldsymbol{\theta}_\rho^{(j)} \mid D_\rho^{(j-1,k)}) \sim [\mathbf{m}_{jk}, \mathbf{C}_{jk}], \tag{5.23}$$

where the vector $\mathbf{m_{jk}}$ and the matrix $\mathbf{C_{jk}}$ can be thought of as direct extensions of Equations (3.13) and (3.14):

$$\mathbf{m_{jk}} = \mathbf{a}_{jk} + \frac{1}{q_{jk}} \mathbf{R}_{jk} \mathbf{F}_{jk} (f_{jk}^* - f_{jk}), \tag{5.24}$$

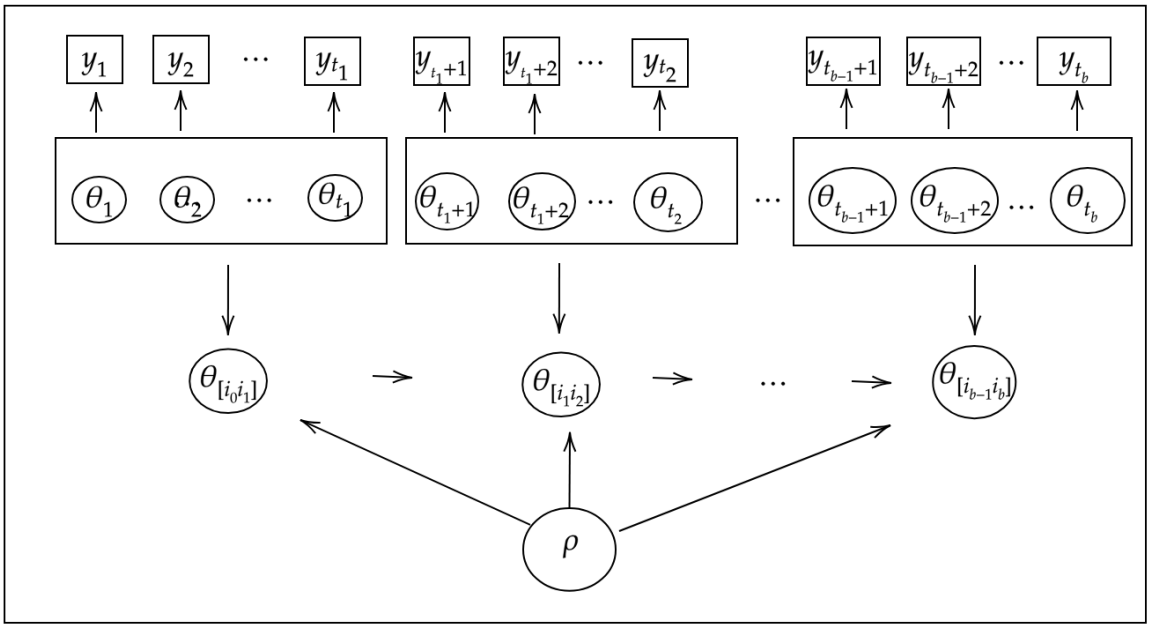


Figure 5.1: Schematic representation of the DGLM via PPM

$$\mathbf{C}_{jk} = \mathbf{R}_{jk} - \frac{1}{q_{jk}} \mathbf{R}_{jk}\mathbf{F}_{jk}\mathbf{F}'_{jk}\mathbf{R}_{jk} \left(1 - \frac{q^*_{jk}}{q_{jk}}\right). \tag{5.25}$$

Figure (5.1) presents a visual representation of the DGLM-PPM. Within each block there is no parametric evolution. The information set available at the beginning of the block $[i_{j-1}i_j]$ is denoted by $D_\rho^{(j-1,0)}$ and the updating procedure is started by taking $\mathbf{a}_{j1} = \mathbf{m}_{(j-1,n_{j-1})}$ and $\mathbf{R}_{j1} = \mathbf{C}_{(j-1,n_{j-1})}$. After each observation is processed, the values of $\mathbf{a}_{jk}$ and $\mathbf{R}_{jk}$ are updated to $\mathbf{a}_{jk} = \mathbf{m}_{(j,k-1)}$ and $\mathbf{R}_{jk} = \mathbf{C}_{(j,k-1)}$. Once every observation in the j-th block is handled, the evolution equation is applied to the process and the inference proceeds to the next block. This cycle goes on until all the observations of every block are processed. Naturally, the complete specification of a DGLM-PPM requires an initial state denoted here by:

$$(\boldsymbol{\theta}_0 \mid D_0) \sim [\mathbf{m_0}, \mathbf{C_0}]. \tag{5.26}$$

5.3.2 Gibbs Sampler

Due to the high computational effort required in the calculation of the posterior relevances, an alternative approach becomes necessary. In this section we present the Gibbs sampling algorithm proposed by Barry & Hartigan (1993) [79] and used by Loschi & Cruz (2002) [80] to obtain inference about the partitions and parameters of a PPM.

Let $\mathbf{U}$ be an auxiliary random vector of length $n-1$ whose i-th component is defined as:

$$U_i = \begin{cases} 1, & \text{if} \quad \theta_i = \theta_{i+1} \\ 0, & \text{if} \quad \theta_i \neq \theta_{i+1} \end{cases}$$

Each element $U_i \in U$ for $i = 1, 2, \cdots, n-1$ is an indicator variable that takes on zero or one whether or not the process goes through a change-point at the i-th observation. Observe that any partition $\rho = \{i_0, i_1, \cdots, i_b\}$ gets completely defined with knowledge of the vector $U = (U_1, \cdots, U_{n-1})^\intercal$.

The Gibbs Sampler is initiated with the vector $\mathbf{U}^0 = (U_1^0, \cdots, U_{n-1}^0)$. In the step $s > 1$ a new vector is generated with the form $U^s = (U_1^s, \cdots, U_{n-1}^s)$, where the r-th element, U_r^s, is generated from the following conditional distribution:

$$f(U_r \mid U_1^s, \cdots, U_{r-1}^s, U_{r+1}^{s-1}, \cdots, U_{n-1}^{s-1}; Y_1, \cdots, Y_n),$$

with $r = 1, \cdots, n-1$.

To generate new samples from the partition ρ, Loschi & Cruz (2002) propose the use of the following ratio:

$$R_r = \frac{P(U_r = 1 \mid A_r^s; Y_1, \cdots, Y_n)}{P(U_r = 0 \mid A_r^s; Y_1, \cdots, Y_n)}, \tag{5.27}$$

where $A_r^s = \{U_1^s = u_1, \cdots, U_{r-1}^s = u_{r-1}, U_{r+1}^{s-1} = u_{r+1}, \cdots, U_{n-1}^{s-1} = u_{n-1}\}$ and $r = 1, \cdots, n-1$. Considering an uniform prior cohesion for the blocks, Equation (5.27) become:

$$R_r = \frac{P(\mathbf{Y} \mid \rho)}{P(\mathbf{Y} \mid \rho^*)}, \tag{5.28}$$

where $\rho = \{U_1 = u_1, \cdots, U_r = 0, \cdots, U_{n-1} = u_{n-1}\}$ and $\rho^* = \{U_1 = u_1, \cdots, U_r = 1, \cdots, U_{n-1} = u_{n-1}\}$. That is, Equation (5.27) becomes a ratio between predictive distributions given two different partitions: one with a change point at the r-th observation and the other without. Remember that the predictive distribution has a closed form expression given by equation (5.13). Thus, a criterion for accepting or rejecting the values obtained for U_r^s is

$$U_r^s = \begin{cases} 1, \text{ if} \qquad R_r \geq \frac{1-u}{u} \\ 0, \text{ otherwise} \end{cases}$$

where the values of u are sampled from a (0,1) Uniform distribution.

Denote by M the size of the chain generated in the Gibbs Sampler and suppose, by the sake of example, that the interest lies on the posterior mean of the state vector of the j-th block of the partition. Estimates for this parameter may be obtained as follows: for each $\rho^{(s)}$, $s = 1, 2, \cdots, M$ generated, first compute the block estimate of the parameter, namely:

$$\tilde{\boldsymbol{m}}_{jk}^{(s)} = \boldsymbol{m}_{jk} \tag{5.29}$$

with $j = 1, \cdots, b$ and $k = 1, 2, \cdots, n_j$. The product estimate may be calculate by averaging this quantity over all the partitions generated, that is:

$$\tilde{\boldsymbol{m}}_{\boldsymbol{jk}} = \frac{\sum_{s=1}^{M} \tilde{\boldsymbol{m}}_{\boldsymbol{jk}}^{(s)}}{M}. \tag{5.30}$$

This procedure can be extended in an analagous fashion to all the other parameters of the DGLM-PPM.

5.3.3 Discount Factor

In the DGLM class, the usual strategy to specify the variance matrix of the evolution equation, as seen in Chapter 3, is via discount factors (denoted by δ). This approach is very appealing since it handles the uncertainty about the state parameters in an elegant and straightforward fashion. A decrease in the discount factor is associated with a larger variance $\mathbf{W}_t$ in the state evolution and, consequently, with a larger uncertainty in the parameters. In a DGLM setting, in most cases, it is reasonable to choose values of discount factors that are close to one since it is expected that a time series should retain most of its information when transitioning from one state to the next.

In the context of the DGLM-PPM class, however, the specification of the matrix $\mathbf{W_j}$ using discount factors is much more difficult. Almeida (2016) [22] uses a grid of discount factors ranging from $\delta = 0.05$ to $\delta = 0.95$ and computes the posterior probability for each specification to determine the optimal discount factor value. The major drawback of this method is that it is very computationally expensive since it requires the repeated use of the Gibbs algorithm described in the previous section. If the grid has too many values for δ, the computational time can make the DGLM-PPM impracticable; if the grid is too short, the estimation will not be reliable. To circumvent this problem we propose, in this work, to append an ARMS step within the Gibbs Sampling procedure to compute estimates for the discount factor. This method, that, as far as we are aware, has never been used in a similar context before, solves the main problems related to the grid approach since it is much faster and also provides more accurate point and interval estimates for δ . The variance matrix identified with the block j can be written in terms of the discount factor δ according to the following expression:

$$\mathbf{W}_j = \frac{(1-\delta)\mathbf{G_j}\mathbf{C_{j,n_j}}\mathbf{G_j'}}{\delta} = \frac{1}{\delta}\mathbf{R}_{j,n_j}.$$

Observe that the prior variance of the state vector is a function of δ and this dependency on the discount factor is carried throughout the whole inference cycle via the hyperparameters α_{jk} and β_{jk}. Thus, given the complete information set and the partition, and since the block predictives given by Equation (5.21) are implicit functions of the discount factor, the posterior distribution for δ will be given by

$$p(\delta \mid D^{(b)}) \propto \left[\prod_{j=1}^{b}\prod_{k=1}^{n_j} p(Y_{jk} \mid D^{(j-1,k-1)}, \rho)\right] p(\delta), \tag{5.31}$$

where $p(\delta)$ represents a prior distribution for the discount factor and $D^{(b)}$ is the complete information set up to the last block b. Since $\delta \in [0, 1]$, the Beta distribution arises as natural candidate to be the prior distribution for the discount factor δ . Samples from Equation (5.31) can be obtained efficiently using the ARMS [26] algorithm shown in Chapter 2.

The interpretation of the discount factor in the DGLM-PPM is also not as straightforward as in the classical DGLM. As already shown, the evolution equation is used to pass on information between blocks. Each block may contain several observations and, within the blocks, there is only a process of aggregation of the observation through the updating steps, thus leading to a smaller variance. For this reason, the block structure is likely to induce lower values of discount factor for the DGLM-PPM in relation to the conventional DGLM since high values for δ would imply very litle uncertainty about the state vector of the next block.

Penalized discount factor

Another possible approach is to weight the discount factor according to a measure of effective block size. The reasoning behind this idea is to penalize the larger components of the partition since they aggregate several $\boldsymbol{\theta_k}$. Under these assumptions, the discount factor associated to the j-th block will be given by:

$$\delta_j = \delta^{z_j}, \tag{5.32}$$

with $z_j = \frac{s_j}{\bar{s}}$ where s_j denotes the number of observations within the j-th block and $\bar{s}$ is the mean number of observations of the b blocks in the partition. Thus, under this specification, the discount factor associated with the DGLM-PPM is dynamic. If $s_j = \bar{s} \; \forall \; j$, then $z_j = 1$ and there is no penalization.

5.3.4 Example: DGLM via PPM Poisson

Consider the time series y_{jk} with $j = 1, 2, \cdots, b$ and $k = 1, 2, \cdots, n_j$. The DGLM via PPM Poisson has the following basic structure:

- **Observation Equation:**

$$p(y_{jk}) = \frac{\mu_{jk}^{y_{jk}} \exp(-\mu_{jk})}{y_{jk}!}, \qquad \text{with } \mu_{jk} > 0;$$

- **Link Equation:**

$$\log(\lambda_{jk}) = \mathbf{F}'\boldsymbol{\theta}_{jk};$$

- **Prior for** μ_{jk}

$$p(\mu_{jk}) = \frac{\beta_{jk}^{\alpha_{jk}}}{\Gamma(\alpha_{jk})} \mu_{jk}^{\alpha_{jk}-1} \exp(-\beta_{jk}\mu_{jk});$$

- **Prior for δ**

$$p(\delta) = \frac{1}{\text{Beta}(r,s)}(1-\delta)^{r-1}\delta^{s-1};$$

- **Initial Conditions:**

$$(\boldsymbol{\theta}_0 \mid D_0) \sim [\mathbf{m_0}, \mathbf{C_0}].$$

The gamma prior for the mean of the process μ_{jk} is the conjugate prior for the Poisson distribution. The hyperparameters α_{jk} and β_{jk} can be elicited in terms of f_{jk} and q_{jk} using the first two moments of the linear predictor and the first order approximations for the digamma and trigamma functions. That yields:

$$\alpha_{jk} = \frac{1}{q_{jk}} \qquad \text{and} \qquad \beta_{jk} = \frac{\exp(-f_{jk})}{q_{jk}}$$

The posterior distribution for μ_{jk} can be easily obtained from the conjugacy assumption:

$$p(\mu_{jk} \mid D_{jk}) \sim \text{Gamma}(\alpha_{jk} + y_{jk}, \beta_{jk} + 1).$$

Thus, the posterior moments of the linear predictor are given by the following equations:

$$f^*_{jk} \approx \log\left(\frac{y_{jk} + \alpha_{jk}}{\beta_{jk} + 1}\right),$$

$$q^*_{jk} \approx \frac{1}{y_{jk} + \alpha_{jk}}.$$

And now, from equation (5.21) we can write the predictive distribution of the model as:

$$p(y_{jk} \mid D^{(j-1,k-1)}) = \frac{\Gamma(\alpha_{jk} + y_{jk})}{y_{jk}!\Gamma(\alpha_{jk})} \frac{\beta_{jk}^{\alpha_{jk}}}{(1+\beta_{jk})^{y_{jk}+\alpha_{jk}}}, j = 1, \cdots, b; k = 1, 2, \cdots, n_j.$$

The posterior distribution for the discount factor can be obtained by direct application of equation (5.31):

$$p(\delta \mid D^{(b)}, \rho) \propto \left[\prod_{i=1}^{b}\prod_{k=1}^{n_j} \frac{\Gamma(\alpha_{jk} + y_{jk})}{y_{jk}!\Gamma(\alpha_{jk})} \frac{\beta_{jk}^{\alpha_{jk}}}{(1+\beta_{jk})^{y_{jk}+\alpha_{jk}}}\right] (1-\delta)^{r-1}\delta^{s-1}.$$

It is important to remember that, in the equation above, α_{jk} and β_{jk} are implicit functions of δ. Updating of the state vector can be done through equations (5.24) and (5.25). Finally, inferences about the partition ρ can be performed using the Gibbs Sampling scheme already shown. In this work, we assume that there is no prior information available regarding the block structure. Thus, we can assign a discrete uniform prior cohesion for the partition, that is: $c_\rho^{(j)} = 1$, $\forall\ j$. This choice implies that the prior distribution for ρ will be given by

$$p(\rho = \{i_1, i_2, \cdots, i_b\}) = \frac{1}{2^n - 1},$$

Observe that $2^n - 1$ represents the total number of possible partitions for a data set containing n observations. The uniform prior implies that the expected number of clusters in the partition will be $\mathsf{E}(B) = \frac{n-1}{2} + 1$ (see Loschi & Cruz (2005) for details [82]). Considering an uniform prior cohesion for ρ, samples from the partition can be obtained using the Gibbs Sampler already described with the ratio R_r of Equation (5.28). So, given the transition matrix $\mathbf{G}_j$, the design vector $\mathbf{F}_j$, the value of δ, the partition U and the initial values for $\mathbf{a}$ and $\mathbf{R}$, the DGLM via PPM Poisson follows the algorithm described in the pseudo-code (3). This same structure can be adapted when the observation equation follows any other distribution in the Exponential Family

Algorithm 3: Basic DGLM via PPM algorithm

Read $y_1, \cdots, y_n$
Set initial values for a, R, δ and U
$N \leftarrow length(Y)$
Function `DGLMPPM(`a_0, R_0, δ, U`)`**:**
  for $i \leftarrow 1$ **to** N **do**
    $a[i] \leftarrow a_0$
    $R[i] \leftarrow R_0$
    $f[i] \leftarrow Fa[i]$
    $q[i] \leftarrow F'R[i]F$
    $\alpha[i] \leftarrow \frac{1}{q[i]}$
    $\beta[i] \leftarrow \frac{\exp(-f[i])}{q[i]}$
    $pred[i] \leftarrow \frac{\Gamma(y[i]+alpha[i])}{y[i]!\Gamma(\alpha[i])} \frac{\beta[i]}{(1+\beta[i])^{y[i]+\alpha[i]}}$
    $f^*[i] \leftarrow \log(y[i] + \alpha[i]) - \log(\beta[i] + 1)$
    $q^*[i] \leftarrow \frac{1}{y[i]+\alpha[i]}$
    $m[i] \leftarrow a[i] + \frac{1}{q[i]} R[i]F(f^*[i] - f[i])$
    $C[i] \leftarrow R[i] - \frac{1}{q[i]} \left[R[i]FF'R[i] \left(1 - \frac{q^*}{q[i]}\right)\right]$
    if $U[i] == 1$ **then**
      $a_0 \leftarrow m[i]$
      $R_0 \leftarrow C[i]$
    else
      $a_0 \leftarrow Gm[i]$
      $R_0 \leftarrow \frac{GC[i]G'}{\delta}$
    end
  end
  $LL \leftarrow \sum(\log(pred))$ `// Observed Log-Likelihood`
  return *LL, a, R, m, C*
End Function

5.4 Simulation Study

Aiming to evaluate the properties of the DGLM via PPM class a simulation study was carried out. We generated $L = 500$ samples of size $n = 100$ observations from a Poisson Local Level Model (PLLM). With the objective of obtaining more well-behaved samples, we multiplied the auto-regressive term of the model by a constant $\phi = 0.99$, that is

$$y_t \sim \text{Poisson}(\exp(\gamma_t))$$
$$\gamma_t = \phi\gamma_{t-1} + \eta_t, \qquad \eta_t \sim N[0, W_t],$$

where γ_t represents the level of the process at time t. With this specification the PLLM becomes stationary. In order to create the change-points, jumps were inserted in the level γ_t at observations y_{25}, y_{50} and y_{75}, thereby creating three breaks in the samples. These jumps were designed so that the expected value of the process shifts by 15 units. For this experiment, we also considered the variance of the evolution equation W to be constant and set its value at 0.01. All simulations were run using the version 3.5.2 of the software R Core Team (2017) [63]. Figure (5.2) presents an example of a time series simulated according to the setting just described.

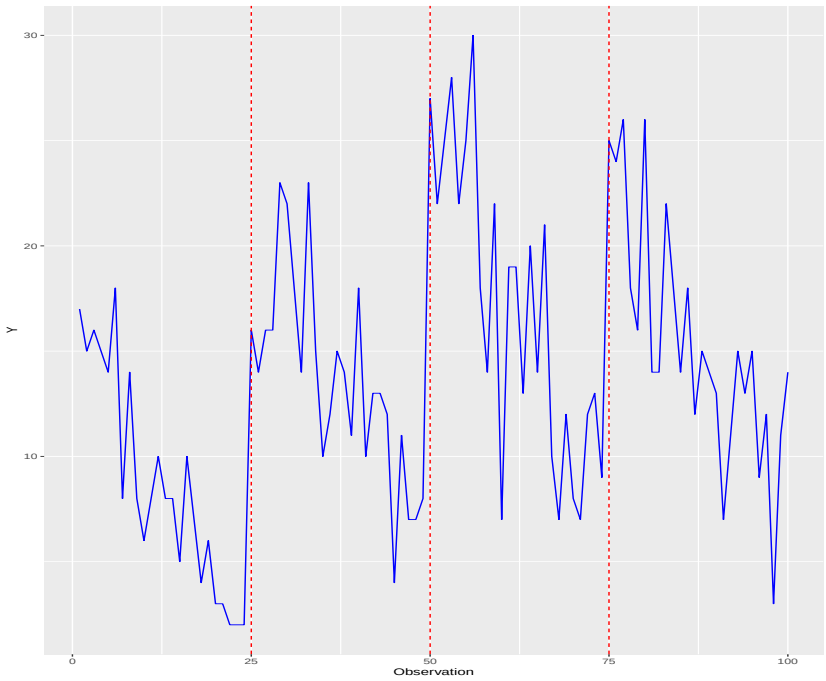


Figure 5.2: Process simulated from a Poisson LLM. The dashed lines indicate the points where the change-points were introduced.

In this simulation study we chose to work only with the simpler scenario of a non-penalized discount factor. Because of the lengthy computational time of the Gibbs Sampling scheme described in previous sections, we opted for generating 4000 samples of the partition $\mathbf{U}$ and of the discount factor δ. Preliminary analysis using the tools available on the R 'coda' package [66] showed that a burn-in of 1000 samples should be enough for the chains to achieve convergence.

To avoid auto-correlation among chains we considered a lag of 3. Thus, we are working here with a net sample size of $M = 1000$. Loschi & Cruz (2002) [80] also report fast convergence and low correlation for the Gibbs algorithm in a similar setting. We assume in this experiment no prior information about the values of the discount factor. For this reason, we assigned a Beta$(1, 1)$ prior distribution for δ. Note that, with this parametrization, the Beta distribution is equivalent to a Uniform (0,1). Estimates of the probability that, at any given point y_r of the time series, there is a change-point can be computed by taking sample averages over the partitions generated in the MCMC procedure, that is:

$$Pr^{(i)}(y_r \text{ is a change point} \mid \mathbf{Y}, \varrho) = \frac{\sum_{i=1}^{M} U_r^{(i)}}{M}$$

where $\varrho = \{\rho^{(1)}, \rho^{(2)}, \cdots, \rho^{(i)}, \cdots, \rho^{(M)}\}$ and $U_r^{(i)}$ corresponds to the r-th observation of the i-th partition (ρ_i) sampled by the Gibbs algorithm.

For a better visualization, the probabilities of regime switching for each observation from the 500 Monte Carlo replications were organized in box plots. The results are displayed in Figure (5.3). As expected, the DGLM-PPM attributes larger probabilities of structural change for those observations in which the jumps in the mean of the process were inserted.

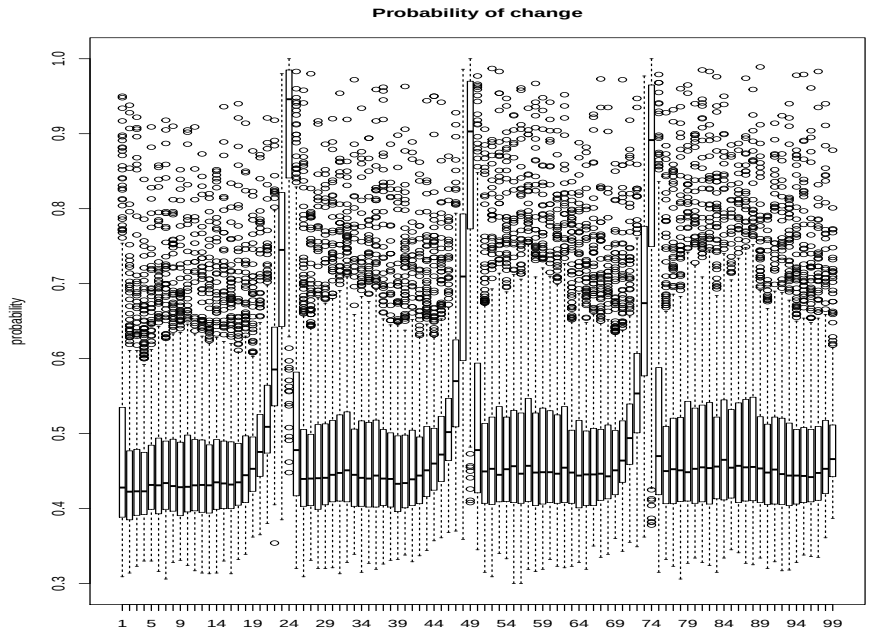


Figure 5.3: Probability of change associated with each observation according to the DGLM-PPM.

We are also interested in the behaviour of the discount factors of the DGLM and of the DGLM-PPM. The samples drawn from the posterior distribution (5.31) via ARMS were used to calculate point and interval estimates for the δ. The Bayes estimator used in this work was the posterior mean (from here on, denoted by BE-Mean). The credible intervals of 95% Highest Posterior Density (HPD) were calculated with the help of the tools available in the 'coda' package [66]. The superior and inferior limits were computed as sample averages over the 500 Monte

Carlo replicas. To measure the variability associated to the BE-mean, the Empirical Standard Error (ESE) was used. Table (5.1) presents the results obtained

	$\hat{\delta}$	CI	ESE
DGLM-PPM	0.31	[0.40, 0.61]	0.047
DGLM	0.51	[0.19, 0.43]	0.075

Table 5.1: Point and interval estimates for the discount factor of the DGLM-PPM and DGLM

The point estimates show a higher discount factor for the conventional DGLM, an outcome that was expected and is in conformity with the discussion presented in Section (5.3.3). We also observe a higher ESE for the estimation of δ in the conventional DGLM in relation to the DGLM-PPM, implying a smaller variability in the estimates for the later model. To visualize this difference the posterior distributions of the discount factor associated with the DGLM and DGLM-PPM are displayed as boxplots in Figure (5.4). It is noteworthy from the plots that the estimation of δ for the conventional DGLM presents a slightly higher variability in relation to what is observed for the DGLM-PPM, thus confirming the ESE results.

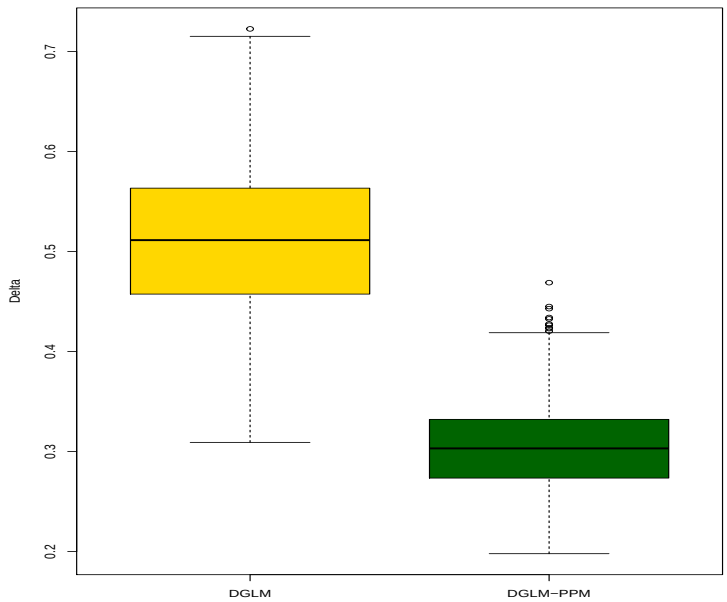


Figure 5.4: Comparison of the discount factors estimated via ARMS. In yellow the estimates for the conventional DGLM, in green the estimates for the DGLM via PPM

The DGLM via PPM also allow us to obtain posterior estimates for the number of blocks in the partition. Figure (5.5) displays the empirical distribution of the posterior expected number of blocks considering the 500 Monte Carlo samples generated for this experiment. The boxplot shows that the mean number of blocks is concentrated around 50. This result is not expected and is not satisfactory from an inferential point of view since we expected values much smaller than those observed. We believe that this behavior is an outcome of the uniform prior cohesion

chosen for the model. As already shown, the prior expected number of blocks for a time series of size $n = 100$ under a uniform prior would be $\mathsf{E}(B) = \frac{n-1}{2} + 1 = 50.5$. Thus, the results of this experiment show that the posterior number of blocks is dominated by the prior cohesion and has litle influence from the data. An alternative to stimulate the formation of blocks with a larger number of observations would be to assume the prior cohesion defined by Yao (1984) [85].

$$c_{\rho}^{(j)} = \begin{cases} p(1-p)^{i-j-1}, & \text{if} \quad j < n \\ (1-p)^{i-j-1}, & \text{if} \quad j = n, \end{cases}$$

where $0 \leq p \leq 1$ represents the probability that a change-point occurs at the instant t. Values of p closer to zero would induce the formation of partitions with a small value for the prior expected number of blocks. Another approach, would be to define the prior cohesion as $c_{\rho}^{(j)} = n_j$, $\forall\, j = 1, \cdots, b$, where n_j is the number of observation in the j-th cluster. This way, the model would assign higher probabilities to partitions with lesser components. The study of these different prior cohesions is left as a future work.

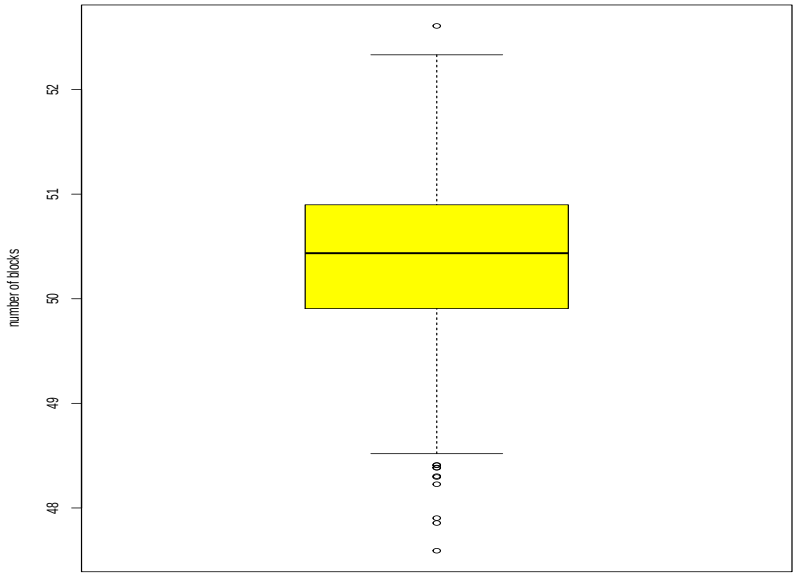
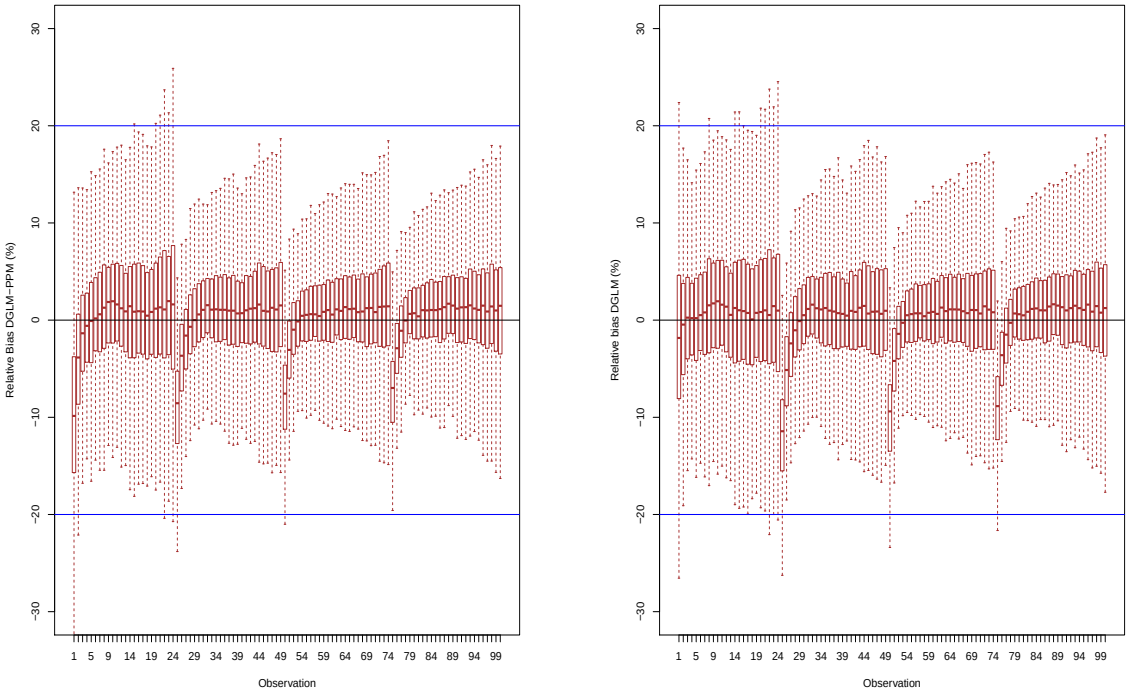


Figure 5.5: Empirical distribution of the posterior expected number of blocks in the partition.

To assess the degree of adherence between the parameters estimated by the DGLM-PPM and the real underlying process we proceeded to compare the estimates obtained for the posterior mean of the state vector to the real ones generated for this study. Remember that the state vector is only partially specified in the terms of its first two moments, that is: $\theta_{jk} \sim [\mathbf{m}_{jk}, \mathbf{C}_{jk}]$. Our interest here lies on the posterior mean $\mathbf{m}_{jk}$. Denote the estimates of $\mathbf{m}_{jk}$ by $\hat{\mathbf{m}}_{\mathbf{jk}}$, the Relative Bias (RB) associated to this estimator, in percent, will be given by:

$$RB(\%) = \frac{100(\hat{\mathbf{m}}_{\mathbf{jk}} - \mathbf{m}_{jk})}{|\mathbf{m}_{jk}|} \tag{5.33}$$

We computed the RB in the estimation of $\mathbf{m}_{jk}$ for each observation across all Monte Carlo samples. For comparison purposes, the same procedure was repeated for the conventional DGLM using the same simulated data. The results obtained were organized in box plots and are displayed in the Figures (5.6a) and (5.6b) for the DGLM-PPM and the conventional DGLM, respectively. We call attention to the following facts related to the DGLM-PPM estimates: i) all observations considered, the mean RB is small (0.34%). Besides, the individual biases rarely surpass 20% except for the change-points, suggesting good adjustment of the DGLM via PPM; ii) as indicated by the black horizontal line set at $RB = 0$, even though the RB are small, they are systematically positive as the process moves away from the breaks; iii) the biases are much larger on the first three observations following a change-point.



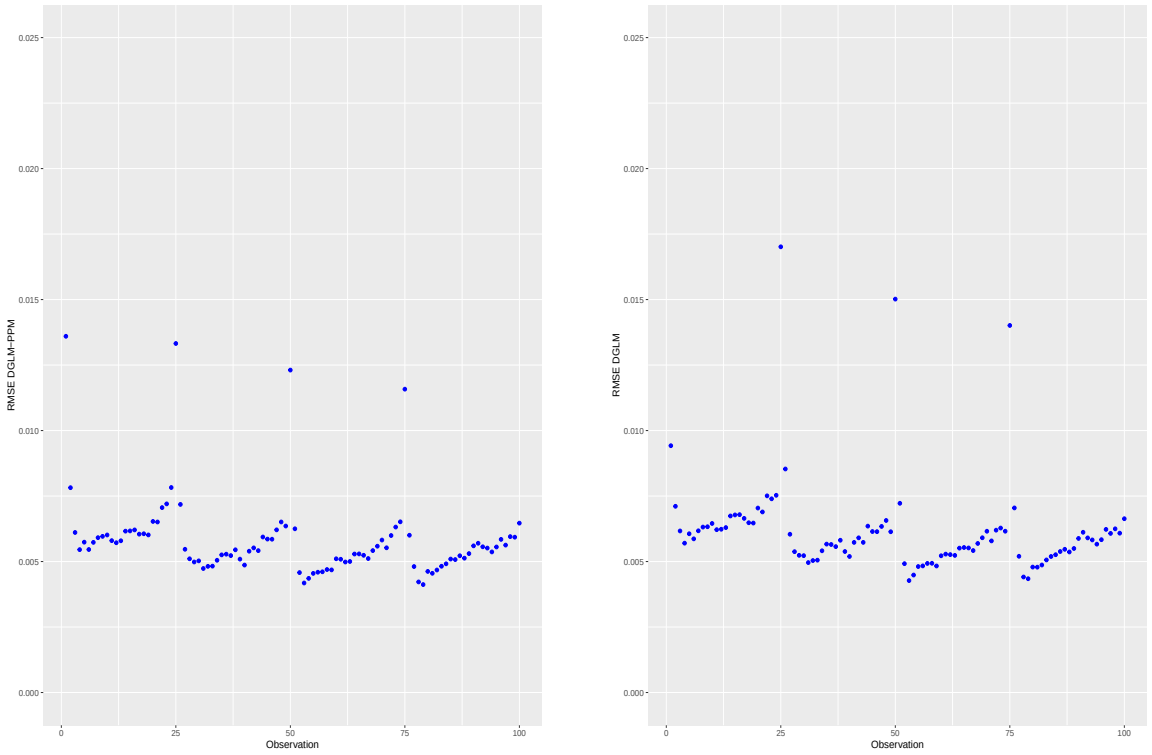
(a) (b)

Figure 5.6: Box Plots of the Relative Bias associated to the estimation of the mean of the posterior state vector for the DGLM-PPM (a) and the DGLM (b). The horizontal blue lines indicate the levels of $RB = 20\%$ and $RB = -20\%$. The horizontal black line is set at $RB = 0$.

Point (iii) is easy to understand since it is natural that the model takes some time to adapt to a regime switch. Observe, however that, around the change-points, the variability associated with the DGLM-PPM estimates is slightly smaller than that of the conventional DGLM. To further illustrate this fact we present, in Figures (5.7a) and (5.7b), the RMSE [1] related to the estimates of $\mathbf{m_{jk}}$ for the DGLM-PPM and the conventional DGLM. It is possible to spot a small, but noticeable, difference in the RMSE in favor of the DGLM-PPM at the change-points, thus

[1] $RMSE = \sqrt{\frac{\sum_{i=1}^{500}(\hat{\mathbf{m}}_{\mathbf{jk}} - \mathbf{m}_{\mathbf{jk}})^2}{500}}$

giving our model an edge over the conventional DGLM specially when the time series under study has structural breaks.

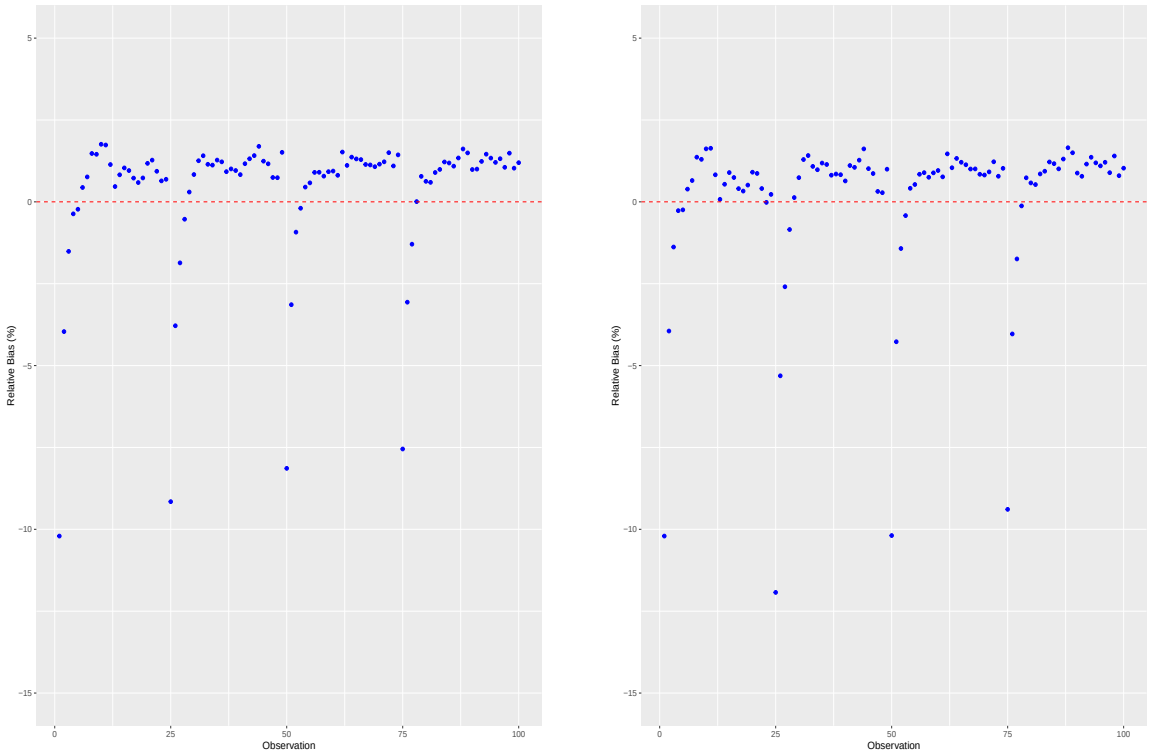


(a) (b)

Figure 5.7: RMSE calculated for the estimates obtained with the DGLM-PPM and DGLM.

As for point (ii), we believe that the persistent estimation bias can be caused by one, or both, of the following two reasons: the first is that the updating of the state vector is not exact and rely on several linear approximations that can cause the observed bias; the second is that we are working with a relatively small sample of 100 observations and the convergence to RB levels close to zero in the DGLM class might be slow, perhaps even more so in the presence of change-points. We proceeded, then, to investigate if, despite the persistent bias, the estimates obtained with the DGLM-PPM have any advantage over those obtained with the conventional DGLM. To do that, we examined the mean RB associated to each observation, that is: $\overline{RB_{jk}} = \frac{1}{500}\sum_{i=1}^{500} RB_{jk_i}$ – for the two models. Figures (5.8a) and (5.8b) show the results obtained. Observe that the plots are very similar – in both cases a small ($< 2\%$) positive RB is evident for the observations that are in-between change-points. We can also see from the figures that, for both models, there is a kind of learning process in which the large biases induced by the change-points get progressively smaller as the observations of the time series are processed. This effect, however, is more noticeable in the DGLM-PPM: not only the model fits the observations immediately after the jumps better than the conventional DGLM, but it also converges back to the levels of lower bias faster. This outcome is probably due to the lower discount factor induced by the block structure that allows the model to better adapt to the regime switches. For this reason, the DGLM-PPM is likely to outperform the conventional DGLM in the presence of

change-points. This advantage will become more relevant as the number of jumps in the data increases.



(a) (b)

Figure 5.8: Mean RB for the estimates of the posterior mean of the state vector with the DGLM-PPM and the DGLM. The dashed red lines indicate $RB = 0$.

5.5 Real data application

This section is reserved for the application of the DGLM via PPM to real data sets. Two well known time series were selected: the coal mining accidents counts and the Nile River flow. To obtain inference about the partitions we used the Gibbs Sampling scheme described in the Section 3. Since the applications do not require the repeated use of the Gibbs Sampler algorithm as in the simulation experiment, we opted to work with longer chains. For each data set analyzed a single chain of size 45000 was generated and, after a burn-in of 5000 and a lag of 10, 4000 posteriors samples were obtained. The final configuration was determined after some tests using different settings and with the use of the diagnostic tools for MCMC offered by the 'coda' package from R [66]. The prior cohesion for the partition is assumed to be uniform, that is, $c_\rho^{(j)} = 1, j = 1, \cdots, b$. No explanatory variables are available for the two data sets chosen for these applications; besides that, a visual inspection of the time series plots does not indicate any signs of trend or seasonality. Thus, in both examples, the models were defined as Local Level Models. The specification of the variance of the evolution equation is made via discount factor

as discussed earlier. In the first application both the penalized and non-penalized δ are tested while in the second one only the non-penalized version is used. The estimation of δ was made via ARMS [26] according to the method already described in this chapter. The results obtained with the DGLM-PPM were compared to those of the conventional DGLM using the Posterior Model Probability as an in-sample measure of model performance and the MAE and MSE as an out-of-sample measure.

5.5.1 Coal mining data

The coal mining data reports the annual number of coal mining disasters involving more than ten men in the UK from 1851 to 1962 totaling 112 observations. This time series has been extensively used in the literature for the study of change-points in discrete data. Some examples include Raftry and Akman (1986) [17], Carlin, Gelfan and Smith (1992) [86], Santos, Franco and Gammerman (2010) [87], Lai and Xing (2011) [77] and da Silva and da-Silva (2016) [19]. Figure (5.9) shows the coal mining disasters counts. Observe that the annual number of accidents decreases at the end of the nineteenth century, indicating a possible change-point around the period that extends from 1885 to 1895.

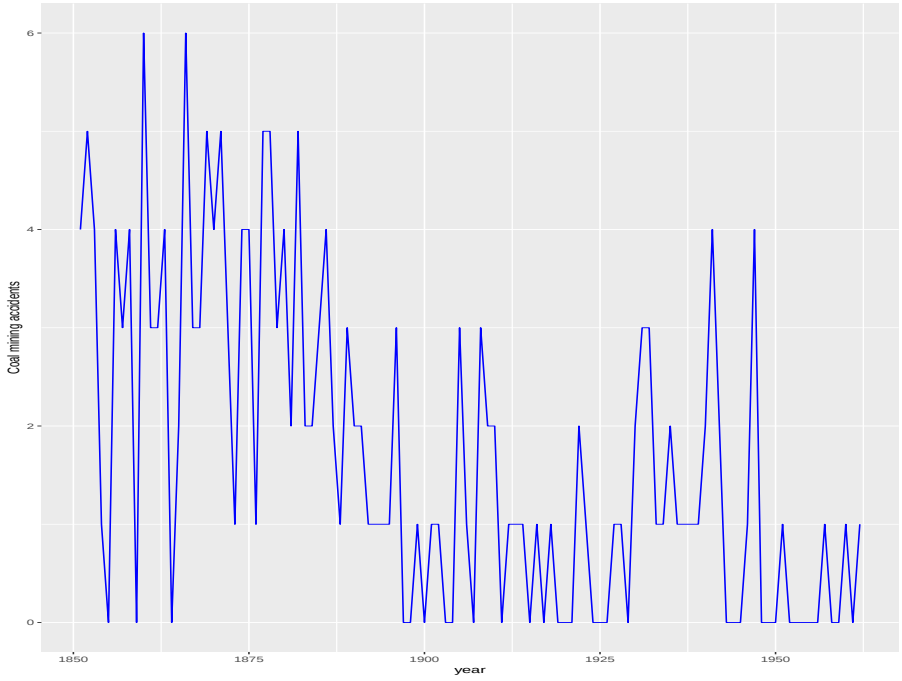


Figure 5.9: Annual number of coal mining disaster involving more than 10 men in the UK from 1851 to 1962.

Since we are dealing with a time series of counts, a natural modeling choice is the Poisson distribution. Three models are in consideration for this exercise: the conventional Poisson DGLM and the Poisson DGLM-PPM with non-penalized and penalized discount factor. Table (5.2) presents the estimated values of δ associated with each of the models tested. We also display the most probable number of blocks of the partition according to the DGLM-PPM in

both cases studied. The results are shown along with 95% credible intervals. We observe that the estimated discount factors for the PPM models are lower than those of the conventional DGLM – a result that is in accordance with the simulation study and the analysis carried out in the last section. Also, both specifications of the PPM assign a high number of blocks to the partition. As already discussed, this unexpected result was detected in the simulation study and is probably a consequence of the uniform prior cohesion chosen for this exercise.

	DGLM	**DGLM via PPM**	**DGLM via PPM with penalized discount factor**
Estimated δ	0.8248 (0.7368, 0.9012)	0.6944 [0.5521, 0.8316]	0.5933 [0.3169, 0.8728]
Number of blocks	-	56.3448 [45, 65]	56.3875 [46, 66]

Table 5.2: Estimated discount factor and posterior most probable number of blocks for the three models adjusted to the coal mining data set.

Figures (5.10) and (5.11) offer a visualization of the empirical distributions of the posterior samples obtained for δ and the number of blocks in the partitions, respectively. As usual, the results are organized in boxplots.

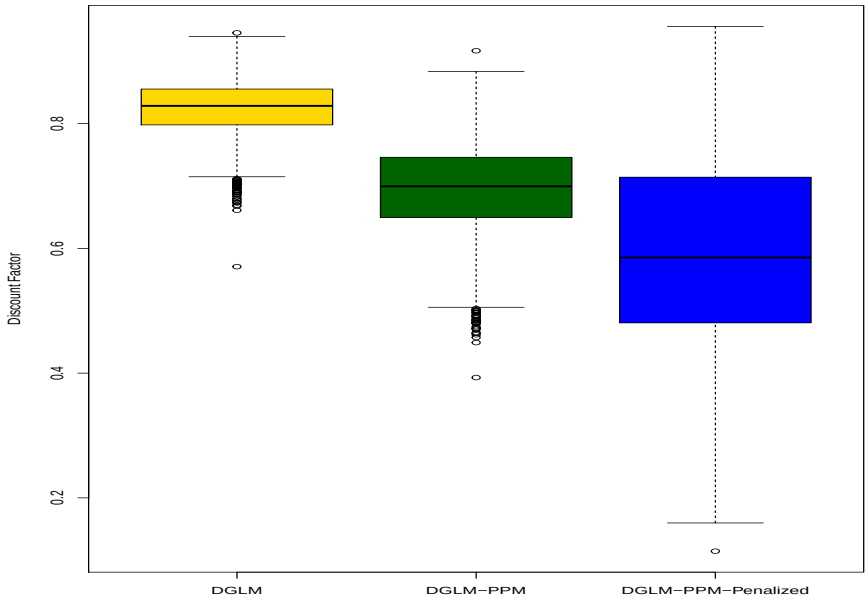


Figure 5.10: Boxplots of the posterior samples of the discount factor obtained for each of the models tested

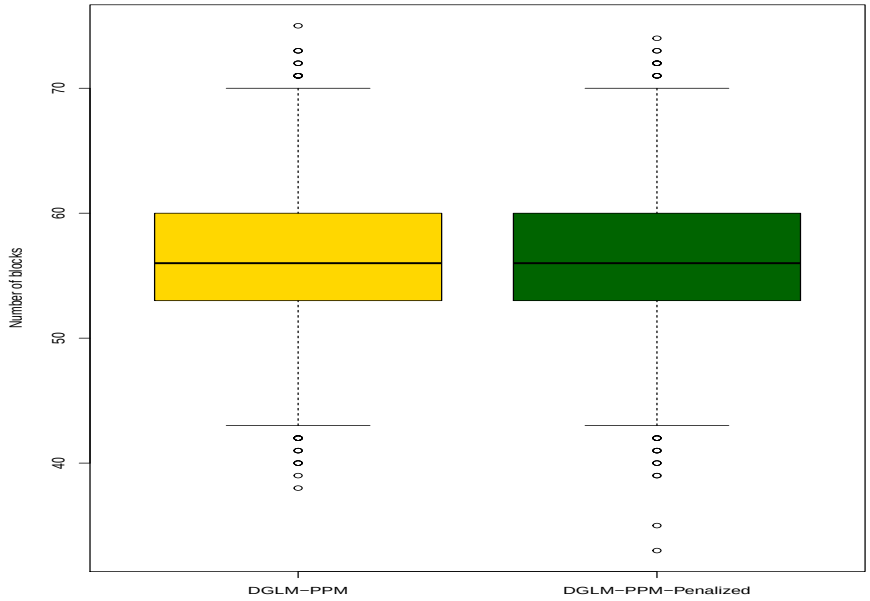


Figure 5.11: Boxplots of the posterior empirical distribution of the number of blocks for the DGLM-PPM and the DGLM-PPM with penalized discount factor.

From (5.10), it is noteworthy that the discount factors estimated by the DGLM-PPM with penalized δ have a much higher variability than its non-penalized counterpart. We believe that this result is a consequence of the asymmetry in the sizes of the blocks that comprise the partition. The following two scenarios must be examined separately:

1. Most blocks in the partition have a size (s_j) smaller than the mean size ($\bar{s}$);

2. Most blocks in the partition have a size larger than the mean size;

In the first case, the exponent z_j of each block will be smaller than one for most blocks in the partition. What this means is that the effective discount factor associated with these blocks will be larger than the value of δ estimated for the whole partition. In this case the penalization will cause the estimated discount factor to be smaller than the one that would be obtained without the penalization. In the second situation, the z_j will be larger than on and the opposite effect will be observed. Thus, the variability in the estimation of the discount factors is a direct consequence of the asymmetry of the partitions sampled by the Gibbs algorithm. A better understanding of this behavior would require the conduction of a simulation study similar to the one presented in the last section. As for the most probable number of blocks, Figure (5.11) shows very similar profiles whether the discount factor is specified with a penalization factor or not; thus suggesting that weighting the components according to their sizes does not alter the most probable number of blocks in the partition. Also, it should be pointed out that the estimated number of cluster is much larger than what is suggested by the data, a problem that was already detected in the simulation study of Section 5.4.

To assess in-sample and out-of-sample performance of the models three criteria were used: the posterior probability to evaluate in-sample adjustment and the MAE and MSE to measure prediction accuracy. The forecasts were calculated as the mean of the forecasts obtained for each of the partitions sampled by the Gibbs scheme. In order to compute the Posterior Probability it was assumed equal prior probabilities for each model. Table (5.3) shows the results obtained.

	MSE	**MAE**	**Posterior Probability**
DGLM	1.816738	1.046794	0.2402
DGLM via PPM	1.7972	1.0396	0.4544
DGLM via PPM with penalized discount factor	1.7997	1.0432	0.3054

Table 5.3: Performance summary of each of the models tested.

The three specifications used performed similarly, even though the DGLM via PPM is the clear winner among them since it presents the highest forecast accuracy and also the highest Posterior Probability. The DGLM via PPM specified with penalized discount factor has a worse performance than the DGLM via PPM but it also slightly outperform the conventional DGLM. The computational cost of the penalized specification is higher in relation to the non-penalized DGLM via PPM so, for this time series, we will be better off with the regular specification of δ. Figure (5.12) displays the one-step-ahead point forecasts made by the DGLM via PPM along with 95% credible intervals. Observe that the predictions line (in black) follows the data points smoothly. Also, all the data points are within the boundaries set by the credible intervals (dashed red lines), indicating a good fit.

One of the main advantages of the DGLM via PPM is that it allows the online inference provided by the DGLM class together with estimates of the change-point probability for every observation in the time series under study. Figure (5.13) displays the probability of change estimated for each point of the coal mining accidents data set (bottom half) along with the original time series plot (upper half). The vertical lines were inserted to serve as a visual guide for the observations where a regime change are most likely to occur. At least four peaks of high probability can be observed: the first around 1887, as expected; the second in 1930 and the third and fourth peaks by the 1940's and early 1950's. A close inspection at the time series reveals that around these periods there is a sudden rise on the number of accidents. The results obtained here are in line with those reported in the literature for the same time series. Lai & Xing (2011) use a Bayesian model for multiple breaks to estimate a posterior probability of change in the order of 40% for the points around 1950 and of 20% for the observations near 1890; these two periods being the ones with highest probability of regime switch. Also, in the work of da-Silva and da Silva (2016) the authors use a model based on the Chopin Filter that detects two possible change-points around the years of 1890 and 1945. Note, however, that the posterior probabilities estimated in this work still do not allow us to state with absolute certainty that a change-point

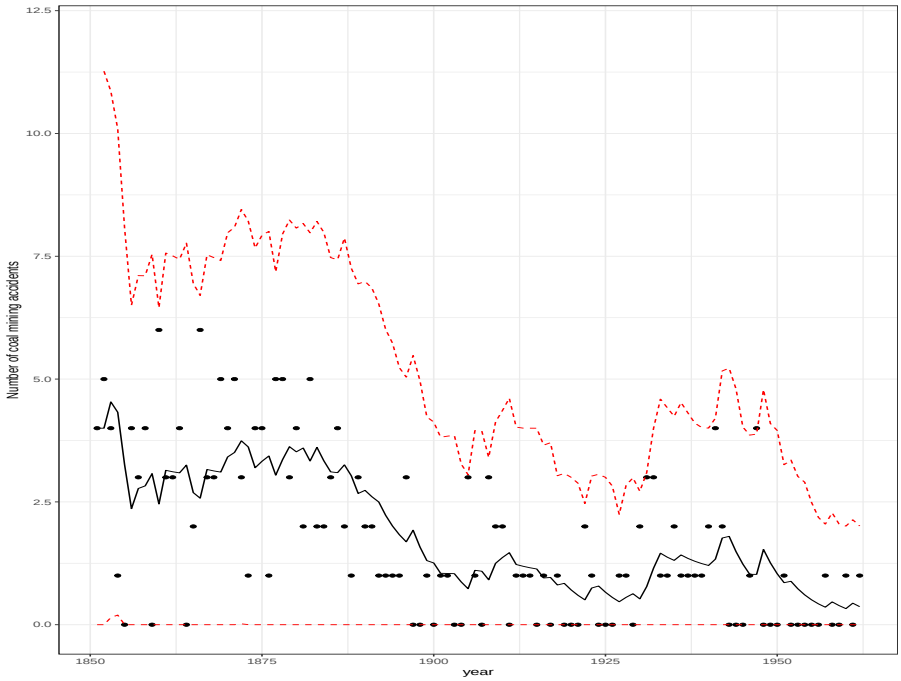


Figure 5.12: One-step-ahead forecasts for the coal mining accidents data using the DGLM via PPM. The solid black line represent the predictions; the dashed red lines indicate the 95% credible intervals.

occurred on the aforementioned years. This is because, as observed on the simulation study, there are cases of false positives in which probability changes around 60% are reported.

5.5.2 Nile River data

The Nile River data consists of 100 observations relative to the annual flow (in $10^8 m^3$) of the Nile River from 1871 to 1970. This data set first appeared in the literature in the work of Cobb (1978) [88] and, ever since, has been examined by several authors in the context of change-point analysis [2]. As can be seen from the figure (5.14), there is a permanent drop in the level of the time series from 1899 onwards. The proposed explanation for this phenomenon is that the Nile River experienced a decline in volume due to the construction of the first Aswan dam in 1898.

To model the Nile time series we used the Normal DGLM-PPM (or simply DLM-PPM) with the precision of the observation equation (ϕ) considered unknown. The basic structure of the Normal DGLM was outlined in section 3 of this work and the extension to a PPM is straightforward following the methodology described in this chapter. Based on the results obtained with the Coal mining data and considering the computational cost of the specification with penalized discount factor, we chose to test only the DGLM-PPM with non-penalized δ. Table (5.4) displays the estimates for the discount factor, posterior number of blocks and precision along with 95% credible intervals.

[2] From a Bayesian standpoint see, for example, the work of Moreno, Casella and Garcia-Ferrer (2005) [89]

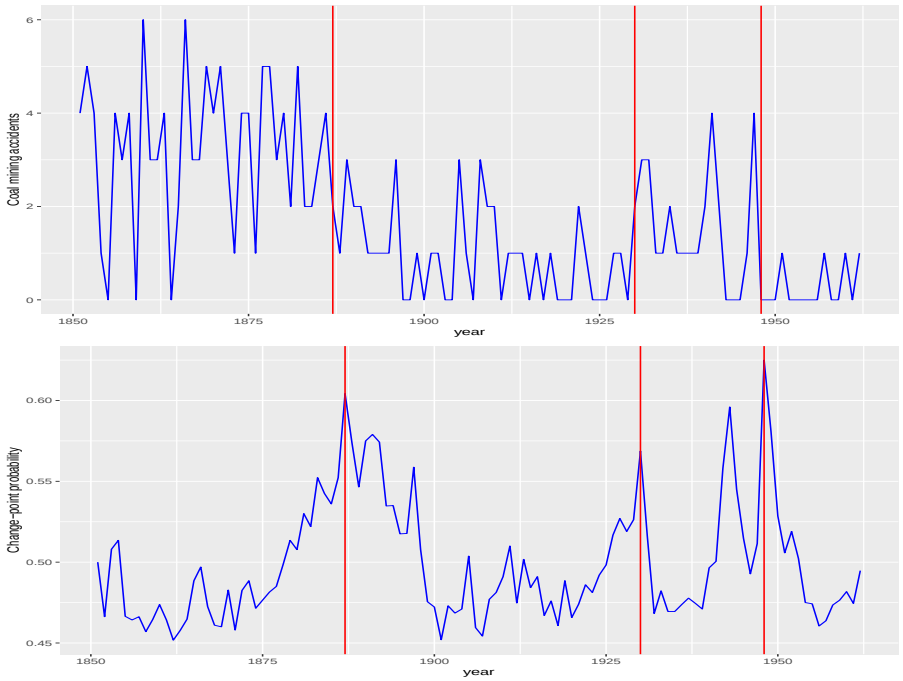


Figure 5.13: Change-point probabilities for the coal mining accidents time series in the bottom and the original time series plot in the upper half. The red vertical lines were inserted at points where the model assigned highest probability of a regime switch: 1887, 1930 and 1948.

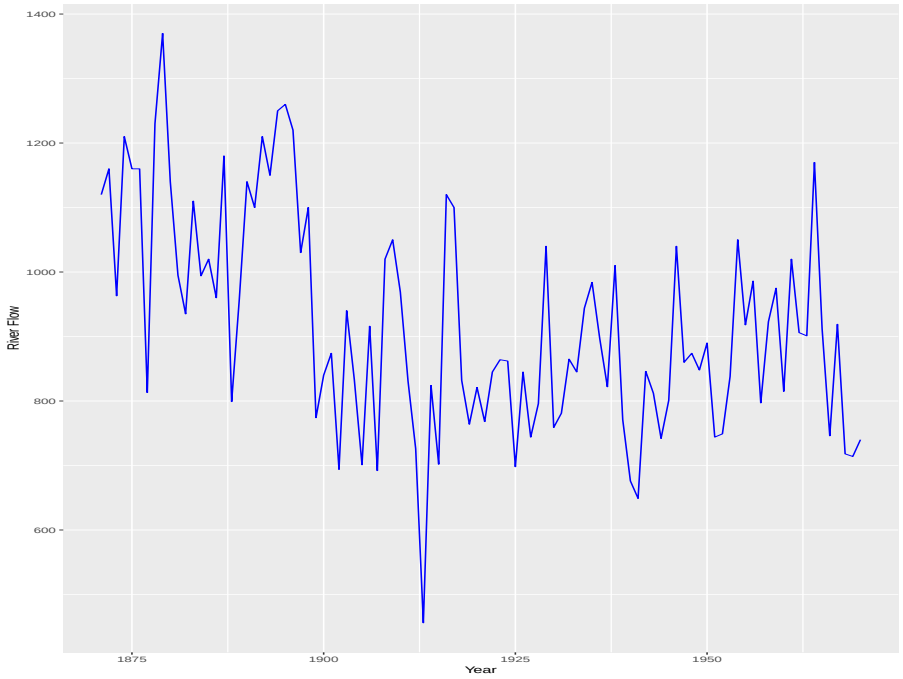


Figure 5.14: Annual flow of the Nile River from 1871 to 1970.

As expected, and in accordance with the simulation experiment, the discount factors estimated for the DGLM-PPM and the conventional DGLM are different; the later one having the highest value. We also observe that the estimates for the precision parameter are very small in both models and that the most probable number of blocks is around 50, thus making the average

	DGLM	DGLM-PPM
Estimated δ	0.6815749 (0.45487, 0.8766)	0.5282 (0.2642, 0.7834)
Estimated ϕ	7.250881e-05 (5.3903e-05, 9.3831e-05)	7.2088e-05 (5.3590e-05, 9.3286e-05)
Number of blocks	-	50.07925 (40, 59)

Table 5.4: Point and interval estimates for the discount factor, precision and most probable number of blocks of the DGLM and DGLM-PPM applied to the Nile River time series.

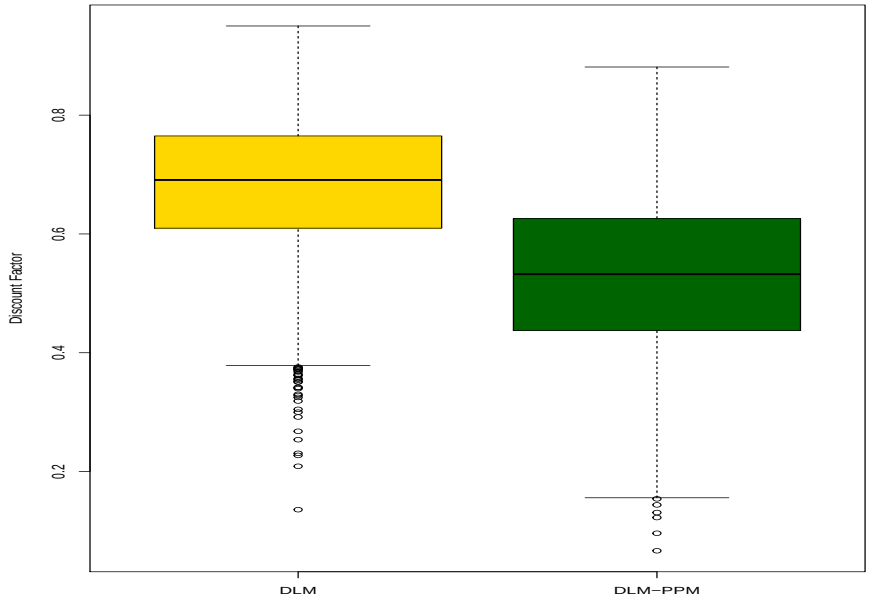


Figure 5.15: boxplots of the posterior samples of the discount factor obtained by each model for the Nile River data set.

	MAE	MSE	Posterior Probability
DGLM	113.1046	20596.39	0.1886
DGLM-PPM	110.2723	19605.87	0.8114

Table 5.5: Performance summary of the DGLM and DGLM-PPM applied to the Nile River time series.

number of observations per block equal to two. Figure (5.15) displays the empirical distributions of the posterior samples of δ for both the Normal DGLM and the Normal DGLM-PPM.

To compare between the two models the same measures of in-sample and out-of-sample performance utilized in the coal mining accidents example were used. The results are summarized in table (5.4). We observe, once again, a consistent advantage for the DGLM-PPM over the conventional DGLM as can be seen by the larger posterior probability and the smaller forecast errors.

Figure 5.16: One-step-ahead forecasts for the Nile River time series using the Normal DGLM-PPM. The solid black line indicates the point predictions and the dashed lines represent the 95% credible interval.

Figure (5.16) shows the one-step-ahead forecasts for the Nile River flow time series using the Normal DGLM-PPM along with 95% credible intervals (CI). The fact that very few observations fall outside the upper and lower bands determined by the CI levels indicates a good adjustment of the model. We also display the change-point probabilities for each observation in the data set (figure 5.17). The vertical red line indicates 1898 – the year with the highest break probability ($\sim 83\%$). This result suggests that the construction of the Aswan dam in fact caused a permanent decline in the river flow confirming the conclusions of other works in the literature.

5.6 Conclusions

Intervention analysis in a major concern in time series analysis. To address this problem we presented in this chapter the DGLM-PPM – a new class of models that incorporates the Product Partitions Models of Barry & Hartigan (1990) [20] into the theory of Dynamic Generalized Linear Models. We believe our model presents several improvements over the conventional DGLM. First, it allows the identification of structural breaks by assigning to each observation in the time series a posterior probability that it is a change-point. Unlike many models used to detect change-points, the DGLM-PPM does not require any prior knowledge regarding the observation wherein the regime switch happens, which can be an important advantage in many situations. Also, just like any other model in the DGLM class, the DGLM-PPM permits the estimation of filtered and smoothed states, thus allowing for k-step-ahead forecasts and retrospective analysis.

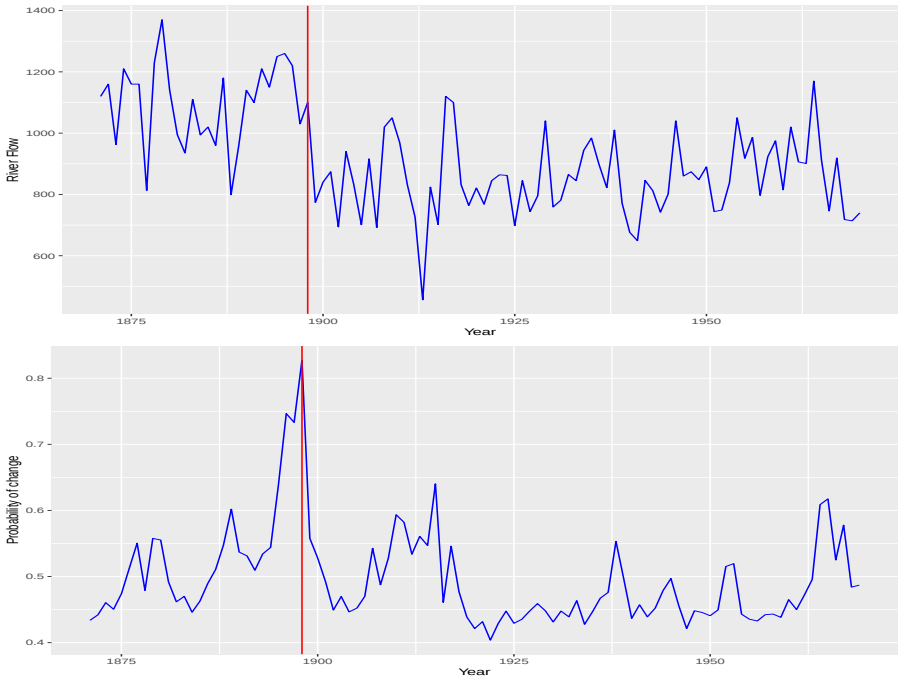


Figure 5.17: Change-point probability. The vertical red line indicates the year of construction of the Aswan dam.

Another important feature of the DGLM-PPM class is the possibility of making inferences about the optimal block structure for a determined data set. Samples of the partition can be obtained in an efficient manner using a Gibbs Sampler scheme proposed by Barry & Hartigan (1993) [79]. Within this Gibbs scheme we included an ARMS [26] step in order to sample from the posterior distribution for the discount factor.

To evaluate the DGLM via PPM we conducted a simulation experiment. From this study several conclusions could be drawn: i) the model does assign the highest probabilities of changes to the points where the jumps were artificially inserted; ii) the discount factor estimated via ARMS is lower for the DGLM via PPM in comparison to the conventional DGLM; iii) the estimates for the posterior first moment of the state vector is slightly biased, a fact that was attributed to the approximations involved in the inference procedure of the DGLM class and to the slow convergence of the model due to poor choices of initial values. Still, we showed that this bias is also present in the conventional DGLM and that the DGLM-PPM adapts better to a change-point; iv) the number of blocks estimated for the partition is larger than we would have expected. This last point is probably a consequence of the uniform prior cohesion specified for the models.

The DGLM-PPM was applied to two real life data sets that are well known in the change-point analysis literature: the UK coal mining accidents time series and the Nile River flow time series. For the coal mining data, the model was specified with penalized and non-penalized discount factors. We compared both specifications to the conventional DGLM using measures of in-sample and out-of-sample adjustment. The DGLM via PPM outperformed the conventional

DGLM in both specifications. The penalization of the discount factor, however, did not improve the model in this case. For the Nile river time series, we opted to use only the non-penalized specification. The results also show that the model we propose here is an improvement over the conventional DGLM. In both examples, the DGLM-PPM detected possible change-points that are in accordance with the findings of other works in the literature.

There are several ways in which the methods presented in this text can be expanded and refined. We consider that the most important problem to be addressed regarding the DGLM-PPM is the number of clusters induced in the partition by the uniform cohesion prior. As already discussed in Section 5.4, the specification of different prior cohesions can affect the expected number of clusters induced in the partition. An evaluation of these cohesions can lead to valuable insights about the structure of the DGLM-PPM. Another extension of this work would be to allow a stochastic evolution within blocks. In this case, two processes would run simultaneously: one intra-block and the other between blocks – each of them with a different discount factor. Other possibilities include a careful simulation study of the penalized discount factor to determine if it can improve the adjustment of DGLM-PPM and the inclusion of a seasonality component into the model. We also verified that the DGLM via PPM is costly from a computational point of view. Thus the implementation of the model using C^{++} would greatly improve its usability.

Chapter 6

Final Considerations

This dissertation presented several extensions of the class of Dynamic Bayesian Models. The work was organized in two strands: one aimed at expanding the class to distribution that are outside the Exponential Family or that belong to the EF but were not properly formalized as a Dynamic Model; the other concerned with incorporating the Product Partition Model class into the DGLM of West, Harrison & Migon (1985) [4] in order to deal with change-points problems.

Within the first strand, several new Dynamic Bayesian Models for single count data time series were introduced. A general method of inference for Bayesian models with uniparametric distributions as observation equations was developed. In this thesis, we worked out the cases of the Borel, Bell, Poisson-Lindley and Yule-Simon distributions as particular cases, but the framework developed here can be extended to any uniparametric distribution, including those with continuous response. We also present a novel Dynamic Negative Binomial Model (DNBM) in which the NB distribution is parametrized as in the Generalized Linear Models framework and the prior distribution for the mean of the process is a Beta Prime of the second kind. One of the main contributions of this work is that two estimators were proposed for the shape fixed parameter of the model: a classical one and a Bayesian one. A Monte Carlo simulation was carried out and the results have shown that, in all scenarios tested, the coverage probabilities were very close to the fixed nominal levels. We also noted that, as the sample sizes increase, the relative biases associated to the estimators decrease, thus indicating good performance. The framework in which the DNBM was built is very straightforward and computationally efficient making it a significant contribution to the literature of count time series. We also showed that this framework can be extended to other biparametric distributions belonging to the Exponential Family. In this work we presented the calculations for the Gamma and Weibull distributions. The conduction of a simulation experiment for these models and applications to real-life data sets were left as future works.

The second part of this thesis dealt with the problem of inference in the presence of change-points. Regime switches are a common feature of time series in many fields and that is why this topic has received a lot of attention in the literature in the last few decades. In this work,

we proposed a new class of models that combines the flexibility of the DGLM class of West, Harrisson and Migon (1985) [4] to the Product Partition Models of Barry and Hartigan(1992) [21]. This new class of models, named as DGLM -PPM, allows forecasting and retrospective analysis as well as inference about the partition. To evaluate the model we carried out a simulation experiment and applied it to two real-life data sets. In both applications, it was shown that the DGLM-PPM outperformed the conventional DGLM. We also observed that the change-point probabilities estimated for the time series in study were compatible with the estimates of other papers available in the literature. This work opens up several possibilities for futures works concerning new developments of the DGLM-PPM class. For instance, we can expand it to include observations in all of the distributions belonging to the Exponential Family. Besides that, the model might be improved by exploring new specifications for the prior block cohesion and discount factors. We also believe that including an intra-block stochastic evolution to the model can lead to a considerable gain in forecasting accuracy.

Appendix A

Proof of some results from chapter 4

A.1 Proof of results (4.2.2) and (4.2.2)

Let X be a Beta Random Variable with parameters α_t and β_t. That is:

$$p(X) = \frac{1}{B(\alpha_t, \beta_t)} X^{\alpha_t - 1}(1-X)^{\beta_{t-1}}$$

Using properties of the Exponential Family da Silva et al (2011) [13] showed that:

$$\mathsf{E}\left[log\left(\frac{X}{1-X}\right)\right] = \psi(\alpha_t) - \psi(\beta_t).$$

Now define $\mu_t = \frac{X}{1-X}$, then $X = \frac{\mu_t}{1+\mu_t}$ and $dX = \frac{d\mu_t}{(\mu_t+1)^2}$. Thus:

$$\begin{aligned}
\mathsf{E}\left[log\left(\frac{X}{1-X}\right)\right] &= \frac{\Gamma(\alpha_t+\beta_t)}{\Gamma(\alpha_t)\Gamma(\beta_t)} \int_0^1 log\left(\frac{X}{1-X}\right) X^{\alpha_t-1}(1-X)^{\beta_t-1} dX \\
&= \frac{\Gamma(\alpha_t+\beta_t)}{\Gamma(\alpha_t)\Gamma(\beta_t)} \int_0^\infty log(\mu_t) \left(\frac{\mu_t}{1+\mu_t}\right)^{\alpha_t-1} \left(1-\frac{\mu_t}{1+\mu_t}\right)^{\beta_t-1} \frac{d\mu_t}{(\mu_t+1)^2} \\
&= \frac{\Gamma(\alpha_t+\beta_t)}{\Gamma(\alpha_t)\Gamma(\beta_t)} \int_0^\infty log(\mu_t)\mu_t^{\alpha_t-1}(1+\mu_t)^{-\alpha_t-\beta_t} d\mu_t \\
&= \psi(\alpha_t) - \psi(\beta_t).
\end{aligned}$$

This proves relation (4.2.2). To find (4.2.2) it is suffice to repeat the same procedure using the following result also demonstrated by da Silva (2011):

$$\mathsf{VAR}\left[log\left(\frac{X}{1-X}\right)\right] = \psi'(\alpha_t) + \psi'(\beta_t)$$

A.2 Proof of results (4.24) and (4.25)

Result (4.24) :

$$\mathsf{E}[log(\mu_t) \mid \kappa, D_{t-1}] = \psi(\alpha_t) - \psi(\beta_t) - log(\kappa)$$

Proof. Let X be a Beta Prime random variable with parameters α_t e β_t, that is:

$$p(X \mid \alpha_t, \beta_t) = \frac{1}{B(\alpha_t, \beta_t)} \frac{X^{\alpha_t - 1}}{(1+X)^{\alpha_t + \beta_t}}$$

Now define $\mu_t = \frac{X}{\kappa}$. By the Jacobian transformation μ_t has the following distribution:

$$\begin{aligned} p(\mu_t \mid \kappa, \alpha_t, \beta_t) &= \frac{1}{B(\alpha_t, \beta_t)} \frac{(\kappa\mu_t)^{\alpha_t - 1}}{(1+\kappa\mu_t)^{\alpha_t + \beta_t}} \kappa \\ &= \frac{\kappa^{\mu_t}}{B(\alpha_t, \beta_t)} \frac{\mu_t^{\alpha_t - 1}}{(1+\kappa\mu_t)^{\alpha_t + \beta_t}} \end{aligned}$$

Therefore μ_t is distributed according to a generalized Beta of the second kind.

$$\mathsf{E}[\log(\mu_t \mid \kappa, \alpha_t, \beta_t)] = \mathsf{E}\left[\log\left(\frac{X}{\kappa}\right) \mid \alpha_t, \beta_t\right] = \mathsf{E}[\log(X)] - \mathsf{E}[\log(\kappa)]$$
$$= \psi(\alpha_t) - \psi(\beta_t) - \log(\kappa)$$

□

Result (4.25) :

$$\mathsf{VAR}[\log(\mu_t) \mid \alpha_t, \beta_t, \kappa] = \psi'(\alpha_t) + \psi'(\beta_t)$$

Proof.

$$\begin{aligned} var[log(\mu_t) \mid \alpha_t, \beta_t, \kappa] &= var[log(X\kappa^{-1}) \mid \alpha_t, \beta_t] \\ &= var[log(X) \mid \alpha_t, \beta_t] \\ &= \psi'(\alpha_t) + \psi'(\beta_t) \end{aligned}$$

□

A.3 Proof of Expressions (4.35) and (4.36)

Result (4.35):

$$\mathsf{E}[log(\mu_t) \mid D_t) = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{1}{y_t + \kappa^{-1}} \frac{1}{B(y_t + 1, \kappa^{-1})} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)}$$
$$\times \left[\psi(\alpha_t + y_t) - \psi(\beta_t + \kappa^{-1}) - \log(\kappa)\right] \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa.$$

Proof. Using the posterior distribution for μ_t it is possible to write:

$$\mathsf{E}[log(\mu_t \mid D_t) = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty log(\mu_t) \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{\kappa^{\alpha_t + y_t}}{B(\alpha_t, \beta_t)} \frac{\mu_t^{\alpha_t + y_t - 1}}{(1+\kappa\mu_t)^{y_t + \kappa^{-1} + \alpha_t + \beta_t}} \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa d\mu_t$$

Notice that it is possible to invoke Fubini's Theoerem to invert the order of integration on the expression above. This procedure will allow the integration over μ_t to be performed. This yields:

$$\mathsf{E}[log(\mu_t \mid D_t) = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{\kappa^{\alpha_t + y_t}}{B(\alpha_t, \beta_t)} \frac{s^r \kappa^{r-1} \exp(-sk)}{\Gamma(r)} \left[\int_0^\infty log(\mu_t) \frac{\mu_t^{\alpha_t + y_t - 1}}{(1 + \kappa\mu_t)^{y_t + \kappa^{-1} + \alpha_t + \beta_t}} d\mu_t \right] d\kappa$$

$$\mathsf{E}[log(\mu_t \mid D_t) = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{\kappa^{\alpha_t + y_t}}{B(\alpha_t, \beta_t)} \frac{s^r \kappa^{r-1} \exp(-sk)}{\Gamma(r)} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{\kappa^{\alpha_t + y_t}} (\psi(\alpha_t + y_t) - \psi(\beta_t + \kappa^{-1}) - log(\kappa)) \; d\kappa$$

$$\mathsf{E}[log(\mu_t \mid D_t) = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)} \psi(\alpha_t + y_t) - \psi(\beta_t + \kappa^{-1}) - log(\kappa) \frac{s^r \kappa^{r-1} \exp(-sk)}{\Gamma(r)} d\kappa$$

□

Result (4.36):

$$\begin{aligned}\mathsf{E}[log^2(\mu_t) \mid D_t] = & \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{1}{y_t + \kappa^{-1}} \frac{1}{B(y_t + 1, \kappa^{-1})} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)} \\ & \times \; (\psi(\alpha_t + y_t) - \psi(\beta_t + \kappa^{-1}) - log(\kappa))^2 + \psi'(\alpha_t + y_t) + \psi'(\beta_t + \kappa^{-1}) \\ & \times \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa.\end{aligned}$$

Proof. Again, using the posterior for μ_t and Fubini's theorem we have that:

$$E[log^2(\mu_t) \mid D_t] = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty log^2(\mu_t) \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{\kappa^{\alpha_t + y_t}}{B(\alpha_t, \beta_t)} \frac{\mu_t^{\alpha_t + y_t - 1}}{(1 + \kappa\mu_t)^{y_t + \kappa^{-1} + \alpha_t + \beta_t}} \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa d\mu_t$$

$$E[log^2(\mu_t) \mid D_t] = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{\kappa^{\alpha_t + y_t}}{B(\alpha_t, \beta_t)} \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) \left[\int_0^\infty log^2(\mu_t) \frac{\mu_t^{\alpha_t + y_t - 1}}{(1 + \kappa\mu_t)^{y_t + \kappa^{-1} + \alpha_t + \beta_t}} d\mu_t \right] d\kappa$$

$$E[log^2(\mu_t) \mid D_t] = \frac{1}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{(y_t + \kappa^{-1})^{-1}}{B(y_t + 1, \kappa^{-1})} \frac{B(\alpha_t + y_t, \beta_t + \kappa^{-1})}{B(\alpha_t, \beta_t)} (\psi(\alpha_t + y_t) - \psi(\beta_t + \kappa^{-1}) - log(\kappa))^2 + \psi'(\alpha_t + y_t) + \psi'(\beta_t + \kappa^{-1}) \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa$$

□

Appendix B

Dynamic Bayesian Models for Positive Asymmetric Data

The framework introduced in Chapter for modeling time series of counts using the NB distribution is very general and can be extended to other distributions. In this appendix, we will present extensions of the Dynamic Bayesian Models for observations with positive response. This kind of data is very common in many fields such as economics, survival analysis, reliability among others. In the GLM framework, non-correlated positive observations are, usually, modelled by a Gamma Regression. A natural extension of the Gamma Regression is the Gamma DGLM (see for example Lindsay & Lambert, 1995). Other distributions, such as the Weibull, Log-Normal and Inverse Gaussian are also adequate choices for modeling and predicting time series when the observations are restricted to R^+. In the next sections we will work out full dynamic Bayesian models with Gamma and Weibull responses. The inference procedure for these models can be replicated with few modifications for any other observation equation within the EF.

B.1 Gamma Bayesian Dynamic Model

B.1.1 Basic Structure of the Model

Let $y_t, t = 1, \cdots, T$ be a time series in discrete time such that $y_t \in \mathsf{R}^+$. The Dynamic Bayesian Gamma Model (DBGM) has the following structure:

- **Observation equation:**

$$p(y_t \mid, \mu_t, \kappa) = \frac{1}{\Gamma(\kappa)} \left(\frac{\kappa y_t}{\mu_t}\right)^{\kappa} exp\left(-\frac{\kappa y_t}{\mu_t}\right) \frac{1}{y_t}. \tag{B.1}$$

- **Link Function:**

$$\lambda_t = g(\mu_t) = \mathbf{F_t}'\theta_\mathbf{t} = \log(\mu_t). \tag{B.2}$$

- **Prior for μ_t:**

$$p(\mu_t \mid D_{t-1}) = \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \mu_t^{\alpha_t - 1} \exp\left[-\frac{\beta_t}{\mu_t}\right]. \tag{B.3}$$

- **Evolution equation:**

$$\boldsymbol{theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim [0, \mathbf{W}_t]. \tag{B.4}$$

- **Initial Information:**

$$(\boldsymbol{\theta}_0 \mid D_0) \sim [\mathbf{m}_0, \mathbf{C}_0]. \tag{B.5}$$

- **Prior for κ:**

$$p(\kappa \mid r, s) = \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk). \tag{B.6}$$

As usual, the parametrization of the observation was chosen so that $\mathsf{E}(y_t \mid \mu_t, \kappa) = \mu_t$. Also, it is easy to prove that $\mathsf{VAR}(y_t \mid \mu_t, \kappa) = \mu_t^2$. The density in (B.1) belongs to the EF with $\phi_t = \kappa$, $\eta_t = -\frac{1}{\mu_t}$, $c(y_t, \phi_t) = \kappa \log(\kappa) + \kappa \log(y_t) - \log(y_t) - \log(\Gamma(\kappa))$ and $a(\eta_t) = \log(\mu_t)$. The prior distribution (B.3) for the mean μ_t of the process is an Inverse Gamma distribution that conjugates with the Gamma observation equation. The prior for the shape parameter is a Gamma distribution with parameters r and s. This choice was motivated by the flexibility of the Gamma distribution but any other distribution with support on R^+ would be adequate. As already shown, the classical estimation of the model can be thought of as a particular case of the full Bayesian specification when the prior for κ is uniform. For this reason, from here on, we will omit the inference via Maximum Likelihood Estimation (or MAP). Prior distributions for the state vector θ_t and the linear predictor λ_t are given by (3.9) and (3.10) respectively. To solve f_t and q_t in terms of α_t and β_t we must, first, find $\mathsf{E}[\log(\mu_t)]$ and $\mathsf{VAR}[\log(\mu_t)]$. Since the distribution of the logarithm of an Inverse Gamma Random Variable is identical to the distribution of the negative of the logarithm of a Gamma random variable all we need are the relations below:

$$\mathsf{E}[-X] = -\mathsf{E}[X],$$
$$\mathsf{VAR}[-X] = \mathsf{VAR}[X],$$

where X follows the Gamma density. Since the moments of the logarithm of a Gamma random variable are well known we can write:

$$f_t = \mathsf{E}(\mu_t) = -[\psi(\alpha_t) - \log(\beta_t)] = \log(\beta_t) - \psi(\alpha_t),$$

$$q_t = \mathsf{VAR}[\mu_t] = \psi^{'}(\alpha_t),$$

Now, applying the first order Taylor approximations to the digamma and trigamma functions, we get:

$$\alpha_t = \frac{1}{q_t},$$

$$\beta_t = \frac{\exp(f_t)}{q_t},$$

B.1.2 Predictive

The predictive distribution $p(y_t \mid D_{t-1})$ can be obtained marginally from:

$$\begin{aligned}
p(y_t \mid D_{t-1}) &= \int_0^\infty \int_0^\infty p(y_t \mid D_{t-1}, \kappa, \mu_t) p(\mu_t \mid D_{t-1}) p(\kappa \mid r, s) d\mu_t d\kappa \\
&= \int_0^\infty \int_0^\infty \frac{1}{\Gamma(\kappa)} \left(\frac{\kappa y_t}{\mu_t}\right)^{\kappa} \exp\left(-\frac{\kappa y_t}{\mu_t}\right) \frac{1}{y_t} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \mu_t^{\alpha_t - 1} \exp\left(-\frac{\beta_t}{\mu_t}\right) \\
&\times \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) \\
&= \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \frac{1}{y_t} \int_0^\infty \frac{\kappa^\kappa \kappa^{r-1} y_t^\kappa}{\Gamma(\kappa)} \exp(-s\kappa) \left[\int_0^\infty \frac{1}{\mu_t^{\kappa+\alpha_t+1}} \exp\left(-\frac{\kappa y_t + \beta_t}{\mu_t}\right) d\mu_t \right] d\kappa \\
&= \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \int_0^\infty \frac{\kappa^\kappa \kappa^{r-1} y_t^{\kappa-1}}{\Gamma(\kappa)} \exp(-s\kappa) \frac{\Gamma(\alpha_t + \kappa)}{(\kappa y_t + \beta_t)^{(\alpha_t+\kappa)}} d\kappa \\
&= \beta_t^\alpha \int_0^\infty \frac{\kappa^\kappa y_t^{\kappa-1}}{B(\alpha_t, \kappa)(\kappa y_t + \beta_t)^{(\alpha_t+\kappa)}} \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa
\end{aligned} \tag{B.7}$$

A point estimator for the one-step-ahead prediction of the process is given by the expectation of the predictive. Using the Law of Total Probability it is easy to prove that: $\hat{y}_t \mid D_{t-1} = \mathsf{E}[y_t \mid D_{t-1}] = \frac{\beta_t}{\alpha_t - 1}$ for $\alpha_t > 1$. In the context of positive continuous data, however, it is preferable, sometimes, to use the predictive median as point estimator for the predictions.

B.1.3 Updating of μ_t

Given a new observation y_t, the mean of the process can be update using Bayes's theorem,

$$\begin{aligned}
p(\mu_t, \kappa \mid D_t) &= \frac{p(y_t \mid \mu_t, \kappa, D_{t-1}) p(\mu_t \mid D_{t-1}) p(\kappa)}{p(y_t \mid D_{t-1})} \\
&= \frac{1}{p(y_t \mid D_{t-1})} \frac{1}{\Gamma(\kappa)} \left(\frac{\kappa y_t}{\mu_t}\right)^{\kappa} exp\left(-\frac{\kappa y_t}{\mu_t}\right) \frac{\beta_t^{\alpha_t}}{y_t \Gamma(\alpha_t)} \mu_t^{\alpha_t - 1} exp\left(-\frac{\beta_t}{\mu_t}\right) \\
&\times \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) \\
&= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \frac{y_t^{\kappa-1}}{\Gamma(\kappa)} \frac{\kappa^\kappa \kappa^{r-1}}{\mu_t^{\kappa+\alpha_t+1}} exp\left(-\frac{\kappa y_t + \beta_t}{\mu_t}\right) \exp(-s\kappa).
\end{aligned}$$

Thus, integrating over κ we obtain the posterior of μ_t given the information up to the instant t:

$$p(\mu_t \mid D_t) = \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \int_0^\infty \frac{y_t^{\kappa-1}}{\Gamma(\kappa)} \frac{\kappa^\kappa \kappa^{r-1}}{\mu_t^{\kappa+\alpha_t+1}} \exp\left(-\frac{\kappa y_t + \beta_t}{\mu_t}\right) \exp(-s\kappa) d\kappa. \tag{B.8}$$

B.1.4 Updating of λ_t

By definition, the posterior moments of the linear predictor are given by equation (4.34) where $p(\mu_t \mid D_t)$ is represented by expression (B.8), that is:

$$\begin{aligned} \mathsf{E}[\log(\mu_t)^n \mid D_t] = \frac{s^r \beta_t^{\alpha_t}}{p(y_t \mid D_{t-1})\Gamma(r)\Gamma(\alpha_t)} \int_0^\infty [\log(\mu_t)]^n \\ \times \int_0^\infty \frac{y_t^{\kappa-1}}{\Gamma(\kappa)} \frac{\kappa^{\kappa+r-1}}{\mu_t^{\kappa+\alpha_t+1}} \exp\left(-\frac{\kappa y_t + \beta_t + s\kappa}{\mu_t}\right) d\kappa d\mu_t. \end{aligned}$$

Thus, setting $n = 1$ and using the Fubini theorem to invert the order of integration, we have:

$$\begin{aligned} f_t^* &= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \int_0^\infty \frac{y_t^{\kappa-1} \kappa^\kappa \kappa^{r-1} \exp(-s\kappa)}{\Gamma(\kappa)} \left[\int_0^\infty \frac{\log(\mu_t)}{\mu_t^{\kappa+\alpha_t+1}} \exp\left(-\frac{(\kappa y_t + \beta_t)}{\mu_t}\right) d\mu_t \right] d\kappa \\ &= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \int_0^\infty \frac{y_t^{\kappa-1} \kappa^\kappa \kappa^{r-1} \exp(-s\kappa)}{\Gamma(\kappa)} \frac{\Gamma(\kappa+\alpha_t)}{(\kappa y_t + \beta_t)^{\alpha_t+\kappa}} [\log(\kappa y_t + \beta_t) - \psi(\alpha_t + \kappa)] d\kappa \\ &= \frac{\beta_t^{\alpha_t}}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{y_t^{\kappa-1} \kappa^\kappa}{(\kappa y_t + \beta_t)^{\alpha_t+\kappa} B(\alpha_t, \kappa)} [\log(\kappa y_t + \beta_t) - \psi(\alpha_t + \kappa)] \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa \\ &= \frac{\beta_t^{\alpha_t}}{p(y_t \mid D_{t-1})} \int_0^\infty g(\kappa) f(\kappa) d\kappa. \end{aligned} \tag{B.9}$$

Proceeding in a similar fashion for $n = 2$ we get

$$\begin{aligned} \mathsf{E}[(\log(\mu_t))^2] &= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \int_0^\infty \frac{y_t^\kappa \kappa^{\kappa+r-1} \exp(-s\kappa)}{y_t \Gamma(\kappa)} \\ &\times \left[\int_0^\infty \frac{(\log(\mu_t))^2}{\mu_t^{\kappa+\alpha_t+1}} \exp\left(-\frac{(\kappa y_t + \beta_t)}{\mu_t}\right) d\mu_t \right] d\kappa \\ &= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \int_0^\infty \frac{y_t^\kappa \kappa^{\kappa+r-1} \exp(-s\kappa)}{y_t \Gamma(\kappa)} \frac{\Gamma(\kappa+\alpha_t)}{(\kappa y_t + \beta_t)^{\alpha_t+\kappa}} [\psi'(\alpha_t) + log^2 \beta_t \\ &- 2\psi(\alpha_t) log(\beta_t) + \psi^2(\alpha_t)] d\kappa \\ &= \frac{\beta_t^{\alpha_t}}{p(y_t \mid D_{t-1})} \int_0^\infty \frac{y t^{\kappa-1} \kappa^\kappa}{(\kappa y_t + \beta_t)^{\alpha_t+\kappa} B(\alpha_t, \kappa)} [\psi'(\alpha_t) + log^2 \beta_t - 2\psi(\alpha_t) log(\beta_t) + \psi^2(\alpha_t)] \times \\ &\frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-s\kappa) d\kappa \\ &= \frac{\beta_t^{\alpha_t}}{p(y_t \mid D_{t-1})} \int_0^\infty g(\kappa) f(\kappa) d\kappa. \end{aligned} \tag{B.10}$$

Equations (B.7), (B.9) and (B.10) may be solved by numerical or Monte Carlo integration techniques. In this work we verified that the later is a more reliable method specially for non-stationary time series.

B.1.5 Updating of the state vector

The updating of $\boldsymbol{\theta}_t$ can be done using the LBE approximation through the following equations:

$$\mathbf{m}_t = \mathbf{a}_t + \frac{1}{q_t}\mathbf{R}_t\mathbf{F}_t(f_t^* - f_t),$$

$$\mathbf{C}_t = \mathbf{R}_t - \frac{1}{q_t}\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t\left(1 - \frac{q_t^*}{q_t}\right).$$

B.1.6 Estimation of κ

The posterior distribution for κ given the whole set of information D_T can be calculated from Bayes's theorem as:

$$\begin{aligned} p(\kappa \mid D_T) &= \frac{\prod_{t=1}^{T} p(y_t \mid D_{t-1}, \kappa)\, p(\kappa)}{\int_0^\infty \prod_{t=1}^{T} p(y_t \mid D_{t-1}, \kappa)\, p(\kappa) d\kappa} \\ &= \frac{\prod_{t=1}^{T} \frac{\beta^{\alpha_t}}{B(\alpha_t,\kappa)} \frac{\kappa^\kappa y_t^{\kappa-1}}{(\kappa y + t + \beta_t)^{\alpha_t+\kappa}} \frac{s^r}{\Gamma(r)}\kappa^{r-1}\exp(-s\kappa)}{\int_0^\infty \prod_{t=1}^{T} \frac{\beta^{\alpha_t}}{B(\alpha_t,\kappa)} \frac{\kappa^\kappa y_t^{\kappa-1}}{(\kappa y + t + \beta_t)^{\alpha_t+\kappa}} \frac{s^r}{\Gamma(r)}\kappa^{r-1}\exp(-s\kappa) d\kappa}. \end{aligned} \tag{B.11}$$

Under quadratic loss, the Bayesian estimator is the posterior mean, that is: $\mathsf{E}(\kappa \mid D_T)$. Since the integrals involved in the calculation of the expected value have no closed-form we propose to generate samples from (B.11) using the ARMS algorithm in the same way we did for the DBNBM. Another possibility would be to use the posterior mode as the Bayesian Estimator. In this case, the maximum of the posterior distribution can be determined with the help of some optimization algorithm such as the Nelder-Mead. The big advantage of the ARMS procedure over the mode estimator is that interval estimates are readily available from the samples obtained.

B.2 Weibull Bayesian Dynamic Model

B.2.1 Basic Structure of the Model

The Weibull Bayesian Dynamic Model is defined by the following equations:

- **Observation Equation**

  The Weibull distribution is widely used in many fields such as survival analysis, reliability, engineering meteorology, among others. The probability density function of a Weibull random variable y_t is given by:

$$p(y_t \mid \kappa, \lambda_t) = \frac{\kappa}{\lambda_t} \left(\frac{y_t}{\lambda_t}\right)^{\kappa-1} \exp\left(-\left(\frac{y_t}{\lambda_t}\right)^{\kappa}\right), \qquad y_t > 0,$$

where

$$\mathsf{E}[y_t \mid \kappa, \lambda_t] = \lambda_t \Gamma\left(1 + \frac{1}{\kappa}\right) \quad \text{and} \quad \mathsf{VAR}[y_t \mid \lambda_t, \kappa] = \lambda_t^2 \left[\Gamma\left(1 + \frac{2}{\kappa}\right) - \Gamma\left(1 + \frac{1}{\kappa}\right)^2\right].$$

Defining $\lambda_t = \frac{\mu_t}{\Gamma(1+\frac{1}{\kappa})}$ we can rewrite the equation above as:

$$p(y_t \mid \mu_t, \kappa) = \frac{\kappa}{y_t} \left(\frac{y_t \Gamma(1+\frac{1}{\kappa})}{\mu_t}\right)^{\kappa} \exp\left[-\left(\frac{y_t \Gamma(1+\frac{1}{\kappa})}{\mu_t}\right)^{\kappa}\right], \tag{B.12}$$

with $\mathsf{E}[y_t \mid \kappa, \mu_t] = \mu_t$.

- **Link Function:**

$$\lambda_t = g(\mu_t) = \mathbf{F}_t' \boldsymbol{\theta}_t = \log(\mu_t). \tag{B.13}$$

- **Prior for μ_t:**

$$p(\mu_t \mid \alpha_t, \beta_t, \kappa, D_{t-1}) = \kappa \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \mu_t^{-\kappa\alpha_t - 1} \exp\left[-\frac{\beta_t^{\frac{1}{\kappa}}}{\mu_t}\right]. \tag{B.14}$$

Expression (B.14) is a particular case of the Inverse Generalized Gamma as defined in Mead (2015) [90]. This choice for prior distribution is justified by the fact (B.14) is the natural conjugate of the observation equation (B.12).

- **Evolution equation:**

$$\boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \omega_t \sim [0, \mathbf{W}_t]. \tag{B.15}$$

- **Initial Information:**

$$(\theta_{\mathbf{0}} \mid D_0) \sim [\mathbf{m}_0, \mathbf{C}_0]. \tag{B.16}$$

- **Prior for κ:**

$$p(\kappa \mid r, s) = \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk). \tag{B.17}$$

The reason why we chose a Gamma prior distribution for the shape parameter is the same as in the DBGM. Again, we call attention to the fact that other distributions with support on the positive real line would also fit. Priors for the state vector and the linear predictor are defined as usual. To equate the parameters α_t and β_t the following relations were used:

$$f_t = \mathsf{E}\left[\log(\mu_t \mid D_{t-1})\right] = \frac{1}{\kappa}\left[\log(\beta_t) - \psi(\alpha_t)\right], \tag{B.18}$$

$$q_t = \mathsf{VAR}[\log(\mu_t \mid D_{t-1})] = \frac{1}{\kappa}\left[\log(\beta_t) - \psi(\alpha_t)\right], \tag{B.19}$$

Proof. The Generalized Inverse Gamma distribution as defined in Mead (2015) [90] is given by:

$$f(x;\alpha,\kappa,\lambda,\theta,\gamma) = \frac{\gamma\theta^{\alpha\gamma}}{\Gamma_\lambda(\alpha,\kappa)} x^{-\alpha\gamma-1} \left(\frac{\theta}{x}^{\gamma} + \kappa\right)^{-\lambda} \exp\left(-\frac{\theta}{x}^{\gamma}\right). \tag{B.20}$$

Taking $\lambda = 0$, (B.20) reduces to:

$$f(x;\alpha,\kappa,\theta,\gamma) = \frac{\gamma\theta^{\alpha\gamma}}{\Gamma(\alpha)} x^{-\alpha\gamma-1} \exp\left(-\frac{\theta}{x}^{\gamma}\right) \tag{B.21}$$

We are interested in the expectation and variance of $\log(x)$. From the definition of expected value we have:

$$\begin{aligned}
\mathsf{E}[\log(x)] &= \frac{\gamma\theta^{\alpha\gamma}}{\Gamma(\alpha)} \int_0^\infty log(x) x^{-\alpha\gamma-1} \exp\left(-\frac{\theta}{x}^{\gamma}\right) dx \\
&= \frac{\gamma\theta^{\alpha\gamma}}{\Gamma(\alpha)} \left(-\frac{1}{\gamma}\frac{\partial}{\partial\alpha} \int_0^\infty x^{-\alpha\gamma-1} \exp\left(-\frac{\theta}{x}^{\gamma}\right) dx\right) \\
&= \frac{\theta^{\alpha\gamma}}{\Gamma(\alpha)} \left(-\frac{\partial}{\partial\alpha} \frac{\Gamma(\alpha)}{\gamma\theta^{\alpha\gamma}}\right) \\
&= \frac{1}{\gamma}[\gamma log(\theta) - \psi(\alpha)]
\end{aligned}$$

$$\begin{aligned}
\mathsf{E}[\log^2(x)] &= \frac{\gamma\theta^{\alpha\gamma}}{\Gamma(\alpha)} \int_0^\infty log^2(x) x^{-\alpha\gamma-1} \exp\left(-\frac{\theta}{x}^{\gamma}\right) dx \\
&= \frac{\gamma\theta^{\alpha\gamma}}{\Gamma(\alpha)} \left(\frac{1}{\gamma^2}\frac{\partial^2}{(\partial\alpha)^2} \int_0^\infty x^{-\alpha\gamma-1} \exp\left(-\frac{\theta}{x}^{\gamma}\right) dx\right) \\
&= \frac{\theta^{\alpha\gamma}}{\gamma\Gamma(\alpha)} \left(\frac{\partial^2}{(\partial\alpha)^2} \frac{\Gamma(\alpha)}{\gamma\theta^{\alpha\gamma}}\right) \\
&= \frac{1}{\gamma^2}[(\gamma\log(\theta) - \psi(\alpha))^2 + \psi'(\alpha)].
\end{aligned}$$

Thus:

$$\begin{aligned}\mathsf{VAR}[\log(x)] &= \mathsf{E}[\log^2(x)] - \mathsf{E}[\log(x)]^2 \\ &= \frac{1}{\gamma^2}[(\gamma\log(\theta) - \psi(\alpha))^2 + \psi'(\alpha)] - \frac{1}{\gamma}(\gamma\log(\theta) - \psi(\alpha))^2 \\ &= \frac{1}{\gamma^2}\psi'(\alpha)\end{aligned}$$

Now observe that equation (B.14) has the same form of equation (B.20) with:

$$x = \mu_t, \qquad \gamma = \kappa, \qquad \alpha = \alpha_t, \qquad \theta = \beta^{\frac{1}{\kappa}}$$

So, the expected value and variance of $\log(\mu_t)$ are given, respectively, by:

$$f_t = \mathsf{E}[\log(\mu_t \mid D_{t-1})] = \frac{1}{\kappa}[\log(\beta_t) - \psi(\alpha_t)];$$

$$q_t = \frac{1}{\kappa^2}\psi'(\alpha_t).$$

□

With the help of the usual approximations for the digamma and trigamma functions we can solve for α_t and β_t to obtain:

$$\beta_t = \frac{\exp(\kappa f_t)}{\kappa^2 q_t} \qquad \text{and} \qquad \alpha_t = \frac{1}{\kappa^2 q_t}.$$

B.2.2 Predictive

The joint distribution of y_t and μ_t is given by:

$$p(y_t, \mu_t, \kappa \mid D_{t-1}) = p(y_t \mid \mu_t, D_{t-1}, \kappa)p(\mu_t \mid D_{t-1})p(\kappa).$$

Then, the marginal distribution $p(y_t \mid D_{t-1})$ can be obtained from:

$$
\begin{aligned}
p(y_t \mid D_{t-1}) &= \int_0^\infty \int_0^\infty p(y_t \mid \mu_t, D_{t-1}, \kappa) p(\mu_t \mid D_{t-1}) p(\kappa) d\mu_t d\kappa \\
&= \int_0^\infty \int_0^\infty \frac{\kappa}{y_t} \left(\frac{y_t \Gamma(1+\frac{1}{\kappa})}{\mu_t} \right)^\kappa \exp\left[-\left(\frac{y_t \Gamma(1+\frac{1}{\kappa})}{\mu_t} \right)^\kappa \right] \kappa \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \mu_t^{-\kappa\alpha_t - 1} \exp\left[-\frac{\beta_t^{\frac{1}{\kappa}}}{\mu_t} \right]^\kappa \right] \\
&\times \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) \\
&= \int_0^\infty \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \kappa^2 y_t^{\kappa-1} \Gamma\left(1+\frac{1}{\kappa}\right)^\kappa \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) \\
&\times \int_0^\infty \frac{1}{\mu_t^{\kappa\alpha_t+\kappa+1}} \exp\left[-\frac{y_t \Gamma\left(1+\frac{1}{\kappa}\right)^\kappa + \beta_t}{\mu_t^\kappa} \right] d\mu_t d\kappa \\
&= \int_0^\infty \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \kappa^2 y_t^{\kappa-1} \Gamma\left(1+\frac{1}{\kappa}\right)^\kappa \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) \frac{\Gamma(\alpha_t+1)}{\kappa[(y_t \Gamma\left(1+\frac{1}{\kappa}\right)^\kappa + \beta_t]^{\alpha_t+1}} d\kappa \\
&= \int_0^\infty \frac{\alpha_t}{y_t} \beta_t^{\alpha_t} \frac{\kappa}{(y_t \Gamma\left(1+\frac{1}{\kappa}\right)^\kappa + \beta_t)^{\alpha_t+1}} \frac{y_t \Gamma(1+\frac{1}{\kappa})^\kappa}{} \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) d\kappa.
\end{aligned}
\tag{B.22}
$$

A point estimator for the one-step-ahead prediction is the expected value of the predictive distribution. By repeated application of the law of total probability, we have:

$$
\begin{aligned}
\hat{y}_t \mid D_{t-1} &= \mathsf{E}[y_t \mid D_{t-1}] \\
&= \mathsf{E}[\mathsf{E}[\mathsf{E}[y_t \mid \mu_t, \kappa]] \mid D_{t-1}] \\
&= \mathsf{E}[\mathsf{E}[\mu_t \mid \kappa] \mid D_{t-1}] \\
&= \mathsf{E}\left[\frac{\beta_t^{\frac{1}{\kappa}} \Gamma(\alpha_t - \kappa^{-1})}{\Gamma(\alpha_t)} \mid D_{t-1} \right]
\end{aligned}
\tag{B.23}
$$

Remember that, since α_t and β_t are implicit functions of κ, there is no simple analytic solution for (B.23). It is easy, though, to obtain samples from this expression because κ has a known distribution. Another option would be to use the median or the mode as point estimators.

B.2.3 Updating of μ_t

From Bayes's theorem:

$$
\begin{aligned}
p(\mu_t, \kappa \mid D_t) &= \frac{p(y_t \mid \mu_t, \kappa, D_{t-1}) p(\mu_t \mid D_{t-1}) p(\kappa)}{p(y_t \mid D_{t-1})} \\
&= \frac{1}{p(y_t \mid D_{t-1})} \frac{\kappa}{y_t} \left(\frac{y_t \Gamma(1+\frac{1}{\kappa})}{\mu_t} \right)^{\kappa} \exp\left[- \left(\frac{y_t \Gamma(1+\frac{1}{\kappa})}{\mu_t} \right)^{\kappa} \right] \\
&\times \kappa \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \mu_t^{-\kappa\alpha_t - 1} \exp\left[- \frac{\beta_t^{\frac{1}{\kappa}}}{\mu_t} \right] \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) \\
&= \frac{1}{p(y_t \mid D_{t-1})} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \kappa^2 y_t^{\kappa - 1} \Gamma\left(1 + \frac{1}{\kappa}\right)^{\kappa} \frac{s^r}{\Gamma(r)} k^{r-1} \exp(-sk) \\
&\times \frac{1}{\mu_t^{\kappa\alpha_t + \kappa + 1}} \exp\left[- \frac{y_t^{\kappa} \Gamma\left(1 + \frac{1}{\kappa}\right)^{\kappa} + \beta_t}{\mu_t^{\kappa}} \right].
\end{aligned}
$$

Then, integrating over κ:

$$
\begin{aligned}
p(\mu_t \mid D_t) = \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \int_0^{\infty} & \frac{\beta_t^{\alpha_t}}{y_t \Gamma(\alpha_t)} \frac{y_t \Gamma\left(1+\frac{1}{\kappa}\right)^{\kappa}}{\mu_t^{\kappa\alpha_t + \kappa + 1}} \exp\left[- \frac{y_t^{\kappa} \Gamma\left(1+\frac{1}{\kappa}\right)^{\kappa} + \beta_t}{\mu_t^{\kappa}} \right] \\
& k^{r-1} \exp(-sk) . d\kappa
\end{aligned}
$$

B.2.4 Updating of the linear predictor and state vector

As usual, the updating of the linear predictor can be obtained through the following equations:

$$
f_t^* = \mathsf{E}[\lambda_t \mid D_t] = \mathsf{E}[\log(\mu_t) \mid D_t] \qquad \text{and} \qquad q_t^* = \mathsf{VAR}[\lambda_t \mid D_t] = \mathsf{VAR}[\log(\mu_t) \mid D_t],
$$

where:

$$
\begin{aligned}
\mathsf{E}[log(\mu_t) \mid D_t] = \frac{1}{p(y_t \mid D_{t-1})} \int_0^{\infty} & \frac{\alpha_t \beta_t^{\alpha_t}}{y_t} \left(\frac{y_t \Gamma\left(1+\frac{1}{\kappa}\right)^{\kappa}}{y_t \Gamma\left(1+\frac{1}{\kappa}\right)^{\kappa} + \beta_t} \right)^{\alpha_t + 1} \\
& \times \left[log\left(y_t \Gamma\left(1 + \frac{1}{\kappa}\right)^{\kappa} + \beta_t \right) - \psi(\alpha_t + 1) \right] \frac{s^r}{\Gamma(r)} \kappa^{r-1} \exp(-sk) d\kappa,
\end{aligned}
\tag{B.24}
$$

Proof. Since μ_t and κ belong to R^+ it is trivial to see that the integrand in (B.24) is non-negative. Thus, the conditions of Fubini-Tonelli theorem apply and we can write:

$$
\begin{aligned}
\mathsf{E}[\log(\mu_t) \mid D_t] &= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \int_0^\infty \int_0^\infty \log(\mu_t) \frac{\beta_t^{\alpha_t}}{y_t \Gamma(\alpha_t)} \frac{y_t \Gamma\left(1+\frac{1}{\kappa}\right)^{\kappa}}{\mu_t^{\kappa\alpha_t+\kappa+1}} \\
&\times \exp\left[-\frac{y_t^\kappa \Gamma\left(1+\frac{1}{\kappa}\right)^{\kappa} + \beta_t}{\mu_t^\kappa}\right] k^{r-1} \exp(-sk) d\kappa d\mu_t \\
&= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \int_0^\infty \frac{\beta_t^{\alpha_t}}{y_t \Gamma(\alpha_t)} \kappa^2 \left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right)^{\kappa} \kappa^{r-1} \exp(-sk) \\
&\times \left\{ \int_0^\infty \frac{log(\mu_t)}{\mu_t^{\kappa\alpha_t+\kappa+1}} \exp\left[-\frac{y_t^\kappa \Gamma\left(1+\frac{1}{\kappa}\right)^{\kappa} + \beta_t}{\mu_t^\kappa}\right] d\mu_t \right\} d\kappa \\
&= \frac{1}{p(y_t \mid D_{t-1})} \frac{s^r}{\Gamma(r)} \int_0^\infty \frac{\beta_t^{\alpha_t}}{y_t \Gamma(\alpha_t)} \frac{\kappa^2 \left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right)^{\kappa} \Gamma(\alpha_t+1)}{\kappa \left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right)^{\kappa} + \beta_t)^{\alpha_t+1}} \\
&\times \frac{1}{\kappa} \left[log\left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right) + \beta_t \right] - \psi(\alpha_t+1) \\
&\times \kappa^{r-1} \exp(-sk) d\kappa \\
&= \int_0^\infty \frac{\alpha_t \beta_t^{\alpha_t}}{y_t} \frac{\left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right)^{\kappa}}{\kappa \left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right)^{\kappa} + \beta_t)^{\alpha_t+1}} \\
&\times \left[log\left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right) + \beta_t \right] - \psi(\alpha_t+1) \frac{s^r}{\Gamma(r)} \kappa^{r-1} \\
&\times \exp(-sk) d\kappa
\end{aligned}
$$

□

and

$$
\begin{aligned}
\mathsf{E}[\log^2(\mu_t)] &= \int_0^\infty \frac{\alpha_t \beta_t^{\alpha_t}}{y_t} \frac{\left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right)^{\kappa}}{\kappa \left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right)^{\kappa} + \beta_t)^{\alpha_t+1}} \\
&\times \left\{ \left[log\left(y_t \Gamma\left(1+\frac{1}{\kappa}\right)\right) + \beta_t \right] - \psi(\alpha_t+1) \right\}^2 + \psi'(\alpha_t+1) \\
&\times \frac{s^r}{\Gamma(r)} \kappa^{r-1} exp(-sk) d\kappa.
\end{aligned}
\tag{B.25}
$$

Proof.

$$
\begin{aligned}
\mathsf{E}[\log^2(\mu_t)] &= \frac{1}{p(y_t \mid D_{t-1})}\frac{s^r}{\Gamma(r)}\int_0^\infty\int_0^\infty \log^2(\mu_t)\frac{\beta_t^{\alpha_t}}{y_t\Gamma(\alpha_t)}\frac{\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa}{\mu_t^{\kappa\alpha_t+\kappa+1}} \\
&\quad\times\exp\left[-\frac{y_t^\kappa\Gamma\left(1+\frac{1}{\kappa}\right)^\kappa+\beta_t}{\mu_t^\kappa}\right]k^{r-1}\exp(-sk)d\kappa d\mu_t \\
&= \frac{1}{p(y_t \mid D_{t-1})}\frac{s^r}{\Gamma(r)}\int_0^\infty \frac{\beta_t^{\alpha_t}}{y_t\Gamma(\alpha_t)}\frac{\kappa^2}{\kappa}\frac{\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa\Gamma(\alpha_t+1)}{\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa+\beta_t)^{\alpha_t+1}} \\
&\quad\times\frac{1}{\kappa^2}\left\{log\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)+\beta_t\right)-\psi(\alpha_t+1)\right\}^2+\psi'(\alpha_t+1)\Bigg\} \\
&\quad\times\kappa^{r-1}\exp(-sk)d\kappa \\
&= \frac{1}{p(y_t \mid D_{t-1})}\int_0^\infty \frac{\alpha_t\beta_t^{\alpha_t}}{y_t}\frac{\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa}{\kappa\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa+\beta_t)^{\alpha_t+1}} \\
&\quad\times\left\{log\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)+\beta_t\right)-\psi(\alpha_t+1)\right\}^2+\psi'(\alpha_t+1)\Bigg\} \\
&\quad\times\frac{s^r}{\Gamma(r)}\kappa^{r-1}\exp(-sk)d\kappa.
\end{aligned}
$$

□

Naturally, q_t can be obtained from $\mathsf{VAR}[\log(\mu_t) \mid D_t] = \mathsf{E}[\log^2(\mu_t)] - (\mathsf{E}[log(\mu_t) \mid D_t])^2$. Updating of the state vector is done in the usual sense through Equations (3.13) and (3.14). The integrals in (B.24), (B.25) and (B.22) can be solved via numerical or Monte Carlo integration.

B.2.5 Estimation of κ

Samples of the posterior distribution of κ can be obtained via ARMS from the following target distribution:

$$
\begin{aligned}
p(\kappa \mid D_T) &= \frac{\prod_{t=1}^T p(y_t \mid D_{t-1},\kappa)\,p(\kappa)}{\int_0^\infty \prod_{t=1}^T p(y_t \mid D_{t-1},\kappa)\,p(\kappa)d\kappa} \\
&= \frac{\prod_{t=1}^T \frac{\kappa}{y_t}\alpha_t\frac{\beta_t^{\alpha_t}\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa}{[(y_t\Gamma(1+\frac{1}{\kappa}))^\kappa+\beta_t]^{\alpha_t+1}}\frac{s^r}{\Gamma(r)}\kappa^{r-1}\exp(-s\kappa)}{\int_0^\infty \prod_{t=1}^T \frac{\kappa}{y_t}\alpha_t\frac{\beta_t^{\alpha_t}\left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa}{[(y_t\Gamma(1+\frac{1}{\kappa}))^\kappa+\beta_t]^{\alpha_t+1}}\frac{s^r}{\Gamma(r)}\kappa^{r-1}\exp(-s\kappa)d\kappa},
\end{aligned}
\tag{B.26}
$$

where

$$
\begin{aligned}
p(y_t \mid \kappa, D_{t-1}) &= \int_0^\infty p(y_t \mid \kappa, \mu_t, D_{t-1}) p(\mu_t \mid D_{t-1}) d\mu_t \\
&= \kappa^2 \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} y_t^{\kappa-1} \Gamma\left(1+\frac{1}{\kappa}\right) \int_0^\infty \frac{1}{\mu_t^{\kappa\alpha_t+\kappa+1}} \exp\left[-\frac{(y_t\Gamma(1+\frac{1}{\kappa}))^\kappa + \beta_t}{\mu_t^\kappa}\right] d\mu_t \\
&= \frac{\kappa^2}{y_t} \frac{\beta_t^{\alpha_t}}{\Gamma(\alpha_t)} \left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa \frac{\Gamma(\alpha_t+1)}{\kappa[(y_t\Gamma(1+\frac{1}{\kappa}))^\kappa + \beta_t]^{\alpha_t+1}} \\
&= \frac{\kappa}{y_t} \alpha_t \frac{\beta_t^{\alpha_t} \left(y_t\Gamma\left(1+\frac{1}{\kappa}\right)\right)^\kappa}{[(y_t\Gamma(1+\frac{1}{\kappa}))^\kappa + \beta_t]^{\alpha_t+1}}.
\end{aligned}
$$

Again, we suggest to sample from (B.26) via ARMS. The specification of the variance $\mathbf{W}_t$, in this case, must be done with the help of discount factors.

Bibliography

[1] J. J.F. Commandeur and S. J. Koopman. *An Introduction to State Space Time Series Analysis*. Oxford University Press, 2007.

[2] A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1990.

[3] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics, 1997.

[4] Harrison P. J. West, M. and H. S. Migon. Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83, 1985.

[5] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[6] D. Gamerman and M. West. An application of dynamic survival models in unemployment studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 36(2/3):269–274, 1987.

[7] J.K. Lindsey and P. Lambert. Dynamic generalized linear models and repeated measurements. *Journal of Statistical Planning and Inference*, 47(1):129 – 139, 1995.

[8] M. Chiogna and C. Gaetan. Dynamic generalized linear models with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):453–468, 2002.

[9] L. Fahrmeir. Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87(418):501–509, 1992.

[10] Godsill S. Doucet, A. and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[11] K. Triantafyllopoulos. Inference of dynamic generalized linear models: On-line computation and appraisal. *International Statistical Review / Revue Internationale de Statistique*, 77(3):430–450, 2009.

[12] D. Gamerman. Markov chain monte carlo for dynamic generalised linear models. *Biometrika*, 85(1):215–227, 1998.

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

[14] C.Q. Da-Silva and G.S. Rodrigues. Bayesian dynamic dirichlet models. *Communications in Statistics - Simulation and Computation*, 44(3):787–818, 2015.

[15] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 2001.

[16] B. Kedem and K. Fokianos. *Regression Models for Time Series Analysis*. Springer, 2002.

[17] A. E. Raftery and V. E. Akman. Bayesian analysis of a poisson process with a change-point. *Biometrika*, 73(1):85–89, 1986.

[18] J. Whittaker and S. Fruhwirth-Schnatter. A dynamic changepoint model for detecting the onset of growth in bacteriological infections. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(4):625–640, 1994.

[19] P.H.D. da Silva and C.Q. da Silva. Dynamic changepoint detection in count time series: a particle filter approach. *Journal of Statistical Computation and Simulation*, 87(1):42–68, 2017.

[20] J.A. Hartigan. Partition models. *Communications in Statistics - Theory and Methods*, 19(8):2745–2756, 1990.

[21] D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, 20(1):260–279, 1992.

[22] J.A. Almeida. Generalized linear model via product partition model. Master's thesis, Universidade Federal de Minas Gerais, 2016.

[23] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995.

[24] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[25] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992.

[26] W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):455–472, 1995.

[27] S. I. Resnick. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Birkhäuser Basel, 2014.

[28] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[29] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.

[30] B. Oksendal. *Stochastic Differential Equations (3rd Ed.): An Introduction with Applications*. Springer-Verlag, Berlin, Heidelberg, 1992.

[31] Gudmundsson G. Lindgren G. Bondesson L. Harsaae E. Laake P. Juselius K. Lauritzen S. L. Cox, D. R. Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 8(2):93–115, 1981.

[32] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327, 1986.

[33] B. Kedem and K. Fokianos. *Regression Models for Time Series Analysis*. Hoboken, NJ: Wiley, 2002.

[34] R. A. Davis, W. T. M. Dunsmuir, and S. B. Streett Streett. Observation-driven models for poisson counts. *Biometrika*, 90(4):777–790, 2003.

[35] A. Heinen. Modelling time series count data: An autoregressive conditional poisson model. *CORE Discussion Paper*, 62, 2003.

[36] K. Fokianos. Some recent progress in count time series. *Statistics*, 45(1):49–58, 2011.

[37] T. Liboschik, K. Fokianos, and R Fried. tscount: An r package for analysis of count time series following generalized linear models. *Journal of Statistical Software, Articles*, 82(5):1–51, 2017.

[38] Ed. McKenzie. Some simple models for discrete variate time series1. *JAWRA Journal of the American Water Resources Association*, 21(4):645–650, 1985.

[39] M. A. Al-Osh and A. A. Alzaid. First-order integer-valued autoregressive (inar(1) process). *Journal of Time Series Analysis*, 8(3):261–275, 1987.

[40] K. Brannas. Explanatory variables in the ar(1) count data model. *Umea Economic Studies*, (381), 1995.

[41] M. A. Jazi, G. Jones, and C. Lai. Integer valued ar(1) with geometric innovations. *Journal of the Iranian Statistical Society*, 11(2), 2012.

[42] R. Ferland, A. Latour, and D. Oraichi. Integer-valued garch process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

[43] W. Barreto-Souza. Zero-modified geometric inar(1) process for modelling count time series with deflation or inflation of zeros. *Journal of Time Series Analysis*, 36(6):839–852, 2015.

[44] J. Durbin and S.J. Koopman. *Time series analysis by state space methods*. OUP Oxford, 2012.

[45] S. L. Zeger. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988.

[46] K. Brännäs and P. Johansson. Time series count data regression. *Communications in Statistics - Theory and Methods*, 23(10):2907–2925, 1994.

[47] M. J. Campbell. Time series regression for counts: An investigation into the relationship between sudden infant death syndrome and environmental temperature. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(2):191–208, 1994.

[48] K. S. Chan and J. Ledolter. Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252, 1995.

[49] R. A. Davis, W. T. M. Dunsmuir, and Y. Wang. On autocorrelation in a poisson regression model. *Biometrika*, 87(3):491–505, 2000.

[50] J. Durbin and S. J. Koopman. Monte carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, 84(3):669–684, 1997.

[51] D. Gamerman, T. R. Santos, and G. C. Franco. A non-gaussian family of state-space models with exact marginal likelihood. *Journal of Time Series Analysis*, 34(6):625–645, 2013.

[52] A. C. Harvey and C. Fernandes. Time series models for count or qualitative observations. *Journal of Business & Economic Statistics*, 7(4):407–417, 1989.

[53] T. Aktekin, R. Soyer, and F. Xu. Assessment of mortgage default risk via bayesian state space models. *The Annals of Applied Statistics*, 7(3):1450–1473, 2013.

[54] T. Aktekin, N. Polson, and R. Soyer. Sequential bayesian analysis of multivariate count data. *Bayesian Analysis*, 13(2):385–409, 2018.

[55] R. Soyer. Kalman filtering and sequential bayesian analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1438, 2018.

[56] M.A.O. Souza, H.S. Migon, and J.B.M. Pereira. Extended dynamic generalized linear models: The two-parameter exponential family. *Computational Statistics & Data Analysis*, 121(C):164–179, 2018.

[57] F. Castellares, S.L.P. Ferrari, and A.J. Lemonte. On the bell distribution and its associated regression model for count data. *Applied Mathematical Modelling*, 56:172 – 185, 2018.

[58] E. Borel. Sur l'emploi du theoreme de bernoulli pour faciliter le calcul d'un infinite de coefficients. application au probleme de l' attente a un guichet. *Comptes Rendus, Acedemic des Sciences, Series A*, 214:452 – 456, 1942.

[59] Vazquez-Polo F. Gomez-Deniz, E. and V. Garcia. The modified borel–tanner (mbt) regression model. *REVSTAT – Statistical Journal*, 15(3):425–442, 2017.

[60] P. Puig and J. Valero. Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, 101(473):332–340, 2006.

[61] M. Sankaran. 275. note: The discrete poisson-lindley distribution. *Biometrics*, 26(1):145–149, 1970.

[62] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[63] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[64] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[65] G. Petris. An R package for dynamic linear models. *Journal of Statistical Software*, 36(12):1–16, 2010.

[66] M. Plummer, N. Best, K. Cowles, and K. Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[67] M. A. Jazi, G. Jones, and Chin-Diew Lai. First-order integer valued ar processes with zero inflated poisson innovations. *Journal of Time Series Analysis*, 33(6):954–963, 2011.

[68] M. Csorgo and L. Horváth. *Limit Theorems in Change-Point Analysis.* Wiley, 1997.

[69] P. Perron. *Dealing with structural breaks. In Palgrave Handbook of Econometrics 1*. Pallgrave McMillan, 2006.

[70] A. Aue and L. Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16, 2013.

[71] J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

[72] J. Bai. Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, 91(2):299 – 323, 1999.

[73] E. Kurozumi and P. Tuvaandorj. Model selection criteria in multivariate models with multiple structural changes. *Journal of Econometrics*, 164(2):218 – 238, 2011.

[74] P. Preuss, R. Puchstein, and H. Dette. Detection of multiple structural breaks in multivariate time series. *Journal of the American Statistical Association*, 110(510):654–668, 2015.

[75] S. Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221 – 241, 1998.

[76] T.L. Lai, H. Liu, and H. Xing. Autoregressive models with piecewise constant volatility and regression parameters. *Statistica Sinica*, 15(2):279–301, 2005.

[77] T.L. Lai and H. Xing. A simple bayesian approach to multiple change-points. *Statistica Sinica*, 21(2):539–569, 2011.

[78] A. F. Martínez and R. H. Mena. On a nonparametric change point detection model in markovian regimes. *Bayesian Analysis*, 9(4):823–858, 12 2014.

[79] D. Barry and J. A. Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

[80] R. H. Loschi and F. R. B. Cruz. Applying the product partition model to the identification of multiple change points. *Advances in Complex Systems*, 05(04):371–387, 2002.

[81] P.L. Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, Jun 2006.

[82] R.H Loschi and F. R. B. Cruz. Extension to the product partition model: computing the probability of a change. *Computational Statistics & Data Analysis*, 48:255–268, 2005.

[83] Rosângela Loschi, Jeanne G Pontel, and Frederico Cruz. Multiple change-point analysis for linear regression models. *Chilean Journal of Statistics*, 1:93–112, 10 2010.

[84] J.A. Ferreira, R.H. Loschi, and M.A. Costa. Detecting changes in time series: A product partition model with across-cluster correlation. *Signal Processing*, 96:212 – 227, 2014.

[85] Yi-Ching Yao. Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches. *The Annals of Statistics*, 12(4):1434–1447, 12 1984.

[86] Carlin B. P., Gelfand A. E., and Smith A.F.M. Hierarchical bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):389–405, 1992.

[87] T.R Santos, Glaura C. F., and D. Gamerman. Comparison of classical and bayesian approaches for intervention analysis. *International Statistical Review / Revue Internationale de Statistique*, 78(2):218–239, 2010.

[88] G W. Cobb. The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika*, 65(2):243–251, 08 1978.

[89] E. Moreno, G. Casella, and A. Garcia-Ferrer. An objective bayesian analysis of the change point problem. *Stochastic Environmental Research and Risk Assessment*, 19(3):191–204, Aug 2005.

[90] M. E. Mead. Generalized inverse gamma distribution and its application in reliability. *Communications in Statistics - Theory and Methods*, 44(7):1426–1435, 2015.