Pontifícia Universidade Católica
do Rio de Janeiro

Rodrigo Sarlo Antonio Filho

Intermittent demand forecasting in retail: applications of the GAS framework

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Elétrica.

Advisor: Prof. Cristiano Augusto Coelho Fernandes

Rio de Janeiro
June 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Rodrigo Sarlo Antonio Filho

Intermittent demand forecasting in retail: applications of the GAS framework

Dissertation presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Elétrica. Approved by the Examination Committee:

Prof. Cristiano Augusto Coelho Fernandes
Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Denis Borenstein
UFRGS

Prof. Marcelo Cunha Medeiros
Departamento de Economia - PUC-Rio

Prof. Rutger Lit
VU Amsterdam

Rio de Janeiro, June the 10th, 2021

Rodrigo Sarlo Antonio Filho

Rodrigo Sarlo Antonio Filho received his B.Sc. degree in Economics in 2018 from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil.

Bibliographic data

Antonio Filho, Rodrigo

Intermittent demand forecasting in retail: applications of the GAS framework / Rodrigo Sarlo Antonio Filho; advisor: Cristiano Augusto Coelho Fernandes. – 2021.

86 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2021.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Modelos GAS. 3. Séries temporais de contagem. 4. Demanda intermitente. 5. Modelos inflados em zero. 6. Modelos hurdle . I. Fernandes, Cristiano. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Acknowledgments

To my parents and sisters, for the lifetime support.

To my advisor, Cristiano Fernandes, for the dedication, time and valuable lessons during this journey.

To the D-Lab team, for the skills learned along the years and great research environment.

To Rodrigo Gomes and Thiago Qualharini, for the data acquisition efforts that made this dissertation richer.

PUC-Rio and the Electrical Engineering Department, for the support and structure.

This study was financed by the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Antonio Filho, Rodrigo; Fernandes, Cristiano (Advisor). **Intermittent demand forecasting in retail: applications of the GAS framework**. Rio de Janeiro, 2021. 86p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Intermittent demand is defined by periods of zero sales interleaved with positive sales with highly variable quantities. Most stock keeping units at the store level can be characterized as containing such demand. Thus, accurate models for predicting series with intermittent demand have major impacts in relation to inventory management. In this dissertation we propose the use of the GAS framework with the appropriate distributions for count data, in addition to their versions with excess of zeroes, and apply the derived models to real data obtained from a large Brazilian retail chain. We demonstrate that the proposed models with excess of zeros are consistently estimated via maximum likelihood and the distribution of the estimator is asymptotically normal. The performance of the proposed models is compared to adequate benchmarks from the time series literature for count data and intermittent demand forecast. Forecasting is evaluated based on the accuracy of both the entire predictive distribution and point forecasts. Our results show that the proposed models, specially the one derived from hurdle Poisson distribution, perform better than the analyzed benchmarks.

Keywords

GAS models; Count data time series; Intermittent demand; Zero-inflated models; Hurdle models.

Resumo

Antonio Filho, Rodrigo; Fernandes, Cristiano. **Previsão de demanda intermitente no varejo: aplicações do framework GAS**. Rio de Janeiro, 2021. 86p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Demanda intermitente é definida por períodos de vendas nulas intercaladas com vendas positivas e de quantidade altamente variável. A maior parte das unidades de manutenção de estoque (stock keeping units, em inglês) ao nível loja pode ser caracterizada como contendo demanda desse tipo. Assim, modelos acurados para prever séries com demanda intermitente trazem grandes impactos em relação à gestão de estoque. Nesta dissertação nós propomos o uso do framework GAS com as distribuições adequadas para dados de contagem, além de suas versões com excesso de zeros, e aplicamos os modelos derivados a dados reais obtidos com uma grande rede varejista brasileira. Nós demonstramos que os modelos com excesso de zeros propostos são estimados de forma consistente por máxima verossimilhança e a distribuição dos estimadores é assintóticamente normal. A performance dos modelos propostos é comparada com benchmarks adequados das literaturas de séries temporais para dados de contagem e previsão de demanda intermitente. A avaliação das previsões é feita com base tanto na precisão da distribuição preditiva quanto na precisão das previsões pontuais. Nossos resultados mostram que os modelos propostos, em especial o modelo derivado sob distribuição hurdle Poisson, performam melhor do que os benchmarks analisados.

Palavras-chave

Modelos GAS; Séries temporais de contagem; Demanda intermitente; Modelos inflados em zero; Modelos hurdle.

Table of contents

List of figures

List of tables

List of Abreviations

ACF – Auto-correlation Function
AIC – Akaike Information Criterion
ARMA – Autoregressive Moving Average
ASE – Asymptotic Standard Error
BIC – Bayesian Information Criteron
DCS – Dynamic Conditional Score
ERP – Enterprise Resource Planning
EWMA – Exponentially Weighted Moving Average
GARCH – Generalized Autoregressive Conditional Heteroskedasticity
GARMA – Generalized Autoregressive Moving Average
GAS – Generalized Autoregressive Score
GLM – Generalized Linear Model
HP – Hurdle Poisson
HNB – Hurdle Negative Binomial
IRLS – Iteratively Reweighted Least Squares
MAE – Mean Absolute Error
MASE – Mean Absolute Scaled Error
MGARCH – Multivariate Generalized Autoregressive Conditional Heteroskedasticity
ML – Maximum Liklihood
MSE – Mean Squared Errors
NB – Negative Binomial
PMF – Probability Mass Function
RMSE – Root Mean Squared Error
RMSSE – Root Mean Squared Scaled Error
SE – Standard Error
SES – Simple Exponential Smoothing
SKU – Stock Keeping Unit
TSL – Time Series Lab
VAR – Vector Autoregression
ZINB – Zero-inflated Negative Binomial
ZIP – Zero-inflated Poisson

1
Introduction

1.1
Motivation

Stock replenishment is one of the main issues in supply chain management. Its objective is to determine the quantity of a certain product needed in stock to meet the future demand for a given time period. Some key components are involved in such decisions: some are exogenous, as the number of days needed by the vendor to deliver the product (i.e. lead time); some are strategical and defined by the company, as the number of days that each order must cover; and there is the quantity demanded by costumers, which can be influenced by actions of the company, such as price reductions, but is not deterministic nor in direct control of the firm.

Demand is a key source of uncertainty in retail operations. As such, there has been a whole body of statistical literature that is concerned about this issue. A comprehensive review was made by Fildes, Ma and Kolassa [1]. In this article, the authors enumerate the possible levels of aggregation for retail's needs and its mostly used forecasting techniques, with usages in strategical decisions (concerning the competitive environment in which the firm is involved), tactical decisions (about how to implement the strategy defined, for example with advertising) and operational decisions (concerning daily operations, such as demand and supply planning processes).

The focus of this dissertation is at the relevant level of demand forecasts needed for stock control: Stock Keeping Unit (SKU) x store level. Accurate forecasts at this granularity can benefit the company in other dimensions besides stock optimization. If the relevant explanatory variables are available, the demand model can also be used for promotions planning and price optimization for each store. In spite of these possible benefits, the main interest of this dissertation is in the development of accurate demand models for inventory management and, with that in mind, relevant evaluation metrics will be used.

As noted by Johnston, Boylan and Shale [2], most of the SKUs at any store are observed to have intermittent demand, i.e. several periods with zero demand interleaved with positive and variable order sizes, a kind of demand difficult to predict. Temporal aggregation is a possible way to relief this difficulty, as proposed in the ADIDA framework from Nikolopoulos et al. [3] In this dissertation, the problem of intermittent demand is tackled without temporal aggregation, so that our data is composed by SKU x store x day observations. As remarked in [2] and confirmed by previous experience from the authors, items with intermittent demand typically generate over 40% of a store income and require about 60% of the investment in stock, so that methods that address this issue directly are extremely important in retail forecasting.

1.2 Contributions

We now state the two main contributions of this dissertation.

First, we have derived score-driven or Generalized Autoregressive Score (GAS) models for distributions not previously studied. Most of the GAS literature focuses on volatility forecasting, see for example the repository that keeps track of GAS related papers: `http://www.gasmodel.com/`. In this case, it is expected that count data distributions do not receive much attention. Indeed, there is a work from Blasques, Holý and Tomanová [4] that studies some of the distributions we present in this work - namely: negative binomial (NB) and zero-inflated negative binomial (ZINB), although in the context of duration modelling. We extend the work for the ZINB distribution by letting the probability of zeroes arising from the Bernoulli process also follow a GAS dynamic. There is also a previous work from Blazsek and Escribano [5] for Poisson distribution under the GAS framework.

We derive the scaled scores from the GAS framework for the following distributions: Poisson, NB, zero-inflated Poisson (ZIP), ZINB, hurdle Poisson (HP) and hurdle negative binomial (HNB). Some of these distributions are mixtures of a Bernoulli process and a non-negative distribution. In this case, we also let the probability of success be time varying with a score-driven dynamic. To the best of our knowledge this is the first time that both components of the mixture are made time varying in a GAS framework.

The second contribution concerns the application of the GAS framework, proposed in [6] and [7], to intermittent demand forecasting problems. We evaluate the derived models and compare them with some benchmarks from the retail literature using the relevant metrics for stock optimization purposes. To the best of our knowledge, it is also the first time that the GAS framework

is applied in such setting.

1.3 Organization

The remainder of this dissertation is organized as follows. Chapter 2 presents the relevant literature in which the present study takes part.

Chapter 3 presents some standard time series models for count data and specific count data models for intermittent data. Chapter 4 focuses on the proposed GAS models to forecast the sales of products with intermittent demand. We present the model structure adopted in this work and detail specific issues in modelling within the GAS framework.

Chapter 5 presents a simulation study developed to address the issue of the distribution of the GAS estimators when working with hurdle and zero-inflated models. We also evaluate the adequacy of the chosen residuals used in diagnostics when working with correctly specified models.

Chapter 6 shows the application and evaluation on real data of the proposed GAS models and compares them with the relevant benchmarks. Chapter 7 wraps up our findings and suggests future research.

2
Literature review

This chapter is divided in three sections. The first section concerns the presentation of the GAS framework. We discuss the idea behind this methodology and provide an overview of this field of study. Intermittent demand is discussed after that. We present Croston's seminal study [8] and detail the advances after that for both mean demand predictions and also for its entire predictive distribution. This is followed by a review of count data models from the time series literature which can also address the intermittent demand forecasting problem.

2.1
Score-driven models

The Generalized Autoregressive Score (or GAS) model framework was independently developed by Creal, Koopman and Lucas [6] and Harvey [7], for the latter under the name of Dynamic Conditional Score (DCS) models. The idea behind the methodology is to provide an unified treatment for dynamic models for any probability distribution, discrete or continuous. These models are classified as observation-driven under the categorization developed by Cox et al. [9]. They provide a generalization of the autoregressive moving average (ARMA) models for non-Gaussian distributions. The analogue for the disturbance, or innovation, in ARMA models is the scaled score, which drives the variation in the parameters being modeled. As previously mentioned, the following website compiles works that use this framework `http://www.gasmodel.com/`

As presented in [6], the GAS framework encompasses many well-known observation driven models as special cases, such as generalized autoregressive conditional heteroskedasticity (GARCH), autoregressive conditional duration, autoregressive conditional intensity, and Poisson count models with time-varying mean. The latter will be discussed in Chapter 4.

Blazsek and Licht [10] present an overview of the applications of score-driven models and relate the presented works with other models from the time series literature. Their first example is based on a previous work by Harvey and Sucarrat [11]. Both works discuss how GAS models with the choice of

adequate fat tail distributions are capable of winsorizing the innovations in the model. This feature is extremely helpful in volatility modelling. The review also presents examples of GAS models for both univariate and multivariate time series, which we will discuss in the ongoing paragraphs.

Concerning multivariate models, there are some different streams of research. Some works resemble Vector Autoregression (VAR) and multivariate generalized autoregressive conditional heteroskedasticity (MGARCH) models, meaning they relate directly future observations to previous ones. Alternatively, the series can be modeled through copulas and/or latent factors. An example of the former is the work from Blazsek, Escribano and Licht [12], that presents a generalization of VAR models that follows a multivariate Student t distribution. The same authors also present a multivariate Student t model that can represent co-integration relations in [13]. The MGARCH analogue previously mentioned was developed by Creal, Koopman and Lucas in [14].

Latent factors provide a suitable way to model large panels of time series. Creal et al. [15] develop a dynamic factor model to forecast macroeconomic, credit and loss given default risk variables. The model is also capable of handling time series of different frequencies.

The possibility of using the GAS framework with dynamic copulas was first introduced and exemplified in the paper that proposed the score-driven dynamic [6]. Opschoor et al. [16] model the dependence of 100 stocks through a factor dynamic copula model.

GAS models can be set up using two possible structures: an ARMA-like structure and an unobserved components structure, akin to the structural models of Harvey [17]. Examples of the former are also presented in [6]. The idea of the former is to relate the forecast of the parameters directly to previous forecasts of the same parameters and also to the innovations of the model, that is, the scaled score. The latter, which is the structure of our choice for this dissertation, relates the forecast of the parameters with unobserved components that have an interpretation, such as trend, seasonality and cycle. Harvey and Luati [18] employ such structure and relate its use in the GAS framework with the Gaussian unobserved component model estimated through the Kalman filter. A similar work, but with different distributions, is presented by Caivano, Harvey and Luati [19].

Other works that relate more directly to this dissertation are the previously mentioned works [4] and [5]. The former work, by Blasques, Holý and Tomanová, compares the ZINB GAS model (with a static Bernoulli variable) with other models in the context of duration forecasting. The objective is to accurately forecast the length of the time interval between two successive trans-

actions of stocks of the Dow Jones index. The ZINB GAS model is compared with benchmarks relevant for duration modelling. As mentioned, we extend their work by making the Bernoulli variable also dynamic, but in a different context. They also present the GAS negative binomial model.

Blazsek and Escribano [5] study patents registered at year x firm level in the US. They propose a GAS fixed effects panel data model with Poisson distribution. As in standard panel models, the parameters are collectively estimated for the regressors, the GAS dynamics parameters and the initialization of the recurrence relations. The GAS component is responsible for introducing serial correlation at firm level.

Another work that deals with excessive zeroes using the GAS framework is Harvey and Ito [20], although with a different distribution. The authors are focused on augmenting the probability of observing zeroes with continuous variables. The probability of observing zeroes is also dynamic, but not driven by the scaled score. Instead they use the dynamic parameter of the continuous distribution as the driver of this probability.

2.2 Intermittent demand

Croston [8] was the first to recognize that the widely used Simple Exponential Smoothing (SES), or Exponentially Weighted Moving Average (EWMA), method was not suitable for intermittent demand forecasting. His proposed solution breaks down the forecast of retail series in two components: the interval between positive demands and order size (given that an order occurs), and employs one SES recursion to forecast each component using the same smoothing constant for both methods. The predicted demand is simply the division of predicted order size by predicted interval between demands.

Croston's method became the standard for intermittent demand forecasting, and is still widely used in Enterprise Resource Planning (ERP) softwares. It was later recognized by Syntetos and Boylan [21] that the method produces biased mean forecasts. To solve this shortcoming, the authors propose an approximately unbiased modification, which involves multiplying Croston's forecast by a constant.

Another issue with the method is that the demand prediction is only updated when a new order is observed. To overcome this, Teunter, Syntetos and Babai [22] propose another method inspired by Croston's work in which the demand prediction is updated every period, having observed an order or not.

There are some other proposed methods that are not simple modifications of Croston's method, but the idea of separating observed demand into its constituent elements is kept. One of such works was developed by Gutierrez, Solis and Mukhopadhyay [23]. The authors develop a neural network with, basically, the same demand constituents of Croston's method. Willemain, Smart and Schwarz [24] develop a bootstrap algorithm to simulate the lead time demand cumulative distribution based on the same ideas.

Aside from proposing new methods, there is a stream of research that focuses on demand categorization. The aim is to define sets of rules for choosing the forecasting method for an observed demand time series based on descriptive statistics from the series. This is important for a firm that needs demand predictions for large numbers of SKUs, in which the computational burden of multiple pseudo-out-of-sample evaluations could be prohibitive. This idea was first introduced in Johnston and Boylan [25]. The authors argue that the classifications should be based on comparative forecasting performance of the methods and the rules defined based on which set of series characteristics each method performs better.

After this work, other rules have been proposed, as the one discussed in Syntetos, Boylan and Croston [26]. This rule is based on theoretical results for Mean Squared Errors (MSE) of different methods. The rule was later evaluated on real data. Kostenko and Hyndman [27] provided a correction for the rule developed and, later, the correction was further validated by Heinecke, Syntetos and Wang [28].

Except for the bootstrap algorithm previously mentioned, none of the presented works focuses on predictive distributions, providing only point forecasts. Syntetos, Babai and Altay [29] and Johnston, Boylan and Shale [2] analyse some candidate distributions and evaluate its adequacy on real data. Snyder, Ord and Beaumont [30] evaluate SES-like recursions estimated with Poisson, hurdle shifted Poisson and NB distributions to forecast the mean parameters of these. Hydman et al. [31] previously mentioned the same possibility.

The work presented in [30] also highlights that evaluating the predictions of low count data methods only with point forecast measures is inadequate. The solution given by the authors is to also evaluate the methods based on the entire predictive distribution. The same argument was presented by Kolassa [32] and also by Czado, Gneiting and Held [33].

We follow [30] and evaluate the proposed models with scale-free error measures for point forecasts and metrics for the entire predictive distribution. Also, our benchmarks from intermittent demand literature are the ones pre-

sented in [30]. These points will be further discussed later in our work.

2.3 Count data

Hilbe [34] provides a good introduction to count data models, though the work and examples focus on cross section and panel data. It also addresses issues typical of econometric literature, as censored data and instrumental variables. Besides standard Poisson and NB distribution, the author also presents the hurdle and zero-inflated static versions of both models. A review focused on count data models with excessive zeroes is developed by Greene [35].

For time series data, some of the earlier count data models are part of a framework presented by Benjamin, Rigby and Stasinopoulos [36]. The methodology is called generalized autoregressive moving average (GARMA), and the term "generalized" here means that the class of models developed are a generalization of the ARMA models to variables with distributions that belong to the exponential family. These models are also categorized as observation-driven in Cox's classification.

It is interesting to note that, for suitable choices in each framework, the GAS and GARMA methodologies can give the same recursive equations for the time varying parameters. An example of this equivalence is shown in Section 4.2.1.

Some of the models proposed before in count data literature that are part of the GARMA framework are presented in Davis [37] and Fokianos and Tjøstheim [38], for Poisson distribution, and Davis and Wu [39], for NB distribution.

3
Count data models

Here we present the models that will be used in this dissertation when comparing the forecast accuracy of our proposed GAS models. They are divided in two categories: the first encompasses those models originated from the count time series literature, and the second is specific for models associated with the intermittent demand forecasting literature.

Within each section, we present the model equations, the hypothesised distribution (when necessary), the estimation procedure, the way explanatory variables are included, and the forecasting algorithm.

3.1
Standard count data models

GARMA models provide a flexible framework that extends ARMA models to distributions of the exponential family. Poisson and NB distributions are suitable for count data and are members of the exponential family, making GARMA models natural benchmarks in the present context. The general form of a GARMA(p,q) model, as presented by Benjamin, Rigby and Stasinopoulos [36], is detailed in the ongoing paragraphs.

If y_t is the variable with distribution belonging to the exponential family, we model its time varying mean as a function of past observations:

$$\begin{aligned} \mu_{t|t-1} =& E(y_t|Y^{t-1}) \\ \text{where: } & Y^{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_1) \end{aligned} \tag{3-1}$$

Then the evolution of $\mu_{t|t-1}$ is given by:

$$g(\mu_{t|t-1}) = x_t^{'}\beta + \sum_{i=1}^{p} \phi_i \mathcal{A}(y_{t-i}, x_{t-i}, \beta) + \sum_{j=1}^{q} \theta_j \mathcal{M}(y_{t-j}, \mu_{t-j|t-j-1}) \tag{3-2}$$

Where $g(.)$ is a link function that specifies how the conditional mean of the chosen distribution, $\mu_{t|t-1}$, evolves in time and is affected by the explanatory variables. $\mathcal{A}$ is a function that represents the autoregressive terms

and the function $\mathcal{M}$ is for the moving average terms. x_t is a vector of explanatory variables. The fixed and unkown parameters in the model are $[\{\phi_i\}_{i=1}^{p}, \{\theta_j\}_{j=1}^{q}, \beta]$, with ϕ_i´s associated with autorregressive terms, θ_j´s for moving averages and β being the vector of coefficients of the explanatory variables.

The authors discuss some models presented in time series literature that are special cases of this general form. We implement a special case of the above equation, a GARMA(1,0) model, with the following structure:

$$\ln(\mu_{t|t-1}) = x_t^{'}\beta + \phi y_{t-1} \tag{3-3}$$

As discussed in [36], the inclusion of explanatory variables is natural in this framework. Note that a GARMA(0,0) model is a generalized linear model (GLM) regression. As such, we do not need to worry about initialization of the coefficients, differently from all other models in this study. GLMs have been widely studied in statistics and procedures with years of usage are available for the optimization of the estimation routine. Nevertheless, we need to select among the available regressors which are relevant for forecasting. In this case, we employ the same heuristic that will be presented in Section 4.5 for GAS models.

Estimation in GARMA framework is done through iteratively reweighted least squares (IRLS), the same algorithm used for GLM models. This makes these models very fast to estimate. k-steps-ahead forecasting is based on simulation in a manner similar to the one presented in Section 4.7.

We also implement two other count data models that are not members of the exponential family but have the same structure in terms of the $\mu_{t|t-1}$ equation as given in equation 3-3. These are regression models with ZIP and ZINB distributions. Now there is also the need to specify an equation for $\pi_{t|t-1}$. Inspired by the $p_{t|t-1}$ equation presented in the following section, π_{t-1} is specified as:

$$\ln\left(\frac{\pi_{t|t-1}}{1-\pi_{t|t-1}}\right) = \alpha + \delta z_{t-1} \tag{3-4}$$

where z_{t-1} is a dummy variable that is equal to zero if $y_{t-1} > 0$ and one otherwise.

As these distributions are not members of the exponential family, they are not estimated via IRLS. In this case, estimation can be done via EM algorithm, as detailed in Section 4.3.1. We opt to work directly with the maximum

likelihood estimation without the EM algorithm. As will be detailed later, we did not find significant differences in parameter estimates when applying the EM algorithm or the chosen procedure.

As with the other models in this dissertation, k-steps-ahead predictive distributions are available through simulation. The same procedure for variable selection previously mentioned was also used here.

3.2 Intermittent demand count data models

Our main interest is to evaluate model performance for stock control. This objective makes it necessary that the methodology chosen is able to also generate forecasts for the entire predictive distribution. That is why we do not use models that are adaptations of Croston's method aiming to simply correct its bias, or other disadvantages, but can only generate mean forecasts.

The work presented in Snyder, Ord and Beaumont [30] provides our chosen benchmark models from the intermittent demand literature. The presentation in this section is based on their study.

The authors propose the use of two different recursive relations for parameter forecasting, which they call damped and undamped dynamics, as given by:

$$\textit{damped}: \mu_{t|t-1} = (1-\phi-\alpha)\mu + \phi\mu_{t-1|t-2} + \alpha y_{t-1}, \quad \phi, \alpha, \mu > 0, \quad \phi + \alpha < 1 \tag{3-5}$$

$$\textit{undamped}: \mu_{t|t-1} = (1-\alpha)\mu_{t-1|t-2} + \alpha y_{t-1}, \quad 0 < \alpha < 1 \tag{3-6}$$

Note that the undamped dynamic is analogous to a SES equation. The difference lies in the estimation procedure: instead of MSE minimization, the parameters here are estimated through maximum likelihood. The authors highlight that, with the distributions being employed (that we discuss bellow), $\mu_{t+k|t-1}$ stochastically converges to zero as k grows in the undamped dynamics case. With the forecasting horizons analysed in this dissertation we did not observe such effect. The damped dynamic has a long-run mean (μ) to which its predictions converge, therefore avoiding the stochastic convergence to zero mentioned above.

The distributions used for parameter estimation within this context are Poisson, NB and hurdle shifted Poisson. As we will present the first two distributions in the next chapter, we will not repeat them in here. The hurdle shifted Poisson is a little different from the HP presented for GAS

models: instead of a zero-truncated Poisson distribution in case of failure in the Bernoulli trial, the observation y_t is sampled from a Poisson shifted one unit to the right. The distribution is given by the following expression:

$$p(y_t|\mu_{t|t-1}, p_{t|t-1}) = \begin{cases} p_{t|t-1} & \text{if } y_t = 0 \\ (1-p_{t|t-1})\frac{\mu_{t|t-1}^{y_t-1}\exp(-\mu_{t|t-1})}{(y_t-1)!} & \text{else} \end{cases} \tag{3-7}$$
$$p_{t|t-1} \in (0,1), \quad \mu_{t|t-1} > 0$$

We use $\mu_{t|t-1}$ instead of $\lambda_{t|t-1}$ above in order to keep the same notation used in the presentation of the damped and undamped recursive equations. The mean of the above distribution is $p_{t|t-1}(\mu_{t|t-1}+1)$.

The authors also make $p_{t|t-1}$ dynamic with equations analogue to the damped and undamped dynamics presented, but instead of utilizing y_{t-1} for updating $p_{t|t-1}$, a variable x_{t-1} is defined. It is equal to zero if $y_{t-1} = 0$ and equal to one if $y_{t-1} > 0$. The same recursive equations and estimated parameters are always used for both $\mu_{t|t-1}$ and $p_{t|t-1}$, only the starting values are separate for each parameter.

As we will discuss later for GAS models, starting values are needed to initialize the estimation procedure. We initialize ϕ and α with the fixed value 0.2; μ_1 is initialized as the mean of the first six observations, and p_1 is equal to the mean of the first six x_ts. Differently from the GAS models, here we optimize μ_1 and p_1, since these models are much faster to estimate. When working with damped dynamics we set $\mu = \mu_1$ and $p = p_1$, i.e. the model is initialized with the long run mean.

The inclusion of explanatory variables can be done here by simply adding a $\zeta^{'}X_t$ term to the damped and undamped dynamics, just as will be done for GAS models in Section 4.5. We follow exactly the same process for variable selection and initialization from Chapter 4 here with the intermittent demand models.

The forecasting algorithm for these models is also analogue to the one presented in Chapter 4. The distribution for k-steps-ahead ($k \geq 2$) predictions is only available via simulation. The algorithm simulates m paths of k successive iterations of sampling from the predicted distribution and calculating the updated parameters. The difference here is that there are no unobserved states, so the calculation is directly done for the parameters.

4
Score-driven models for time series of counts

This chapter discusses in detail the derivation of models suitable for intermittent demand forecasting using the GAS framework. We begin presenting the form in which parameters evolve in time. After that we present the chosen distributions and the scaled score derived to drive the variation in the GAS framework. We then discuss how to estimate the model. This is followed by a presentation of the heuristics proposed for initializing both the unobserved components and the static parameters that we need to estimate. The next step is to present the inclusion of explanatory variables in the model, followed by the presentation of the diagnostics. The chapter is concluded with the description of the forecasting algorithm for k-steps-ahead, $k \geq 2$.

4.1
GAS and unobserved components

The presentation in this section is based on the previously mentioned works of Creal, Koopman and Lucas [6] and Harvey and Luati [18].

The basic setup for the GAS framework works as follows: let y_t be the variable of interest, $f_{t|t-1}$ the time-varying parameter of its conditional distribution and θ a vector of static parameters. Define $Y^t = \{y_1, \ldots, y_t\}$ and $F^t = \{f_0, \ldots, f_{t|t-1}\}$. The available information set at time t consists of $\{f_{t|t-1}, \mathcal{F}_{t-1}\}$, where:

$$\mathcal{F}_{t-1} = \{Y^{t-1}, F^{t-1}\}, \quad t = 1, \ldots, n \tag{4-1}$$

and y_t is assumed to be generated by the conditional density/probability mass function (PMF):

$$y_t \sim p(y_t | f_{t|t-1}, \mathcal{F}_{t-1}; \theta) \tag{4-2}$$

Here, we assume that the time-varying parameter $f_{t|t-1}$ is modeled as a function of unobserved components appropriate for daily retail time series:

$$h_f(f_{t|t-1}) = \mu_{t|t-1} + \gamma_{t|t-1} \tag{4-3}$$

$$\mu_{t|t-1} = \phi\mu_{t-1|t-2} + \rho_1\tilde{s}_{f,t-1|t-2}, \quad |\phi| < 1, \quad \rho_1 > 0 \tag{4-4}$$

$$\alpha_{t|t-1} = \alpha_{t-1|t-2} + \kappa_{t-1}\tilde{s}_{f,t-1|t-2} \tag{4-5}$$

$$\gamma_{t|t-1} = z_t^{'}\alpha_{t|t-1} \tag{4-6}$$

Where $h_f(.)$ is a suitable link function that will ensure that, whatever the values of $\mu_{t|t-1}$ and $\gamma_{t|t-1}$, $f_{t|t-1}$ will remain on its domain. More on link functions later. In our case, $\mu_{t|t-1}$ is a stationary trend component and $\gamma_{t|t-1}$ is a seasonal component. The specific choices of components presented were made based on experimentation with the dataset presented in Chapter 6, and a similar structure is reproduced in all models for a coherent model comparison. In other settings, different forms of for the trend could be used. We will discuss later the reason for this choice.

Equations 4-3 to 4-6 can be seen as a filter algorithm, and one similar to the single source of error state space model discussed in Hyndman et al. [31] in the sense that the driver of the variation in all components is the same: $s_{f,t|t-1}$. The definition of $s_{f,t|t-1}$ is presented now:

$$s_{f,t|t-1} = S_t * \nabla_{f,t}, \quad \nabla_{f,t} = \frac{\partial \ln p(y_t|f_{t|t-1}, \mathcal{F}_{t-1}; \theta)}{\partial f_{t|t-1}}, \quad S_t = S(t, f_{t|t-1}\mathcal{F}_{t-1}; \theta) \tag{4-7}$$

where $\nabla_{f,t}$ is the score of the conditional distribution hypothesised for y_t with respect to $f_{t|t-1}$, and $S()$ is a matrix of appropriated dimension used for scaling the score of the distribution. The term $s_{f,t|t-1}$ can be regarded as an innovation: recall that $E_{t-1}[\nabla_{f,t}] = 0$ and so $\{s_f\}$ forms a martingale difference sequence. A natural and widely used choice for S_t is:

$$S_t = \mathcal{I}_{t|t-1}^{-d}, \quad d = \left\{0, \frac{1}{2}, 1\right\} \tag{4-8}$$

$$\mathcal{I}_{t|t-1} = E[\nabla_f \nabla_f^{'}|\mathcal{F}_{t-1}] \overset{\text{def}}{=} E_{t-1}[\nabla_f \nabla_f^{'}] = -E_{t-1}\Big[\frac{\partial^2 \ln p(y_t|f_{t|t-1}, \mathcal{F}_{t-1}; \theta)}{\partial f_{t|t-1} f_{t|t-1}^{'}}\Big] \tag{4-9}$$

where $\mathcal{I}_{t|t-1}$ is Fisher information matrix.

In our applications, we set $d = 0$ because this choice gave us the most stable results.

For some of the distributions used in this dissertation, and presented in the next section, it is further necessary to specify π - the probability on a Bernoulli variable that accounts the excess of zeroes typically observed in retail time series. A possible choice is to estimate a fixed value for π, as in [4]. Here we opt to make π also a score-driven processes, which we call $\pi_{t|t-1}$. Its recursive equation is presented now:

$$h_\pi(\pi_{t|t-1}) = \delta + \beta h_\pi(\pi_{t-1|t-2}) + \rho_2 \tilde{s}_{\pi,t-1|t-2}, \quad |\beta| < 1, \quad \rho_2 > 0 \tag{4-10}$$

With $s_{\pi,t|t-1}$ defined similarly to $s_{f,t|t-1}$ and $h_\pi(.)$ a proper link function that will keep $\pi_{t|t-1} \in (0, 1)$. The dynamic specified is analogue to an AR(1) model.

Most of the distributions present parameters with natural constraints, that is, they can only take values in subsets of the real line. This being the case, it may be advisable to adopt specific parametrizations that ensure that those constraints will be obeyed.

We now show how the link functions can be formally introduced into the GAS framework. This is easily accomplished by use of the chain rule. Let $\tilde{f}_{t|t-1} = h(f_{t|t-1})$, where $h(.)$ is a continuous and invertible function that maps the real number $f_{t|t-1}$ to the relevant subspace of $\mathbb{R}$ for the distribution being used. Let $\dot{h} = \frac{\partial h(f_{t|t-1})}{\partial f_{t|t-1}}$, which is deterministic given $\mathcal{F}_{t-1}$, the information set. Then it can be shown that:

$$\tilde{\nabla}_t = \left(\dot{h}'\right)^{-1} * \nabla_t \tag{4-11}$$

$$\tilde{\mathcal{I}}_{t|t-1} = \left(\dot{h}'\right)^{-1} * \mathcal{I}_{t|t-1} * \left(\dot{h}\right)^{-1} \tag{4-12}$$

For the distributions so far discussed, the choices of functions $h(.)$ are:

$$\tilde{f}_{t|t-1} = h_f(f_{t|t-1}) = \ln(f_{t|t-1}) \tag{4-13}$$

$$\tilde{\pi}_{t|t-1} = h_\pi(\pi_{t|t-1}) = \ln\left(\frac{\pi_{t|t-1}}{1 - \pi_{t|t-1}}\right) \tag{4-14}$$

The log link function ensures that $f_{t|t-1} \in \mathbb{R}^+$, and the logit link makes $\pi_{t|t-1} \in (0, 1)$.

We now discuss the choice of the seasonal component ($\gamma_{t|t-1}$), which is the same presented in [18]. Recalling the equations of the seasonal component that is part of the $f_{t|t-1}$ parameter (equations 4-5 and 4-6), we have:

$$\begin{aligned}\alpha_{t|t-1} =& \alpha_{t-1|t-2} + \kappa_{t-1} s_{f,t-1|t-2} \\ \gamma_{t|t-1} =& z_t^{'} \alpha_{t|t-1}\end{aligned}$$

$\alpha_{t|t-1}$ is a vector that collects the seasonal components (one for each seasonal period), and $\alpha_{t-1|t-2}$ collects the previous forecast for the same values. κ_{t-1} is a time-varying vector that has an unknown parameter $\kappa > 0$. The elements of this vector are constrained to sum zero and this, together with an adequate initialization of the seasonal components in α, makes $\gamma_{t|t-1}$ have a zero sum during an entire seasonal cycle.

The elements $\kappa_{j,t-1}, j = 1, \ldots, m$ of the vector κ_{t-1} are such that $\kappa_{j,t-1} = \kappa$ when in season j, and $\kappa_{i,t-1} = -\frac{\kappa}{(m-1)}$ for $i \neq j$. We work with weekly seasonality, so $m = 7$.

z_t is a vector that selects the relevant seasonal forecast for time t, so that $\gamma_{t|t-1}$ is a single number representing the seasonal part of the parameter $f_{t|t-1}$. For example, if in period t we need to select the third component of $\alpha_{t|t-1}$, then $z_t^{'} = [0, 0, 1, 0, \ldots, 0]$.

If more than one seasonal cycle is present, to capture other intrayear seasonal effects, it is possible to add other stochastic seasonal components in order to capture these effects. One possible approach is to replicate the same structure employed in equations (4-5) and (4-6) for the other seasonal cycles. Alternatively, trigonometric seasonal components (as presented in Harvey [17]) can be used. In some cases these will provide a more parsimonious representation of the various seasonal effects, as fewer harmonics might be needed to adequately represent some of the seasonal effects. Finally, in case one of the components can be represented as a deterministic seasonal factor, this can be treated as an explanatory variable and estimated in the form presented in Section 4.5.

4.1.1 Comments on the choice of score scaling

As mentioned in Section 2.1, GAS models are typically used to forecast financial variables, which take values on the whole real line and the interest usually lies in volatility forecasting. In such setting, the values chosen for d in $s_{t|t-1} = \mathcal{I}_{t|t-1}^{-d} \nabla_t$ are usually 0.5 or 1. Authors claim that these values give

more stable forecasts and estimates across multiple windows.

Our application, on the other hand, concerns discrete non-negative variables (daily sales). To define the most appropriate value of d to use in this dissertation, we have done a small experiment. We have selected 5 time series from the dataset described in Chapter 6 and compared both in and out-of-sample performance, analysing log-likelihood and mean absolute error respectively, of the GAS Poisson model estimated for $d \in \{0, \frac{1}{2}, 1\}$. In all cases, $d = 0$ produced the best results in terms of both metrics being analysed.

The experiment was then confirmed by reestimating the same models using the Time Series Lab (TSL) software (available at `https://timeserieslab.com/`) developed by Lit, Koopman, Harvey and Gorgi and evaluating the same metrics. We have also validated the final estimates of the fixed parameters: using the same initialization and number of steps taken by the optimization procedure, we have arrived at the same estimated values produced by the TSL.

4.2 Distributions

We now present the distributions chosen to model intermittent demand time series using the GAS framework. All of the following distributions are discrete and have non-negative support. For each distribution, we present the respective PMF, followed by its mean and variance. Then we present the derived score and Fisher information matrix and the reparametrizations used.

4.2.1 Poisson

The Poisson distribution is the simplest among all of the discrete distributions suitable for count data time series. It has just one parameter:

$$y_t \sim Poisson(\lambda_{t|t-1}) \tag{4-15}$$

$$p(y_t|\lambda_{t|t-1}) = \frac{\lambda_{t|t-1}^{y_t} \exp(-\lambda_{t|t-1})}{y_t!}, \quad \lambda_{t|t-1} > 0, \quad y_t \geq 0 \tag{4-16}$$

The most well known property of the Poisson distribution is called equidispersion. It means that the mean and variance are the same:

$$E_{t-1}[y_t] = \lambda_{t|t-1} = Var_{t-1}[y_t] \tag{4-17}$$

This is a restrictive property when modelling real data. It is usually observed that $Var_{t-1}[y_t] > E_{t-1}[y_t]$, what is called overdispersion. This deficiency is what makes (sometimes) necessary to work with the negative binomial distribution, which will be presented in the next subsection.

Now we present the derivation of the components of the GAS specification:

$$\ln p(y_t|\lambda_{t|t-1}) = -\lambda_{t|t-1} + y_t \ln(\lambda_{t|t-1}) - \ln(y_t!) \tag{4-18}$$

$$\nabla_{\lambda_{t|t-1}} = \frac{\partial \ln p(y_t|\lambda_{t|t-1})}{\partial \lambda_{t|t-1}} = \left(\frac{y_t - \lambda_{t|t-1}}{\lambda_{t|t-1}}\right) \tag{4-19}$$

$$\mathcal{I}_{t|t-1} = E_{t-1}[\nabla^2_{\lambda_{t|t-1}}] = \frac{1}{\lambda_{t|t-1}} \tag{4-20}$$

Reparameterizing these components:

$$\tilde{\lambda}_{t|t-1} = h(\lambda_{t|t-1}) = \ln(\lambda_{t|t-1}) \Rightarrow \lambda_{t|t-1} = \exp(\tilde{\lambda}_{t|t-1}) \tag{4-21}$$

$$\dot{h} = \frac{\partial \ln(\lambda_{t|t-1})}{\partial \lambda_{t|t-1}} = \frac{1}{\lambda_{t|t-1}} \Rightarrow \dot{h}^{-1} = \lambda_{t|t-1} \tag{4-22}$$

$$\tilde{\nabla}_{\lambda_{t|t-1}} = \dot{h}^{-1} \nabla_{\lambda_{t|t-1}} = (y_t - \lambda_{t|t-1}) \tag{4-23}$$

$$\tilde{\mathcal{I}}_{t|t-1} = \dot{h}^{-2} \mathcal{I}_{t|t-1} = \lambda_{t|t-1} \tag{4-24}$$

So that the general form of the scaled score is given by:

$$\tilde{s}_{t|t-1} = \frac{y_t - \lambda_{t|t-1}}{\lambda^d_{t|t-1}} \tag{4-25}$$

In our particular study setting $d = 0$, results in $\tilde{s}_{t|t-1} = (y_t - \lambda_{t|t-1})$.

For the special case of the Poisson distribution, we are working exactly with a single source of error state space model described in Hyndman et al. [31], but estimated with a Poisson PMF.

An interesting property of the derived model is the equivalence between the GAS Poisson and the GARMA models presented in Chapter 3 - the derivations for the NB distribution works in an analogous form. If instead of adopting an unobserved component dynamic, we had used an ARMA dynamics, the resulting GAS(1,1) model will be given by:

$$f_{t|t-1} = \omega + \alpha f_{t-1|t-2} + \beta s_{t-1|t-2} \tag{4-26}$$

In the case of the Poisson distribution we have $\lambda_{t|t-1} = f_{t|t-1}$. Setting $d = 1$ in a GAS model without reparameterization, we have:

$$\lambda_{t|t-1} = \omega + \alpha\lambda_{t-1|t-2} + \beta(y_{t-1} - \lambda_{t-1|t-2}) \Rightarrow \lambda_{t|t-1} = \omega + (\alpha - \beta)\lambda_{t-1|t-2} + \beta y_{t-1}$$

This is the final form for the GAS model. Now we turn to equation 3-2 in a GARMA(1,1) specification. If we choose $g(x) = x$, i.e. the identity link, and make $x_t = [1, \ldots, 1]'$ - only an intercept - we have:

$$\mu_{t|t-1} = \beta_0 + \phi\mathcal{A}(y_{t-1}, \beta_0) + \theta\mathcal{M}(y_{t-1}, \mu_{t-1|t-2})$$

If we define $\mathcal{A}(y_{t-1}, x_{t-1}, \beta) = (y_{t-1} - x'_{t-1}\beta)$ and $\mathcal{M}(y_{t-1}, \mu_{t-1|t-2}) = (y_{t-1} - \mu_{t-1|t-2})$, then:

$$\mu_{t|t-1} = \beta_0 + \phi(y_{t-1} - \beta_0) + \theta(y_{t-1} - \mu_{t-1|t-2}) \Rightarrow \mu_{t|t-1} = \beta_0(1 - \phi) + (\phi + \theta)y_{t-1} - \theta\mu_{t-1|t-2}$$

So we have that for suitable choices of d in the GAS model framework, and $g(.)$, $\mathcal{A}$ and $\mathcal{M}$ in GARMA models, the models are equivalent.

4.2.2 Negative binomial (NB)

There are some possible formulations for the negative binomial distribution. In this work, we chose the one in which the mean is a parameter:

$$y_t \sim NB(\mu_{t|t-1}, \alpha) \tag{4-27}$$

$$p(y_t|\mu_{t|t-1}, \alpha) = \frac{\Gamma(y_t + \alpha)}{\Gamma(y_t + 1)\Gamma(\alpha)} \left(\frac{1}{1 + \frac{\mu_{t|t-1}}{\alpha}}\right)^{\alpha} \left(1 - \frac{1}{1 + \frac{\mu_{t|t-1}}{\alpha}}\right)^{y_t} \tag{4-28}$$

$$\mu_{t|t-1},\ \alpha > 0, \quad y_t \geq 0$$

For the NB distribution there is also a relation between the mean and the variance, but a more flexible one:

$$E_{t-1}[y_t] = \mu_{t|t-1} \tag{4-29}$$

$$Var_{t-1}[y_t] = \mu_{t|t-1}\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right) \tag{4-30}$$

It is interesting to note that $\left(1 + \frac{\mu}{\alpha}\right) > 1$, so that we have $Var_{t-1}[y_t] > E_{t-1}[y_t]$, that is, overdispersion.

Now we present the derivation of the score and Fisher Information for this model:

$$\begin{aligned}\ln p(y_t|\mu_{t|t-1}, \alpha) =& \ln(\Gamma(y_t + \alpha)) - \ln(\Gamma(y_t + 1)) - \ln(\Gamma(\alpha)) - \\ &- \alpha \ln\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right) + y_t \ln(\mu_{t|t-1}) - y_t \ln(\mu_{t|t-1} + \alpha)\end{aligned} \tag{4-31}$$

$$\nabla_{\mu_{t|t-1}} = \frac{\partial \ln p(y_t|\mu_{t|t-1}, \alpha)}{\partial \mu_{t|t-1}} = \left(\frac{y_t - \mu_{t|t-1}}{\mu_{t|t-1}\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right)}\right) \tag{4-32}$$

$$\mathcal{I}_{t|t-1} = E_{t-1}[\nabla^2_{\mu_{t|t-1}}] = \frac{1}{\mu_{t|t-1}\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right)} \tag{4-33}$$

Reparameterizing to ensure that $\mu_{t|t-1} > 0$, we have that:

$$\tilde{\mu_{t|t-1}} = h(\mu_{t|t-1}) = \ln(\mu_{t|t-1}) \tag{4-34}$$

$$\tilde{\nabla}_{\mu_{t|t-1}} = \dot{h}^{-1}\nabla_{\mu_{t|t-1}} = \left(\frac{y_t - \mu_{t|t-1}}{\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right)}\right) \tag{4-35}$$

$$\tilde{\mathcal{I}}_{t|t-1} = \dot{h}^{-2}\mathcal{I}_{t|t-1} = \frac{\mu_{t|t-1}}{\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right)} \tag{4-36}$$

The scaled score for the NB distribution is:

$$\tilde{s}_{t|t-1} = \left(\frac{y_t - \mu_{t|t-1}}{\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right)}\right)\left(\frac{\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right)}{\mu_{t|t-1}}\right)^d \tag{4-37}$$

In our particular case, in which we set $d = 0$, we have $\tilde{s}_{t|t-1} = \left(\frac{y_t - \mu_{t|t-1}}{\left(1 + \frac{\mu_{t|t-1}}{\alpha}\right)}\right)$.

4.2.3 Zero-inflated Poisson (ZIP)

The standard Poisson distribution has a shortcoming when modelling very low count time series: it is sometimes observed that the actual proportion of zeroes is greater than what would be predicted by a Poisson distribution - $p(y = 0|\lambda) = \exp(-\lambda)$.

An usually employed solution to such shortcoming is to add a Bernoulli trial to the Poisson distribution, which is responsible for generating those extra zeroes. The resulting mixture distribution works as follows: first we begin with the trial and, in case of success (what happens with probability π), we observe $y_t = 0$; in case of failure, then we sample from a standard Poisson distribution. Note that we can also have zeroes from this second stage.

Having mentioned the logic behind the zero-inflated Poisson distribution, we now present its definition:

$$p(y_t|\lambda_{t|t-1}, \pi_{t|t-1}) = \begin{cases} \pi_{t|t-1} + (1 - \pi_{t|t-1}) \exp(-\lambda_{t|t-1}) & \text{, if } y_t = 0; \\ (1 - \pi_{t|t-1}) \frac{\lambda_{t|t-1}^{y_t} \exp(-\lambda_{t|t-1})}{y_t!} & \text{, else.} \end{cases} \tag{4-38}$$

$$\lambda_{t|t-1} > 0, \quad \pi_{t|t-1} \in (0, 1), \quad y_t \geq 0$$

Note that, since $\pi + (1 - \pi) \exp(-\lambda) \Rightarrow \pi + \exp(-\lambda) - \pi \exp(-\lambda) > \exp(-\lambda)$, the probability of a zero is higher under the ZIP distribution when compared with the Poisson. It can be shown that:

$$E_{t-1}[y_t] = \lambda_{t|t-1}(1 - \pi_{t|t-1}) \tag{4-39}$$

$$Var_{t-1}[y_t] = \lambda_{t|t-1}(1 - \pi_{t|t-1})(1 + \lambda_{t|t-1}\pi_{t|t-1}) \tag{4-40}$$

As would be expected, the Bernoulli trial added before sampling from the Poisson distribution reduces the mean when compared to the standard Poisson, since $\pi \in (0, 1) \Rightarrow \lambda(1 - \pi) < \lambda$. Another interesting property is that this distribution has overdispersion, since $(1+\lambda\pi) > 1 \Rightarrow Var_{t-1}[y_t] > E_{t-1}[y_t]$.

This and the HP distribution are the simplest discrete distributions to handle the excessive zeroes observed in intermittent demand data. The differences between both will become clear in the HP subsection.

The following equations present the log density, the scores for both λ and π and Fisher information matrix:

$$\ln(p(y_t|\lambda_{t|t-1},\pi_{t|t-1})) = \begin{cases} \ln\{\pi_{t|t-1} + (1-\pi_{t|t-1})\exp(-\lambda_{t|t-1})\} & \text{, if } y_t = 0; \\ \ln(1-\pi_{t|t-1}) - \lambda_{t|t-1} + y_t \ln(\lambda_{t|t-1}) - \ln(y_t!) & \text{, else.} \end{cases} \tag{4-41}$$

$$\nabla_{y_t=0} = \begin{bmatrix} \nabla_{y_t=0}^{\pi_{t|t-1}} \\ \nabla_{y_t=0}^{\lambda_{t|t-1}} \end{bmatrix} = \frac{1}{\pi_{t|t-1} + (1-\pi_{t|t-1})\exp(-\lambda_{t|t-1})} \begin{bmatrix} 1-\exp(-\lambda_{t|t-1}) \\ (\pi_{t|t-1}-1)\exp(-\lambda_{t|t-1}) \end{bmatrix} \tag{4-42}$$

$$\nabla_{y_t>0} = \begin{bmatrix} \nabla_{y_t>0}^{\pi_{t|t-1}} \\ \nabla_{y_t>0}^{\lambda_{t|t-1}} \end{bmatrix} = \begin{bmatrix} \frac{-1}{(1-\pi_{t|t-1})} \\ \frac{y_t-\lambda_{t|t-1}}{\lambda_{t|t-1}} \end{bmatrix} \tag{4-43}$$

$$\begin{aligned} \mathcal{I}_{t|t-1} =& E_{t-1}\begin{bmatrix} (\nabla^{\pi_{t|t-1}})^2 & \nabla^{\pi_{t|t-1}}\nabla^{\lambda_{t|t-1}} \\ \nabla^{\lambda_{t|t-1}}\nabla^{\pi_{t|t-1}} & (\nabla^{\lambda_{t|t-1}})^2 \end{bmatrix} = \frac{1}{\pi_{t|t-1} + (1-\pi_{t|t-1})\exp(-\lambda_{t|t-1})} * \\ & * \begin{bmatrix} \frac{1-\exp(-\lambda_{t|t-1})}{(1-\pi_{t|t-1})} & -\exp(-\lambda_{t|t-1}) \\ -\exp(-\lambda_{t|t-1}) & \frac{\pi_{t|t-1}(1-\pi_{t|t-1})+\exp(-\lambda_{t|t-1})\{(1-\pi_{t|t-1})^2+\lambda_{t|t-1}\pi_{t|t-1}(\pi_{t|t-1}-1)\}}{\lambda_{t|t-1}} \end{bmatrix} \end{aligned} \tag{4-44}$$

Now we reperameterize the components to ensure that $\pi_{t|t-1} \in (0,1)$ and $\lambda_{t|t-1} > 0$:

For $\pi_{t|t-1}$ we choose the logit link function:

$$\tilde{\pi}_{t|t-1} = h(\pi_{t|t-1}) = \ln\left(\frac{\pi_{t|t-1}}{1-\pi_{t|t-1}}\right) \Rightarrow \pi_{t|t-1} = \frac{\exp(\tilde{\pi}_{t|t-1})}{1+\exp(\tilde{\pi}_{t|t-1})} \tag{4-45}$$

$$\dot{h_\pi} = \frac{\partial \ln\left(\frac{\pi_{t|t-1}}{1-\pi_{t|t-1}}\right)}{\partial \pi_{t|t-1}} = \frac{1}{\pi_{t|t-1}(1-\pi_{t|t-1})} \Rightarrow \dot{h_\pi}^{-1} = \pi_{t|t-1}(1-\pi_{t|t-1}) \tag{4-46}$$

While for $\lambda_{t|t-1}$ we choose the log link:

$$\tilde{\lambda}_{t|t-1} = h(\lambda_{t|t-1}) = \ln(\lambda_{t|t-1}) \Rightarrow \dot{h_\lambda}^{-1} = \lambda_{t|t-1} \tag{4-47}$$

$$\text{Define} \quad \dot{H}^{-1} = \begin{bmatrix} \dot{h_\pi}^{-1} & 0 \\ 0 & \dot{h_\lambda}^{-1} \end{bmatrix} = \begin{bmatrix} \pi_{t|t-1}(1-\pi_{t|t-1}) & 0 \\ 0 & \lambda_{t|t-1} \end{bmatrix}, \text{ then}$$

$$\tilde{\nabla}_{y_t=0} = \begin{bmatrix} \tilde{\nabla}_{y_t=0}^{\pi_{t|t-1}} \\ \tilde{\nabla}_{y_t=0}^{\lambda_{t|t-1}} \end{bmatrix} = \dot{H'}^{-1} \nabla_{y_t=0} =$$

$$= \frac{1}{\pi_{t|t-1} + (1-\pi_{t|t-1})\exp(-\lambda_{t|t-1})} \begin{bmatrix} \pi_{t|t-1}(1-\pi_{t|t-1})(1-\exp(-\lambda_{t|t-1})) \\ \lambda_{t|t-1}(\pi_{t|t-1}-1)\exp(-\lambda_{t|t-1}) \end{bmatrix} \tag{4-48}$$

$$\tilde{\nabla}_{y_t>0} = \begin{bmatrix} \tilde{\nabla}_{y_t>0}^{\pi_{t|t-1}} \\ \tilde{\nabla}_{y_t>0}^{\lambda_{t|t-1}} \end{bmatrix} = \dot{H'}^{-1} \nabla_{y_t>0} = \begin{bmatrix} -\pi_{t|t-1} \\ y_t - \lambda_{t|t-1} \end{bmatrix} \tag{4-49}$$

$$\tilde{\mathcal{I}}_{t|t-1} = \dot{H'}^{-1} \mathcal{I}_{t|t-1} \dot{H}^{-1} = \frac{1}{\pi_{t|t-1} + (1-\pi_{t|t-1})\exp(-\lambda_{t|t-1})} *$$

$$* \begin{bmatrix} (\pi_{t|t-1})^2(1-\pi_{t|t-1})(1-\exp(-\lambda_{t|t-1})) & -\exp(-\lambda_{t|t-1})\lambda_{t|t-1}\pi_{t|t-1}(1-\pi_{t|t-1}) \\ -\exp(-\lambda_{t|t-1})\lambda_{t|t-1}\pi_{t|t-1}(1-\pi_{t|t-1}) & \lambda_{t|t-1}[\pi_{t|t-1}(1-\pi_{t|t-1}) + \exp(-\lambda_{t|t-1}) * \\ & *\{(1-\pi_{t|t-1})^2 + \lambda_{t|t-1}\pi_{t|t-1}(\pi_{t|t-1}-1)\}] \end{bmatrix} \tag{4-50}$$

In the case of this and for the next distributions (HP, ZINB and HNB), we have one scaled score for each situation:

$$\tilde{s} = \begin{cases} \tilde{\nabla}_{y=0} * \tilde{\mathcal{I}}^{-d} & \text{, if } y = 0; \\ \tilde{\nabla}_{y>0} * \tilde{\mathcal{I}}^{-d} & \text{, else.} \end{cases} \tag{4-51}$$

But in our study, the chosen scailing with $d = 0$ reduces the scaled score to:

$$\tilde{s} = \begin{cases} \tilde{\nabla}_{y=0} & \text{, if } y = 0; \\ \tilde{\nabla}_{y>0} & \text{, else.} \end{cases} \tag{4-52}$$

It is interesting to note that the derived GAS model for this and the following distributions overcomes one of the Croston's method [8] deficiency: in our framework, the predicted demand is updated every period. This happens since in all of the following distributions the $\tilde{\nabla}_{y=0}$ vector is never null.

4.2.4
Zero-inflated negative binomial (ZINB)

The zero-inflated negative binomial distribution follows the same logic of the ZIP: it is a mixture of a Bernoulli trial followed by a standard negative binomial distribution in the case of failure.

The ZINB distribution is defined as follows:

$$p(y_t|\mu_{t|t-1},\alpha,\pi_{t|t-1}) = \begin{cases} \pi_{t|t-1} + (1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha} & \text{, if } y_t = 0; \\ (1-\pi_{t|t-1})\frac{\Gamma(y_t+\alpha)}{\Gamma(y_t+1)\Gamma(\alpha)}\left(\frac{1}{1+\frac{\mu_{t|t-1}}{\alpha}}\right)^{\alpha}\left(1-\frac{1}{1+\frac{\mu_{t|t-1}}{\alpha}}\right)^{y_t} & \text{, else.} \end{cases} \tag{4-53}$$

$$\mu_{t|t-1},\ \alpha > 0, \quad \pi_{t|t-1} \in (0,1), \quad y_t \geq 0$$

The mean and variance of the ZINB distribution are:

$$E_{t-1}[y_t] = \mu_{t|t-1}(1-\pi_{t|t-1}) \tag{4-54}$$

$$Var_{t-1}[y_t] = \mu_{t|t-1}(1-\pi_{t|t-1})\left(1+\mu_{t|t-1}\left(\pi_{t|t-1}+\frac{1}{\alpha}\right)\right) \tag{4-55}$$

Note that, as with the Poisson and NB distributions, both ZIP and ZINB have the same mean, but the ZINB has a larger variance since $\alpha > 0$ and so $\pi + \frac{1}{\alpha} > \pi$.

Now we present the derivation of the scaled score:

$$\ln(p(y_t|\mu_{t|t-1},\alpha,\pi_{t|t-1})) = \begin{cases} \ln\{\pi_{t|t-1} + (1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}\} & \text{, if } y_t = 0; \\ \{\ln(1-\pi_{t|t-1}) + \ln(\Gamma(y_t+\alpha)) - \ln(\Gamma(y_t+1)) - \\ -\ln(\Gamma(\alpha)) - \alpha\ln\left(1+\frac{\mu_{t|t-1}}{\alpha}\right) + \\ +y_t\ln(\mu_{t|t-1}) - y_t\ln(\mu_{t|t-1}+\alpha)\} & \text{, else.} \end{cases} \tag{4-56}$$

$$\nabla_{y_t=0} = \begin{bmatrix} \nabla_{y_t=0}^{\pi_{t|t-1}} \\ \nabla_{y_t=0}^{\mu_{t|t-1}} \end{bmatrix} = \begin{bmatrix} \frac{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}}{\pi_{t|t-1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \\ \frac{(\pi_{t|t-1}-1)}{\pi_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})} \end{bmatrix} \tag{4-57}$$

$$\nabla_{y_t>0} = \begin{bmatrix} \nabla_{y_t>0}^{\pi_{t|t-1}} \\ \nabla_{y_t>0}^{\mu_{t|t-1}} \end{bmatrix} = \begin{bmatrix} \frac{-1}{(1-\pi_{t|t-1})} \\ \frac{y_t-\mu_{t|t-1}}{\mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})} \end{bmatrix} \tag{4-58}$$

$$\mathcal{I}_{t|t-1} = E_{t-1}\begin{bmatrix} (\nabla^{\pi_{t|t-1}})^2 & \nabla^{\pi_{t|t-1}}\nabla^{\mu_{t|t-1}} \\ \nabla^{\mu_{t|t-1}}\nabla^{\pi_{t|t-1}} & (\nabla^{\mu_{t|t-1}})^2 \end{bmatrix}$$

$$\text{where: } (\nabla^{\pi_{t|t-1}})^2 = \frac{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}}{(1-\pi_{t|t-1})[\pi_{t|t-1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}]}$$

$$\nabla^{\pi_{t|t-1}}\nabla^{\mu_{t|t-1}} = \nabla^{\mu_{t|t-1}}\nabla^{\pi_{t|t-1}} = \frac{-1}{\pi_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})}$$

$$(\nabla^{\mu_{t|t-1}})^2 = \frac{(\pi_{t|t-1}-1)^2}{(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+2}[\pi_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha}+(1-\pi_{t|t-1})]} + \frac{(1-\pi_{t|t-1})}{\mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})} - \frac{(1-\pi_{t|t-1})}{(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+2}} \tag{4-59}$$

Reparameterizing the time varying parameters:

$$\tilde{\pi}_{t|t-1} = h(\pi_{t|t-1}) = \ln\left(\frac{\pi_{t|t-1}}{1-\pi_{t|t-1}}\right) = h_\pi, \quad \tilde{\mu_{t|t-1}} = h(\mu_{t|t-1}) = \ln(\mu_{t|t-1}) = h_\mu$$

$$\tilde{\nabla}_{y_t=0} = \begin{bmatrix} \tilde{\nabla}^{\pi_{t|t-1}}_{y_t=0} \\ \tilde{\nabla}^{\mu_{t|t-1}}_{y_t=0} \end{bmatrix} = \dot{H'}^{-1}\nabla_{y_t=0} = \begin{bmatrix} \frac{\pi_{t|t-1}(1-\pi_{t|t-1})[1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}]}{\pi_{t|t-1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \\ \frac{\mu_{t|t-1}(\pi_{t|t-1}-1)}{\pi_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})} \end{bmatrix} \tag{4-60}$$

$$\tilde{\nabla}_{y_t>0} = \begin{bmatrix} \tilde{\nabla}^{\pi_{t|t-1}}_{y_t>0} \\ \tilde{\nabla}^{\mu_{t|t-1}}_{y_t>0} \end{bmatrix} = \dot{H'}^{-1}\nabla_{y_t>0} = \begin{bmatrix} -\pi_{t|t-1} \\ \frac{y_t-\mu_{t|t-1}}{(1+\frac{\mu_{t|t-1}}{\alpha})} \end{bmatrix} \tag{4-61}$$

$$\tilde{\mathcal{I}}_{t|t-1} = \dot{H'}^{-1}\mathcal{I}_{t|t-1}\dot{H}^{-1} = \begin{bmatrix} h_\pi^{-2}(\nabla^{\pi_{t|t-1}})^2 & h_\pi^{-1}h_\mu^{-1}\nabla^{\pi_{t|t-1}}\nabla^{\mu_{t|t-1}} \\ h_\pi^{-1}h_\mu^{-1}\nabla^{\mu_{t|t-1}}\nabla^{\pi_{t|t-1}} & h_\mu^{-2}(\nabla^{\mu_{t|t-1}})^2 \end{bmatrix} =$$

$$\text{where: } h_\pi^{-2}(\nabla^{\pi_{t|t-1}})^2 = \frac{[1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}][\pi^2_{t|t-1}(1-\pi_{t|t-1})]}{\pi_{t|t-1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}}$$

$$h_\pi^{-1}h_\mu^{-1}\nabla^{\mu_{t|t-1}}\nabla^{\pi_{t|t-1}} = \frac{-\mu_{t|t-1}\pi_{t|t-1}(1-\pi_{t|t-1})}{\pi_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+1}+(1-\pi_{t|t-1})(1+\frac{\mu_{t|t-1}}{\alpha})}$$

$$h_\mu^{-2}(\nabla^{\mu_{t|t-1}})^2 = \frac{(\pi_{t|t-1}-1)^2\mu^2_{t|t-1}}{(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+2}[\pi_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha}+(1-\pi_{t|t-1})]} + \\ + \frac{(1-\pi_{t|t-1})\mu_{t|t-1}}{(1+\frac{\mu_{t|t-1}}{\alpha})} - \frac{(1-\pi_{t|t-1})\mu^2_{t|t-1}}{(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+2}} \tag{4-62}$$

4.2.5 Hurdle Poisson (HP)

Hurdle models also increase the probability of zeroes in discrete count data models but using a different strategy from that adopted by zero-inflated models. We begin by presenting the hurdle Poisson PMF and then discuss the differences from the ZIP model:

$$p(y_t|\lambda_{t|t-1}, \pi_{t|t-1}) = \begin{cases} \pi_{t|t-1} & , \text{if } y_t = 0; \\ (1-\pi_{t|t-1}) \frac{\lambda_{t|t-1}^{y_t} \exp(-\lambda_{t|t-1})}{y_t!} \frac{1}{(1-\exp(-\lambda_{t|t-1}))} & , \text{else.} \end{cases} \tag{4-63}$$

$$\lambda_{t|t-1} > 0, \quad \pi_{t|t-1} \in (0,1), \quad y_t \geq 0$$

As in the case of zero-inflated models, hurdle distributions are also a mixture of the Bernoulli trial and Poisson or NB distributions, but in the present case the observation (y_t) is sampled from a zero-truncated distribution in case of failure in the Bernoulli trial.

The probability of a zero outcome in a hurdle model is $p(y_t = 0|\theta) = \pi$, since zeroes can only come from the Bernoulli trial. Recall that, as previously mentioned, under the standard Poisson distribution, $p(y_t = 0|\lambda) = \exp(-\lambda)$, so the term $\frac{1}{(1-\exp(-\lambda))}$ that multiplies the standard Poisson distribution is responsible for conditioning the distribution on taking positive values.

Differently from the ZIP model, the HP distribution cannot be shown to have a greater probability of zeroes than the Poisson, since the parameter λ is not related to the zeroes in this setting. The probability of zeroes is increased if $p(y = 0|\lambda, \pi) = \pi > \exp(-\lambda_{Poisson}) = p(y = 0|\lambda_{Poisson})$ when both models are used to fit the same dataset.

It can be shown that:

$$E_{t-1}[y_t] = \frac{\lambda_{t|t-1}}{1-\exp(-\lambda_{t|t-1})}(1-\pi_{t|t-1}) \tag{4-64}$$

$$Var_{t-1}[y_t] = \frac{\lambda_{t|t-1}(1-\pi_{t|t-1})}{1-\exp(-\lambda_{t|t-1})} - \frac{\lambda_{t|t-1}^2(1-\pi_{t|t-1})}{\exp(\lambda_{t|t-1})(1-\exp(-\lambda_{t|t-1}))^2} + \frac{\lambda_{t|t-1}^2 \pi_{t|t-1}(1-\pi_{t|t-1})}{(1-\exp(-\lambda_{t|t-1}))^2} \tag{4-65}$$

Note that $\exp(-\lambda) > 0 \Rightarrow \frac{1}{(1-\exp(-\lambda))} > 1$, so that if we assign the same values of λ and π and then sample from a ZIP and a HP distribution, the mean of the latter will be larger.

Now we derive the GAS scaled score components for this model:

$$\ln(p(y_t|\lambda_{t|t-1},\pi_{t|t-1})) = \begin{cases} \ln(\pi_{t|t-1}) & \text{, if } y_t = 0; \\ \Big\{ \ln(1-\pi_{t|t-1}) - \lambda_{t|t-1} + y_t \ln(\lambda_{t|t-1}) - & \\ -\ln(y_t!) - \ln(1-\exp(-\lambda_{t|t-1})) \Big\} & \text{, else.} \end{cases} \tag{4-66}$$

$$\nabla_{y_t=0} = \begin{bmatrix} \nabla_{y_t=0}^{\pi_{t|t-1}} \\ \nabla_{y_t=0}^{\lambda_{t|t-1}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{t|t-1}} \\ 0 \end{bmatrix} \tag{4-67}$$

$$\nabla_{y_t>0} = \begin{bmatrix} \nabla_{y_t>0}^{\pi_{t|t-1}} \\ \nabla_{y_t>0}^{\lambda_{t|t-1}} \end{bmatrix} = \begin{bmatrix} \frac{-1}{(1-\pi_{t|t-1})} \\ \frac{y_t-\lambda_{t|t-1}}{\lambda_{t|t-1}} - \frac{\exp(-\lambda_{t|t-1})}{1-\exp(-\lambda_{t|t-1})} \end{bmatrix} \tag{4-68}$$

$$\begin{aligned} \mathcal{I}_{t|t-1} &= E_{t-1}\begin{bmatrix} (\nabla^{\pi_{t|t-1}})^2 & \nabla^{\pi_{t|t-1}}\nabla^{\lambda_{t|t-1}} \\ \nabla^{\lambda_{t|t-1}}\nabla^{\pi_{t|t-1}} & (\nabla^{\lambda_{t|t-1}})^2 \end{bmatrix} = \\ &= \begin{bmatrix} \frac{1}{\pi_{t|t-1}(1-\pi_{t|t-1})} & 0 \\ 0 & \frac{(1-\pi_{t|t-1})[1-\exp(-\lambda_{t|t-1})-\lambda_{t|t-1}\exp(-\lambda_{t|t-1})]}{(1-\exp(-\lambda_{t|t-1}))^2\lambda_{t|t-1}} \end{bmatrix} \end{aligned} \tag{4-69}$$

Reparameterizing:

$$\tilde{\pi}_{t|t-1} = h(\pi_{t|t-1}) = \ln\left(\frac{\pi_{t|t-1}}{1-\pi_{t|t-1}}\right) = h_\pi, \quad \tilde{\lambda}_{t|t-1} = h(\lambda_{t|t-1}) = \ln(\lambda_{t|t-1}) = h_\lambda$$

$$\tilde{\nabla}_{y_t=0} = \begin{bmatrix} \tilde{\nabla}_{y_t=0}^{\pi_{t|t-1}} \\ \tilde{\nabla}_{y_t=0}^{\lambda_{t|t-1}} \end{bmatrix} = \dot{H'}^{-1}\nabla_{y_t=0} = \begin{bmatrix} (1-\pi_{t|t-1}) \\ 0 \end{bmatrix} \tag{4-70}$$

$$\tilde{\nabla}_{y_t>0} = \begin{bmatrix} \tilde{\nabla}_{y_t>0}^{\pi_{t|t-1}} \\ \tilde{\nabla}_{y_t>0}^{\lambda_{t|t-1}} \end{bmatrix} = \dot{H'}^{-1}\nabla_{y_t>0} = \begin{bmatrix} -\pi_{t|t-1} \\ y_t - \lambda_{t|t-1} - \frac{\lambda_{t|t-1}\exp(-\lambda_{t|t-1})}{1-\exp(-\lambda_{t|t-1})} \end{bmatrix} \tag{4-71}$$

$$\begin{aligned} \tilde{\mathcal{I}}_{t|t-1} &= \dot{H'}^{-1}\mathcal{I}_{t|t-1}\dot{H}^{-1} = \\ &= \begin{bmatrix} \pi_{t|t-1}(1-\pi_{t|t-1}) & 0 \\ 0 & \frac{\lambda_{t|t-1}(1-\pi_{t|t-1})[1-\exp(-\lambda_{t|t-1})-\lambda_{t|t-1}\exp(-\lambda_{t|t-1})]}{(1-\exp(-\lambda_{t|t-1}))^2} \end{bmatrix} \end{aligned} \tag{4-72}$$

Note that, by setting $d = 0$ in the scaled score, we arrive at an expression for $\tilde{s}_t$ in which $\pi_{t|t-1}$ doesn't appear in $\tilde{\nabla}_{y_t}^{\lambda_{t|t-1}}$ and neither $\lambda_{t|t-1}$ is represented in $\tilde{\nabla}_{y_t}^{\pi_{t|t-1}}$ for $y_t = 0$ or $y_t > 0$. Since this is the case, an optimization in

the estimation procedure can be employed, which makes hurdle models much faster to estimate than zero-inflated distributions: we can fit two independent GAS models and only combine them for forecasting. We can estimate a Bernoulli GAS model and a zero-truncated Poisson (or NB) GAS model in two independent (and lighter) function calls, which can also run in parallel if necessary.

If we set $d \in \{\frac{1}{2}, 1\}$ in the scaled score, a different optimization is possible. Note that only λ_t is influenced by π_t via Fisher Information. Then, we can first estimate the Bernoulli GAS model and supply the fitted $\hat{\pi}_{t|t-1}$ to be used when estimating the zero-truncated GAS model.

4.2.6
Hurdle negative binomial (HNB)

The derivation of a hurdle negative binomial distribution follows the same logic used for HP. We begin by presenting the HNB PMF, its mean and variance:

$$p(y_t|\mu_{t|t-1}, \alpha, \pi_{t|t-1}) = \begin{cases} \pi_{t|t-1} & , \text{if } y_t = 0; \\ \left\{ (1-\pi_{t|t-1}) \frac{\Gamma(y_t+\alpha)}{\Gamma(y_t+1)\Gamma(\alpha)} \left(\frac{1}{1+\frac{\mu_{t|t-1}}{\alpha}} \right)^{\alpha} * \right. & \\ \left. * \left(1 - \frac{1}{1+\frac{\mu_{t|t-1}}{\alpha}} \right)^{y_t} \left(\frac{1}{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \right) \right\} & , \text{else.} \end{cases}$$

$$\mu_{t|t-1},\ \alpha > 0, \quad \pi_{t|t-1} \in (0,1), \quad y_t \geq 0 \tag{4-73}$$

It can be shown that:

$$E_{t-1}[y_t] = \frac{\mu_{t|t-1}(1-\pi_{t|t-1})}{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \tag{4-74}$$

$$Var_{t-1}[y_t] = (1-\pi_{t|t-1}) \left[\frac{\mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})}{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} - \left(1+\frac{\mu_{t|t-1}}{\alpha}\right)^{-\alpha} \left(\frac{\mu_{t|t-1}}{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \right)^2 \right] + \\ + \pi_{t|t-1}(1-\pi_{t|t-1}) \left(\frac{\mu_{t|t-1}}{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \right)^2 \tag{4-75}$$

Now we present the derivation of the scaled score:

$$\ln(p(y_t|\mu_{t|t-1},\alpha,\pi_{t|t-1})) = \begin{cases} \ln(\pi_{t|t-1}) & \text{, if } y_t = 0; \\ \left\{ \begin{array}{l} \ln(1-\pi_{t|t-1}) + \ln(\Gamma(y_t+\alpha)) - \ln(\Gamma(y_t+1)) - \\ -\ln(\Gamma(\alpha)) - \alpha\ln\left(1+\frac{\mu_{t|t-1}}{\alpha}\right) + y_t\ln(\mu_{t|t-1}) - \\ -y_t\ln(\mu_{t|t-1}+\alpha) - \ln\left(1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}\right) \end{array} \right\} & \text{, else;} \end{cases} \tag{4-76}$$

$$\nabla_{y_t=0} = \begin{bmatrix} \nabla_{y_t=0}^{\pi_{t|t-1}} \\ \nabla_{y_t=0}^{\mu_{t|t-1}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{t|t-1}} \\ 0 \end{bmatrix} \tag{4-77}$$

$$\nabla_{y_t>0} = \begin{bmatrix} \nabla_{y_t>0}^{\pi_{t|t-1}} \\ \nabla_{y_t>0}^{\mu_{t|t-1}} \end{bmatrix} = \begin{bmatrix} \frac{-1}{(1-\pi_{t|t-1})} \\ \frac{y_t-\mu_{t|t-1}}{\mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})} - \frac{(1+\frac{\mu_{t|t-1}}{\alpha})^{-(\alpha+1)}}{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \end{bmatrix} \tag{4-78}$$

$$\mathcal{I}_{t|t-1} = E_{t-1}\begin{bmatrix} (\nabla^{\pi_{t|t-1}})^2 & \nabla^{\pi_{t|t-1}}\nabla^{\mu_{t|t-1}} \\ \nabla^{\mu_{t|t-1}}\nabla^{\pi_{t|t-1}} & (\nabla^{\mu_{t|t-1}})^2 \end{bmatrix} = \tag{4-79}$$

where: $(\nabla^{\pi_{t|t-1}})^2 = \dfrac{1}{\pi_{t|t-1}(1-\pi_{t|t-1})}$

$\nabla^{\pi_{t|t-1}}\nabla^{\mu_{t|t-1}} = \nabla^{\mu_{t|t-1}}\nabla^{\pi_{t|t-1}} = 0$

$$(\nabla^{\mu_{t|t-1}})^2 = (1-\pi_{t|t-1})\left[\frac{(1+\frac{\mu_{t|t-1}}{\alpha})^{2\alpha+1} - (1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+1} - \mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha}}{\mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{2\alpha+2} + \mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{2} - 2\mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+2}}\right]$$

And the reparametrizations used to map the parameters into the appropriate subspace will be given by:

$$\tilde{\pi}_{t|t-1} = h(\pi_{t|t-1}) = \ln\left(\frac{\pi_{t|t-1}}{1-\pi_{t|t-1}}\right) = h_\pi, \quad \tilde{\mu}_{t|t-1} = h(\mu_{t|t-1}) = \ln(\mu_{t|t-1}) = h_\mu$$

$$\tilde{\nabla}_{y_t=0} = \begin{bmatrix} \tilde{\nabla}^{\pi_{t|t-1}}_{y_t=0} \\ \tilde{\nabla}^{\mu_{t|t-1}}_{y_t=0} \end{bmatrix} = \dot{H'}^{-1}\nabla_{y_t=0} = \begin{bmatrix} (1-\pi_{t|t-1}) \\ 0 \end{bmatrix} \tag{4-80}$$

$$\tilde{\nabla}_{y_t>0} = \begin{bmatrix} \tilde{\nabla}^{\pi_{t|t-1}}_{y_t>0} \\ \tilde{\nabla}^{\mu_{t|t-1}}_{y_t>0} \end{bmatrix} = \dot{H'}^{-1}\nabla_{y_t>0} = \begin{bmatrix} -\pi_{t|t-1} \\ \frac{y_t-\mu_{t|t-1}}{(1+\frac{\mu_{t|t-1}}{\alpha})} - \frac{\mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{-(\alpha+1)}}{1-(1+\frac{\mu_{t|t-1}}{\alpha})^{-\alpha}} \end{bmatrix} \tag{4-81}$$

$$\tilde{\mathcal{I}}_{t|t-1} = \dot{H'}^{-1}\mathcal{I}_{t|t-1}\dot{H}^{-1} =$$

$$= \begin{bmatrix} h_\pi^{-2}(\nabla^{\pi_{t|t-1}})^2 & h_\pi^{-1}h_\mu^{-1}\nabla^{\pi_{t|t-1}}\nabla^{\mu_{t|t-1}} \\ h_\pi^{-1}h_\mu^{-1}\nabla^{\mu_{t|t-1}}\nabla^{\pi_{t|t-1}} & h_\mu^{-2}(\nabla^{\mu_{t|t-1}})^2 \end{bmatrix} =$$

where: $h_\pi^{-2}(\nabla^{\pi_{t|t-1}})^2 = \pi_{t|t-1}(1-\pi_{t|t-1})$

$$h_\pi^{-1}h_\mu^{-1}\nabla^{\pi_{t|t-1}}\nabla^{\mu_{t|t-1}} = 0$$

$$h_\mu^{-2}(\nabla^{\mu_{t|t-1}})^2 = \mu_{t|t-1}(1-\pi_{t|t-1})\left[\frac{(1+\frac{\mu_{t|t-1}}{\alpha})^{2\alpha+1} - (1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+1} - \mu_{t|t-1}(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha}}{(1+\frac{\mu_{t|t-1}}{\alpha})^{2\alpha+2} + (1+\frac{\mu_{t|t-1}}{\alpha})^2 - 2(1+\frac{\mu_{t|t-1}}{\alpha})^{\alpha+2}}\right] \tag{4-82}$$

4.3 Maximum likelihood estimation

In GAS models, the vector of fixed parameters θ is estimated by maximizing the log-likelihood:

$$\hat{\theta} = \arg\max_\theta \sum_{t=1}^{T} l_t \tag{4-83}$$

$$\text{where } l_t = \ln p(y_t|f_{t|t-1}, \mathcal{F}_{t-1}; \theta) \tag{4-84}$$

The expression for $\ln p(y_t|f_{t|t-1}, \mathcal{F}_{t-1}; \theta)$ is readily available, and its evaluation requires only the calculation of $f_{t|t-1}$ - given by the GAS recursion. To solve the resulting non-linear optimization problem we use standard optimization algorithms, such as BFGS, Nelder-Mead, BHHH and others.

It can be shown (Blasques, Koopman and Lucas [40]) that under regularity conditions, the maximum likelihood estimator $\hat{\theta}$ of θ is consistent and:

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, H^{-1}) \tag{4-85}$$
$$\text{where } H = \lim_{T\to\infty} E\left[\left(\frac{\partial l}{\partial \theta}\right)\left(\frac{\partial l}{\partial \theta'}\right)\right]/T, \text{ and } l = \sum_{t=1}^{T} l_t$$

For our models, the θ vector collects the following parameters: $\{\rho_1, \rho_2, \phi, \kappa, \delta, \beta\}$. Besides these, initial values for states $\{\pi_0, \mu_0, \alpha_0\}$ are also needed. They can either be incorporated in the θ vector and be optimized, or be estimated heuristically in order to avoid the extra computational burden of optimizing eight more parameters. This latter option is employed in this dissertation.

4.3.1 EM algorithm

As mentioned in Lambert [41], zero-inflated distributions require a further modification in the estimation procedure. For both ZIP and ZINB, when we observe $y_t = 0$, we cannot distinguish if the zero arises from the Bernoulli trial or the Poisson/NB distribution.

If we knew that the observed zero arises from the Bernoulli trial, we could rewrite the likelihood in a more tractable form. Suppose we could observe a variable Z_t, that is $Z_t = 1$ if the observed zero is from the Bernoulli trial, and $Z_t = 0$ otherwise. Then we can express the likelihood for both distributions in a different form. In the ZIP case we would write:

$$\ln(p(y_t|\lambda_{t|t-1}, \pi_{t|t-1})) = Z_t \ln(\pi_{t|t-1}) + (1 - Z_t)\{-\lambda_{t|t-1} + y_t \ln(\lambda_{t|t-1}) - \ln(y_t!)\} \tag{4-86}$$
$$\text{for observation } t$$

Through the EM algorithm (Dempster, Laird, Rubin [42]) we iteratively estimate Z_t given the current estimates of $\hat{\theta}^{k-1}$ (E step) and then, keeping $\hat{Z}_t^k$ fixed, we maximize the rewritten likelihood to find $\hat{\theta}^k$ (M step). We iterate these steps until convergence.

Z_t is estimated through its expected value, given by:

$$\hat{Z}_t^k = E[\pi_{t|t-1} = 1|y_t, \hat{\theta}^{k-1}] = \\ = \frac{P[y_t = 0|\pi_{t|t-1} = 1]P[\pi_{t|t-1} = 1]}{P[y_t = 0|\pi_{t|t-1} = 1]P[\pi_{t|t-1} = 1] + P[y_t = 0|\pi_{t|t-1} = 0]P[\pi_{t|t-1} = 0]} \tag{4-87}$$

In the case of ZIP distribution, this simplifies to:

$$= \begin{cases} \frac{\pi_{t|t-1}}{\pi_{t|t-1}+(1-\pi_{t|t-1})\exp(-\lambda_{t|t-1})} & \text{, if } y_t = 0; \\ 0 & \text{, else.} \end{cases} \tag{4-88}$$

The estimation of $\hat{\theta}^k$ works as before, with the use of standard optimization routines.

To assess if the EM algorithm would produce different estimates from that obtained by directly maximizing the unmodified likelihood, we ran a small experiment: for ZIP and ZINB distributions, we estimated the same model presented in Chapter 5 in the dataset used in our application and compared the estimates for both procedures (EM algorithm and unmodified likelihood). The maximum difference of parameter estimates in all 752 time series was not larger than 10%. This being the case, we opted to work with the unmodified likelihood, since this is the fastest routine.

4.4 Initialization

This section describes the procedures employed for initializing both the elements of the θ vector in the iterative optimization algorithm, and the initial values of the components used to describe the dynamic of the time varying parameters, in which we condition the optimization of the model.

For some parameters in θ, fixed initial values were used. These were defined based on experimentations with the dataset discussed in Chapter 6.

The parameters associated with the scaled scores (ρ_1, ρ_2, κ) are initialized with the value 0.1. The initial value chosen for ϕ is 0.5. A frequently adequate initial value for α in NB distributions is 5, which we use for all GAS models containing a NB component and also for the intermittent demand models presented in Chapter 3.

For the distributions that involve the π process, we estimate an AR(1) model for an indicator variable that is equal to one if $y = 0$ and zero otherwise. The estimated intercept is used as initial value for δ and the AR coefficient is

used for β. π_0 is initialized as the unconditional mean of the AR(1) process, that is, $\pi_0 = \frac{\delta}{1-\beta}$.

The initialization of μ_0 and α_0 was inspired by an analogous procedure described in Hyndman et al. [43]. The adaptation employed here is defined in the sequel:

1. Compute a moving average of order 3 with the first four weeks of data, call this $\{r_t\}, t = 1, \ldots, 26$. This is the first estimate of the level - as we have mentioned previously, the series employed here do not exhibit trend.

2. Set $\mu_0 = \frac{1}{26}\sum_{t=1}^{26} r_t$.

3. Extract the seasonal component δ_t of the original series. We set $\delta_t = y_t - r_t$ using the first 26 observations of y_t and r_t.

4. Fit a constrained least squares (with no intercept) to the δ_t series to estimate the elements of α_0. The imposed restriction is that $\mathbb{I}'\alpha_0 = 0$, where $\mathbb{I}$ is a vector of ones with appropriate dimension.

5. Set α_0 equal to the parameters estimated in the step above.

We run this procedure in $\ln(1+y)$ to mimic the behaviour of the log link function.

4.5 Explanatory variables

The inclusion of regressors in the GAS framework is straightforward. Let X_t be a $T \times k$ matrix of explanatory variables. Let ζ be the vector of the k coefficients to be estimated. Then we can include the explanatory variables as:

$$h_f(f_{t|t-1}) = \mu_{t|t-1} + \gamma_{t|t-1} + \zeta' X_t \tag{4-89}$$

With $\mu_{t|t-1}$ and $\gamma_{t|t-1}$ as defined in equations 4-4 up to 4-6. This represents a small modification in equation 4-3 that does not affect the scaled scores defined in Section 4.2. The coefficients in ζ need to be incorporated in θ to be estimated.

It is also possible to include regressors in the $\pi_{t|t-1}$ dynamic in a similar manner, but we choose not to do so in this work.

An issue that arises in applications is the choice of the appropriate subset of variables of X_t that has to be included in the final model. We perform

variable selection via adaptive Lasso, from Zou [44]. This procedure has the oracle property - that is, under some specified conditions, as the sample size grows, the probability that the adaptive Lasso selects the correct subset of X_t, the "true" model, converges to 1.

The adaptive Lasso is an extension of Lasso [45] - a l_1 penalized regression - that adds a weight to each coefficient to be estimated. The variable selection and estimation in adaptive Lasso is performed by solving the following optimization problem:

$$\arg\min_{\beta} \|\boldsymbol{y} - \sum_{j=1}^{k} \boldsymbol{x_j}\beta_j\|^2 + \lambda \sum_{j=1}^{k} w_j \beta_j \tag{4-90}$$

With $\boldsymbol{y}$ and $\boldsymbol{x}$ being vectors of length T, for a given choice of λ and previously selected weights w.

The implementation of the variable selection procedure in this study - that is the same to all models with external regressors - is specified bellow:

1. The dependent variable is $\ln(1+y)$.
   - We do so to mimic the behaviour of the log link function presented in the discussion of the reparametrizations.
2. Given all the variables to be selected (the columns of X_t), we add to these the following extra variables: lag 1 and lag 7 of the dependent variable and seasonal dummies.
   - The objective is to insert simple time series structure in the variable selection problem.
   - Neither the lags of $\ln(1+y_t)$ nor the seasonal dummies are subject to penalization. That is, for these variables we set $w = 0$.
3. The weights in w are selected via Lasso as a first stage without penalizing the time series structure added in the step above. We define $w_j = \frac{1}{|\beta_j^{Lasso}| + \frac{1}{\sqrt{T}}}$.
4. Estimate the adaptive Lasso coefficients with the above defined weights, again with no penalization for the time series structure embedded in the model.
   - For both the first stage Lasso and the adaptive Lasso, λ is selected by minimizing BIC (Bayesian information criterion).

5. The resulting variables with coefficients different from zero are selected to compose the final model.

These selected variables are the ones to be included in the GAS dynamic. It is also necessary to determine starting values for ζ when performing maximum likelihood estimation. The proposed solution for initialization is to run a standard GLM regression (with Poisson or NB distribution) for y using a simple time series structure as described in the topic of variables selected via Adaptive Lasso. The coefficients estimated at this stage are inserted as starting values for the optimization algorithm used for estimation.

4.6 Diagnostics

Regarding model diagnostics, when dealing with non-Gaussian distributions there are several options of residuals suitable for analysis. Hilbe [34] discusses possible choices for count data models. Here we opt to work with the randomized quantile residuals, proposed in Dunn & Smyth [46]. This method provides an adaptation of the quantile residuals, typically used in GLM diagnostics, that is suitable for discrete distributions. We define the randomized quantile residuals in the sequel:

$$\begin{aligned} rq_t &= \Phi^{-1}(u_t) \\ \text{where: } u_t &\sim U[\hat{F}_t(y_t - 1), \hat{F}_t(y_t)] \\ \text{and } \hat{F}_t &= \sum_{i=0}^{y_t} p(y_i|\hat{f}_{t|t-1}; \mathcal{F}_{t-1}; \hat{\theta}) \end{aligned} \tag{4-91}$$

With $\Phi(.)$ being the cumulative distribution function of a standard normal variable.

For continuous distributions, r_t (as defined below) is distributed as a U[0,1]:

$$r_t = F_t^*(y) = \int_{-\infty}^{y_t} p^*(y|\hat{f}_{t|t-1}; \mathcal{F}_{t-1}; \hat{\theta})dy$$

Where f_t^* is the density of y_t. This result also holds if we are able to consistently estimate the parameters of f_t^*. u_t is an adaptation of the the probability integral transform that makes the cumulative distribution of a discrete variable also uniformly distributed in the unit interval.

Kolassa [32] employs u_t itself, which is called randomized probability integral transform in that paper, as a model diagnostic. We follow [46] and further transform this variable, via $\Phi^{-1}(.)$, to assess if our models are correctly specified.

We will employ conventional tests for Gaussian time series data to analyse the fitted models. Specifically, we analyse the adequacy of the chosen distributions via the Jarque-Bera test, and use the Ljung-Box test for auto-correlation and conditional heteroskedasticity for remaining unmodeled time series structure.

To validate the model diagnostic choice, Chapter 5 studies the behaviour of the randomized quantile residuals for hurdle and zero-inflated GAS models under correct specification.

4.7 Forecasting

Given the chosen distribution for y_t, we have the entire predictive distribution for the first-step-ahead, since $\pi_{t|t-1}$ and $f_{t|t-1}$ are fully determined given observations up to t and the estimated parameters $\hat{\theta}$. As is typical with non-Gaussian models, for further steps-ahead there is no closed form expression for the distribution. As an example, we show how to obtain the second step ahead predictive distribution:

$$p(y_{t+2}|f_{t|t-1}, \pi_{t|t-1}, \mathcal{F}_{t-1}; \hat{\theta}) = \sum_{0}^{\infty} p(y_{t+2}|f_{t+1|t}, \pi_{t+1|t}, \mathcal{F}_t; \hat{\theta}) p(y_{t+1}|f_{t|t-1}, \pi_{t|t-1}, \mathcal{F}_{t-1}; \hat{\theta}) \tag{4-92}$$

Although $p(y_{t+1}|f_{t|t-1}, \pi_{t|t-1}, \mathcal{F}_{t-1}; \hat{\theta})$ has a known closed form, it is not possible to ensure that the same is true for the product $p(y_{t+2}|f_{t+1|t}, \pi_{t+1|t}, \mathcal{F}_t; \hat{\theta}) p(y_{t+1}|f_{t|t-1}, \pi_{t|t-1}, \mathcal{F}_{t-1}; \hat{\theta})$, and so we need a Monte Carlo algorithm to estimate the k-step-ahead predictive distribution $p(y_{t+k}|f_{t|t-1}, \pi_{t|t-1}, \mathcal{F}_{t-1}; \hat{\theta})$. We do so using the following routine:

arguments: m, k
input : $\mu_{t|t-1}$, $\alpha_{t|t-1}$, $\pi_{t|t-1}$, $\tilde{s}_{t|t-1}$, $\hat{\theta}$
output : A $m \times k$ matrix with m simulations of the k-step ahead predictive distribution

for $i \leftarrow 1$ **to** m **do**
  Initialize $\mu, \alpha, \pi, \tilde{s}$ with the given inputs;
  for $j \leftarrow 1$ **to** k **do**
    Given $\hat{\theta}$ and current states (μ, α, π), sample $y_{t+j}^{(i)}$;
    Store $y_{t+j}^{(i)}$ on the relevant slot of the output matrix;
    Calculate $\tilde{s}$;
    Update the states (μ, α, π);
  end
end

Algorithm 1: k-step-ahead predictive distribution for GAS model

It follows that each line of the returned matrix is an estimate of the k-step-ahead predictive distribution.

5
Simulation studies

In this chapter the finite sample properties of the maximum likelihood (ML) estimator for GAS models with HP, HNB, ZIP and ZINB distributions are investigated. We also check the adequacy of the randomized quantile residuals for our models with excess of zeroes under correct specification. To address these issues, we begin the chapter with a description of the experimental setup. After that, we analyse the results of our simulation studies and discuss the findings, first for parameter estimation, and then for model residuals.

5.1
Setup

To assess the performance of the ML estimator and randomized quantile residuals of the fitted models, we considered the following data generating process:

$$y_t \sim p(y_t|f_{t|t-1}, \pi_{t|t-1}, \mathcal{F}_{t-1}; \theta) \tag{5-1}$$

$$h_f(f_{t|t-1}) = h_f(f_{t-1|t-2}) + \kappa_f \tilde{\nabla}_{f_{t-1|t-2}}, \quad \kappa_f > 0 \tag{5-2}$$

$$h_\pi(\pi_{t|t-1}) = \delta + \beta h_\pi(\pi_{t-1|t-2}) + \kappa_\pi \tilde{\nabla}_{\pi_{t-1|t-2}}, \quad |\beta| < 1, \quad \kappa_\pi > 0 \tag{5-3}$$

With $p(y_t|f_{t|t-1}, \pi_{t|t-1}, \mathcal{F}_{t-1}; \theta)$ being the mixture distributions previously mentioned - see Sections 4.2.3 up to 4.2.6. In both ZIP and HP cases, $\lambda_{t|t-1} = f_{t|t-1}$ and for the ZINB and HNB cases, $\mu_{t|t-1} = f_{t|t-1}$. Note that we use $d = 0$ in the scaled score to mimic the choice made for the application in real data as we have discussed in Section 4.1.

We evaluate the ML estimator and model residuals for time series with sample sizes $T = \{250, 500, 1000\}$. For each sample size, we simulate the time series and estimate the parameters 500 times, store the estimation results and estimate the residuals as in equation 4-91 for later analysis. The parameters associated with the scores are kept fixed at $\kappa_f = 0.01$ and $\kappa_\pi = 2.25$. For the $\pi_{t|t-1}$ variable, we also fix $\delta = 0.6$ and $\beta = 0.2$. When analysing the ZINB and HNB distributions, we set $\alpha = 4$.

5.2 Results

5.2.1 Parameter estimators

Once obtained the parameter estimates, we report the mean of the estimates, the root mean square error (RMSE) and the standard error (SE), and the mean of the asymptotic standard error (ASE) of each ML estimate. We also provide p-values for the Jarque-Bera test used to assess if the distribution of the parameter estimates is Gaussian and provide histograms of the these. We group Poisson-type distributions in one table and NB-type in other.

We begin the discussion of the results with HP and ZIP distributions. Table 5.1 gathers the estimated metrics for these and Figures 5.1 and 5.2 plot the standardized distributions of the parameter estimates along with a $N(0, 1)$ distribution in blue.

Concerning the bias of the estimates we note that, from the beginning, all parameters exhibit small bias, with the largest estimated bias being from β under ZIP distribution - around 12%. We note that this sample size ($T = 250$) is actually smaller than the one being used to estimate our models in our empirical application in Chapter 6 ($T = 365$, as we mention in 6.2.1). As sample sizes grows, the bias decreases, achieving negligible values for all estimated parameters under $T = 1000$.

With few exceptions, RMSE and SE estimates exactly match. If the asymptotic standard error of the ML estimators was a good estimate of the parameters true standard errors, we would observe its estimate being very close to the observed estimates for RMSE and SE. This is not the case for any of the parameters being studied.

Concerning the Jarque-Bera test for the ML estimators, we note that with sample size $T = 1000$, we do not reject the null hypothesis of normality at a 1% significance level for any parameter studied. In some cases, we do not reject the null hypothesis of normality for some of the parameters with smaller sample sizes.

Table 5.1: Estimated parameters for HP and ZIP GAS models.

	HP				ZIP			
	$\kappa_f = 0.01$	$\delta = 0.6$	$\beta = 0.2$	$\kappa_\pi = 2.25$	$\kappa_f = 0.01$	$\delta = 0.6$	$\beta = 0.2$	$\kappa_\pi = 2.25$
T = 250								
Mean	0.011	0.629	0.197	2.299	0.009	0.616	0.176	2.291
RMSE	0.005	0.225	0.150	0.372	0.009	0.234	0.157	0.384
SE	0.005	0.223	0.150	0.369	0.008	0.234	0.155	0.382
ASE	0.000	0.014	0.009	0.024	0.000	0.014	0.010	0.024
Normality	0.004	0.000	0.929	0.000	0.000	0.000	0.230	0.000
T = 500								
Mean	0.010	0.626	0.204	2.284	0.009	0.601	0.200	2.280
RMSE	0.004	0.152	0.097	0.245	0.006	0.155	0.114	0.260
SE	0.004	0.150	0.097	0.243	0.006	0.155	0.114	0.259
ASE	0.000	0.007	0.005	0.012	0.000	0.007	0.005	0.012
Normality	0.055	0.015	0.061	0.000	0.000	0.000	0.704	0.279
T = 1000								
Mean	0.010	0.606	0.204	2.254	0.010	0.601	0.200	2.260
RMSE	0.003	0.106	0.070	0.189	0.004	0.105	0.077	0.174
SE	0.003	0.106	0.070	0.189	0.004	0.105	0.077	0.174
ASE	0.000	0.003	0.002	0.006	0.000	0.003	0.003	0.006
Normality	0.092	0.313	0.484	0.015	0.076	0.836	0.055	0.154

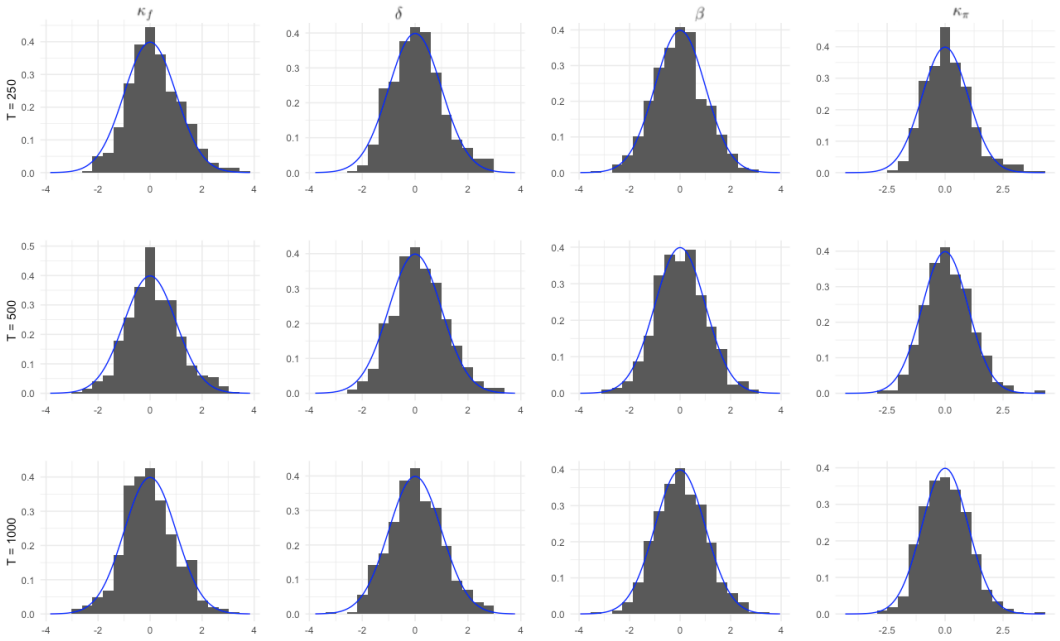

Figure 5.1: Histograms of parameter estimates - HP distribution.

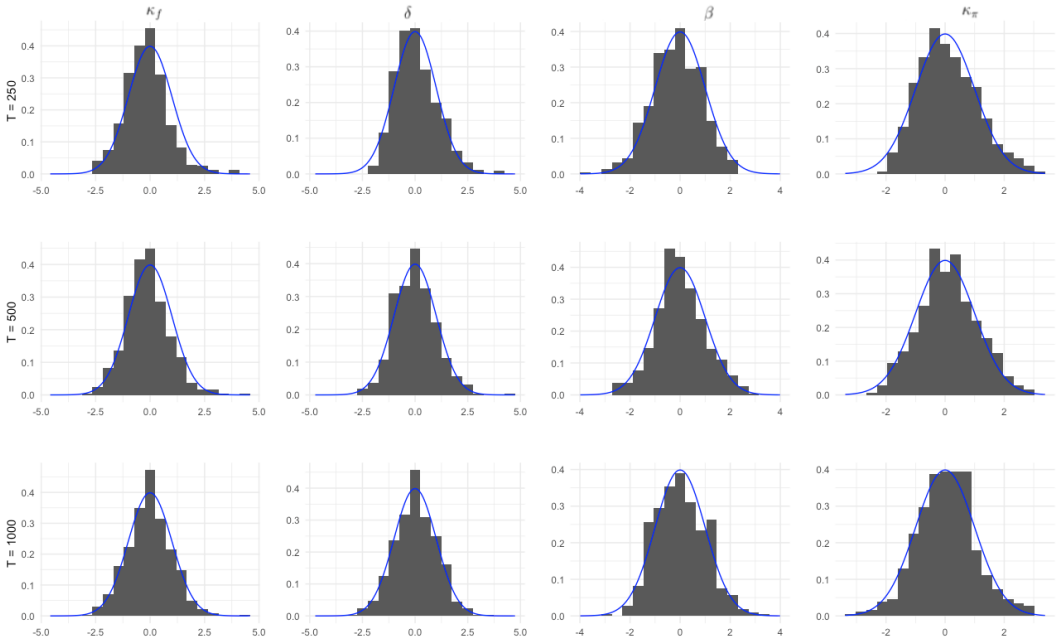

Figure 5.2: Histograms of parameter estimates - ZIP distribution.

The findings obtained for HP and ZIP distributions are also valid for HNB and ZINB. Table 5.2 reports the metrics being analyzed and Figures 5.3 and 5.4 present the histograms of parameter estimates.

We observe a constantly decreasing bias, which becomes negligible with the largest sample size. RMSE and SE show very close estimated values, and ASE is again an inadequate estimate of the true parameter dispersion. We do not reject the null hypothesis of normality for any parameter at 1% significance level for $T = 1000$.

The extra parameter α, which is related to the overdispersion in the NB distribution, has precise estimates, with only a small bias for the sample size of 250 observations. In all other cases, the estimated bias is negligible.

Our results indicate that the ML estimator is consistent and asymptotically normal for the parameters of the excessive zeroes distributions with dynamic Bernoulli process proposed in this dissertation. We also conclude that the asymptotic standard errors (ASE) doesn't provide adequate estimates of true dispersion of the parameters studied. The estimated values are much smaller than the standard errors estimated from the distribution of ML estimators. It seems that a correction for finite samples is needed for proper inference.

Table 5.2: Estimated parameters for HNB and ZINB GAS models.

	HNB				ZINB			
	$\kappa_f = 0.01$	$\delta = 0.6$	$\beta = 0.2$	$\kappa_\pi = 2.25$	$\kappa_f = 0.01$	$\delta = 0.6$	$\beta = 0.2$	$\kappa_\pi = 2.25$	$\alpha = 4$
T = 250									
Mean	0.011	0.640	0.192	2.293	4.151	0.008	0.614	0.190	2.264	4.283
RMSE	0.013	0.220	0.151	0.401	0.899	0.011	0.213	0.145	0.344	0.895
SE	0.013	0.217	0.150	0.399	0.887	0.010	0.213	0.145	0.344	0.850
ASE	0.001	0.014	0.010	0.023	0.055	0.001	0.014	0.009	0.023	0.049
Normality	0.000	0.000	0.020	0.000	0.000	0.000	0.000	0.073	0.003	0.000
T = 500									
Mean	0.011	0.602	0.205	2.272	4.055	0.010	0.603	0.194	2.255	4.106
RMSE	0.010	0.150	0.109	0.264	0.591	0.006	0.155	0.108	0.249	0.541
SE	0.009	0.150	0.109	0.264	0.589	0.006	0.155	0.108	0.249	0.531
ASE	0.000	0.007	0.005	0.011	0.026	0.000	0.007	0.005	0.011	0.023
Normality	0.000	0.028	0.704	0.011	0.000	0.066	0.000	0.506	0.022	0.000
T = 1000									
Mean	0.010	0.601	0.197	2.268	4.043	0.010	0.598	0.201	2.252	4.027
RMSE	0.005	0.108	0.072	0.187	0.398	0.004	0.104	0.075	0.192	0.355
SE	0.005	0.108	0.072	0.187	0.396	0.004	0.104	0.075	0.192	0.354
ASE	0.000	0.003	0.002	0.006	0.013	0.000	0.003	0.002	0.006	0.011
Normality	0.682	0.284	0.446	0.300	0.082	0.064	0.018	0.560	0.012	0.024

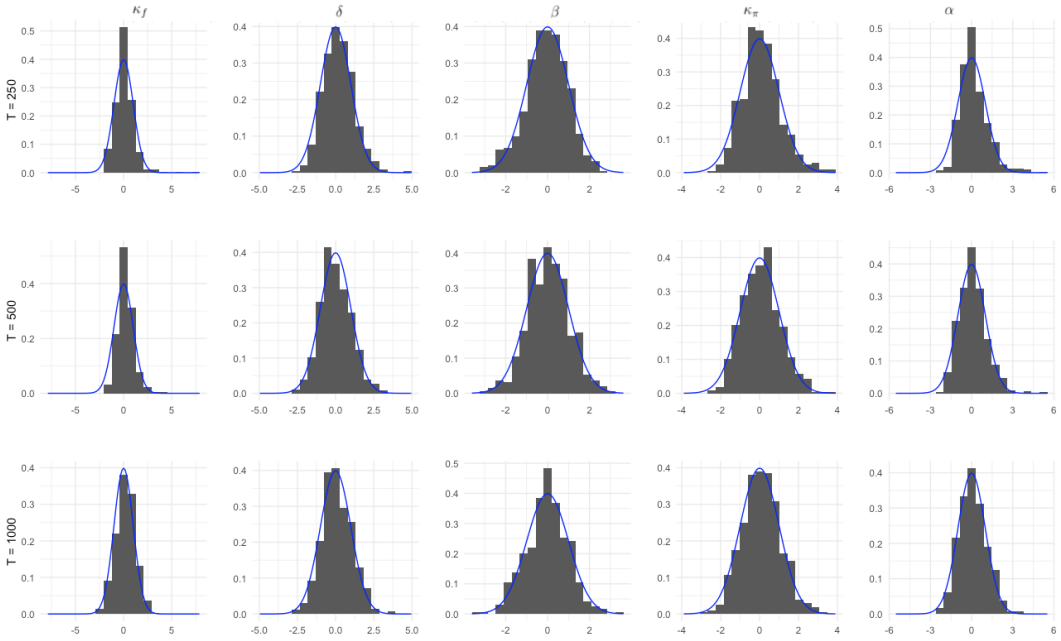

Figure 5.3: Histograms of parameter estimates - HNB distribution.

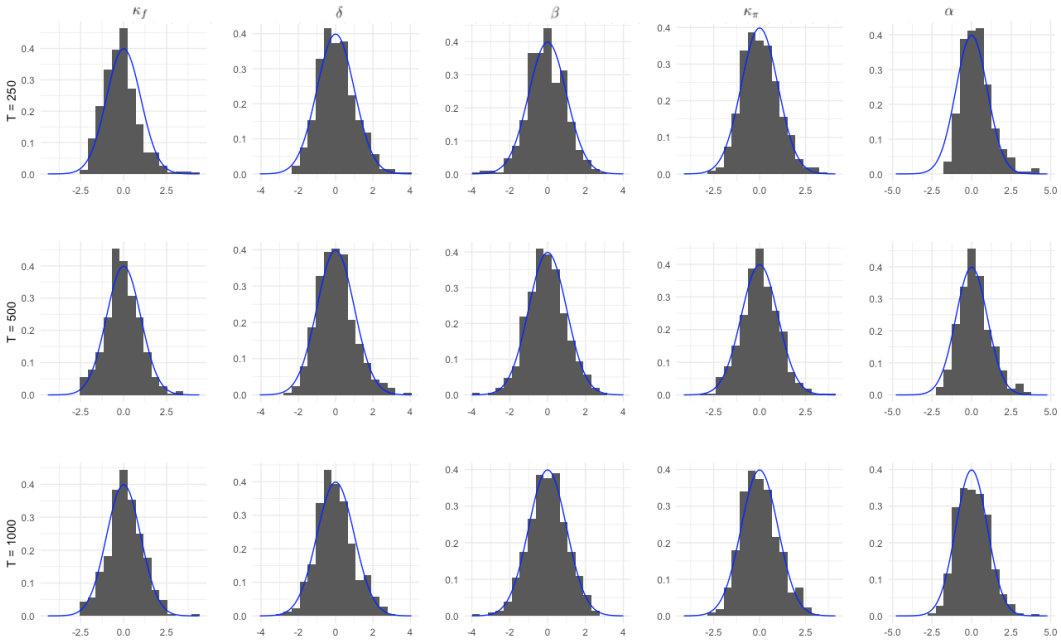

Figure 5.4: Histograms of parameter estimates - ZINB distribution.

5.2.2 Randomized quantile residuals

Having established, via our simulation study, the consistency and asymptotic normality of the ML estimators, we now turn our analysis to the randomized quantile residuals. It is expected that, aside from sample variability, the randomized quantile residuals should be uncorrelated, homoscedastic and follow a Gaussian distribution - as is the case for "standard" quantile residuals under correct specification and consistent estimation method.

This is indeed the result we find. The following tables exhibit, for significance levels $\alpha = \{10\%, 5\%, 1\%\}$ and sample sizes $T = \{250, 500, 1000\}$, the percentage of null hypotheses rejections according to the distribution being studied.

The first table presents the Jarque-Bera test used to assess if the distribution of the randomized quantile residuals is Gaussian. After that, we present the results for Ljung-Box test for auto-correlation, and Ljung-Box test for conditional heteroskedasticity.

Table 5.3: Percentage of rejections - Jarque-Bera test.

	ZIP	HP	ZINB	HNB
T = 250				
$\alpha = 10\%$	7.2%	7.4%	6%	5.2%
$\alpha = 5\%$	3.8%	4%	3.4%	3.2%
$\alpha = 1\%$	1%	2%	0.6%	1.6%
T = 500				
$\alpha = 10\%$	8.4%	6.2%	3.6%	6%
$\alpha = 5\%$	4.8%	3.2%	2.4%	3.2%
$\alpha = 1\%$	2.2%	1.2%	0.2%	1.8%
T = 1000				
$\alpha = 10\%$	8.2%	9.4%	6.8%	4.6%
$\alpha = 5\%$	3.4%	5.6%	3.2%	1.8%
$\alpha = 1\%$	0.2%	1.8%	1%	0.2%

The table above highlights that, for most cases, we have slightly less rejections of the null hypothesis of the Jarque-Bera test than the significance level being used.

The two next tables present the results of Ljung-Box tests with different numbers of lags used to evaluate the null hypotheses. We vary the number

of lags from 1 up to 4 to calculate the test statistic. These numbers seem adequate, since the simulated series do not exhibit a seasonal component.

First, for auto-correlation of the residuals, we note that the number of rejections is also smaller than the significance level specified. This is true for every lag order, sample size and distribution studied. Concerning conditional heteroskedasticity, the percentage of null hypothesis rejections is closer to the significance levels specified, specially for $k = 1$.

Under correct specification, the randomized quantile residuals exhibit the expected properties for the proposed GAS models with excess of zeroes.

Table 5.4: Percentage of rejections - Ljung-Box test for auto-correlation.

	T = 250				T = 500				T = 1000			
	ZIP	HP	ZINB	HNB	ZIP	HP	ZINB	HNB	ZIP	HP	ZINB	HNB
α= 10%												
k = 1	7%	6%	5.8%	4.6%	6%	5.8%	3.6%	4.8%	8%	2.8%	6.8%	4.6%
k = 2	5%	6%	4.8%	3.6%	5%	4.8%	4%	5%	5%	2%	4.6%	4.8%
k = 3	5.8%	7.4%	6%	4.8%	5.6%	5.2%	4.6%	5.8%	5%	4.4%	5%	5.8%
k = 4	6.8%	7.4%	6.6%	5.4%	5.8%	6.6%	6.4%	7.4%	6.4%	5.4%	5.6%	7.4%
α= 5%												
k = 1	2.8%	2.4%	2%	2%	3%	2%	1.8%	1.6%	2.8%	1.4%	2.4%	3.2%
k = 2	2.4%	2.6%	1.8%	1.2%	1.8%	2.4%	1.6%	2.2%	2%	0.6%	1.6%	1.8%
k = 3	2.8%	3.4%	3.2%	2.8%	2.4%	2.8%	2.2%	2.4%	2.4%	2.4%	1.8%	2.8%
k = 4	3.4%	2.8%	3.4%	2.8%	2.2%	3.6%	2.6%	2.4%	2.8%	2.6%	1.8%	4%
α= 1%												
k = 1	1%	0.4%	0.4%	0%	0%	0.2%	0.6%	0.2%	0.6%	0%	0.2%	0.6%
k = 2	0.6%	0.2%	0.6%	0%	0.2%	0.4%	0%	0.2%	0.4%	0%	0%	0.2%
k = 3	0.6%	0.8%	1%	0.2%	0.2%	0.4%	0%	0.2%	0.2%	0.2%	0%	0.6%
k = 4	0.4%	0.8%	1%	0.2%	0.4%	0.4%	0.4%	0.2%	0.4%	0.4%	0%	1%

Table 5.5: Percentage of rejections - Ljung-Box test for conditional heteroskedasticity.

	T = 250				T = 500				T = 1000			
	ZIP	HP	ZINB	HNB	ZIP	HP	ZINB	HNB	ZIP	HP	ZINB	HNB
α= 10%												
k = 1	10.4%	10.2%	8.8%	10.4%	10.2%	6.6%	9.8%	7.8%	9.2%	9.4%	9%	10%
k = 2	8.6%	9.6%	6.2%	7.4%	8.2%	6.4%	7.4%	8%	10.8%	10.4%	10%	10%
k = 3	9%	7.4%	6.2%	9%	9.2%	7.4%	7.6%	9.2%	10%	8.4%	8.4%	9%
k = 4	9%	7.8%	7.2%	9.6%	9.4%	8.2%	8%	9%	12%	8.4%	8.2%	9.8%
α= 5%												
k = 1	5%	4.4%	3.6%	5%	5%	3.2%	4.2%	3.4%	5%	4.4%	4.4%	5.8%
k = 2	3.2%	4.4%	3.2%	4.2%	4.4%	3.2%	4.2%	4.4%	4.8%	4.4%	5%	4.6%
k = 3	5%	4.8%	3.4%	5.6%	5.2%	4%	4%	4.8%	7.2%	4.2%	3.4%	5%
k = 4	5.4%	3%	4.8%	5.8%	5%	5%	4%	3.2%	5.8%	4%	4.6%	3.8%
α= 1%												
k = 1	1.4%	0.6%	1%	1.4%	0.8%	0.4%	0.6%	0.6%	1.2%	0.8%	1%	1%
k = 2	1.4%	0.6%	0.4%	0.8%	1.4%	0.6%	0.4%	0.6%	1%	0.8%	1.2%	0.6%
k = 3	1.4%	0.4%	0.8%	1.4%	1.4%	0.6%	0.6%	0.8%	1.2%	1.4%	0.8%	0.6%
k = 4	1.4%	0.4%	1.2%	1.2%	1.6%	0.8%	0.6%	1.6%	1%	1%	0.8%	1.6%

6
Application

In this chapter we compare our proposed GAS models with the benchmarks from intermittent demand forecasting and count time series literatures. For this, we use real intermittent demand time series obtained from a large Brazilian retailer. The models are compared with metrics suitable for both point and distribution forecasts.

We begin the chapter with the description of the data being analysed. This is followed by a discussion of the setup of our model comparison exercise and the presentation of descriptive statistics of the data used in our comparison. After that, we analyse the models estimated and present its diagnostics. The chapter is concluded with the presentation of the results.

6.1
Data and filters

To test the proposed GAS models, we had access to real daily sales time series from a big Brazilian retailer, a company with more than 1700 stores over the country and presence in all states. The daily data used in this dissertation comes from one particular state, from which we had access to all stores, with observations starting in January 1st 2017 up to the last day of 2019.

The specific SKUs that are analysed come from two different lines of products: hair colour and small appliances. SKU is the most disaggregated unit measure of a product. It represents, for instance, a T-shirt from a certain brand, along with its colour, size and model. This is the relevant unit for inventory management purposes. In theory, prices could be set by SKU, but most typically these are group specif - as is the case for our data. Continuing with the aforementioned example, every T-shirt form the same brand and model has the same price, regardless of colour and size. This common price definition is not taken into account in our models as we do not leverage information across different time series.

Besides daily demand, other variables were also made available by the retailer: daily prices, number of products in stock, a dummy variable indicating if the store is opened and promotions calendar. To these, we added weekdays (for the models with deterministic seasonality) and both national and local

holidays. All of these are used as explanatory variables in all models. For the forecasting horizons being analysed it is reasonable to assume that their values are known in advance, since the ones provided by the company are part of a regular planning routine. The exception is the number of products in stock. In this case, when making predictions, the last observed values are kept fixed for forecasting.

We add holidays to our variables since these have different effects on different kinds of stores: a group of stores may not open on holidays, others have restricted working hours, and some do not have any change in schedule. Even in this case it is important to control for unusual behaviour in costumer flow. With regard to promotions effects, they are extremely variable, since each promotion category may have its own advertising channel, such as TV, newspapers or mobile app notification, for example. We control for promotions in our models, adding up to 10 promotion variables.

Some treatments were needed in order to select the series to be used in our forecasting comparison exercise. The first operation performed was the exclusion of all time series coming from recently opened stores. In the company that provided the data there is a general perception that recently opened stores take about two years to stabilize sales. Because of that, we have chosen to work only with stores older than two years at the beginning of 2017. This explains why we have chosen a stationary level component for our models (equation 4-4).

Other operation performed in our data concerns the removal of two weeks in mid-2018. Brazil faced a general strike of truckers that interrupted supply all over the country. To deal with this unexpected event, we chose to remove the observations during the strike.

We have opted to work with a rolling window scheme for model estimation and forecasting. The year of 2019 was used as our out-of-sample period and a window of size 365 was employed for model fitting.

Not all series have 365 observations available for the first estimation window in 2019. These series were used only when all 365 in-sample observations became available for model estimation. Because of this, the number of out-of-sample observations at our disposal for model comparison is not fixed. In the case of 1-day-ahead forecasts, this translates to a maximum of 365 observations (all year of 2019) used when comparing models, and a minimum of 31 days (only December 2019) in the worst case.

Other criteria used to select which time series to use in our forecasting exercise were the following: we did not use series that exhibit zero sales in 2019; the series used had at least a variance of 0.5; we did not work with SKU/stores

that had more than 60 days with no products in stock (not necessarily in sequence).

After applying all this filters, we are left with a total of 752 daily time series from hair colour and small appliances lines of products, totaling 8285 windows - between 11 and 12 estimation rounds for each one of the 752 time series studied.

6.2 Setup

The experimental setup is now discussed. We provide details about the estimation and prediction procedures and the accuracy measures are shown.

6.2.1 Estimation and prediction

We now describe some decisions concerning the frequency of model reestimation, days at which we evaluate our predictions and number of simulations ran to calculate the k-days-ahead predictive distribution.

As mentioned in Section 6.1, we estimate our models with a fixed window of 365 observations (after applying the described filters). As we have a maximum of one year to evaluate the forecasts, it is important to reestimate the models when possible to update the parameters estimates. We chose to do so on a monthly basis. That is, before predicting the first day of each month in 2019 we reestimate our models using a rolling window and evaluate our predictions at 1, 8, 15-days-ahead.

The benchmark models described in Chapter 3 are estimated incorporating a deterministic seasonality and the first lag of the dependent variable. All other variables are subject to selection. As we employ the same selection procedure for the explanatory variables at each reestimation period, all models for the same time series and month of estimation have the same explanatory variables.

To evaluate the 1, 8, 15-days-ahead predictive distributions, we ran 1000 simulations to calculate the metrics discussed in the next section - $m = 1000, k = 15$ in Algorithm 1. For point forecasts, we calculate the mean of the simulated values for each step ahead when evaluating RMSSE, and use the median of the simulations for MASE. Both of this metrics are defined in the section that follows.

Some particular days are excluded from the calculation of the metrics: days in which the store is not opened, Black Friday and holidays. In this last case we also exclude the day before and the day after the event. For

the Black Friday, this particular retailer has 4 days of promotions. In this case, excluding just the days of promotions is sufficient to account for atypical behaviour. Holidays and Black Fridays are not evaluated since we have too few observations in-sample to make a reasonable prediction. Excluding holidays (and the days around it) and Black Friday reduces our out-of-sample period by 58 observations. Days of closed stores vary, but on average we have the stores opened at 95% of the remaining days. So that from the year of 2019, we utilize around 290 days to evaluate the forecasts.

6.2.2 Accuracy measures

The accuracy measures utilized in this dissertation evaluate both the adequacy of the predictive distribution, and the performance of point predictions. For the former, we follow the discussion in Kolassa [32] and evaluate the accuracy of the predictive distributions via proper scoring rules. Czado, Gneiting, and Held [33] also argument that these should be the forecasting accuracy metrics to be analysed in a count data context. Both works present some possible scoring rules from which we choose Brier and spherical scoring rules to compare the models. Concerning point predictions, we use Mean Absolute Scaled Error (MASE), proposed in Hyndman and Koehler [47], a scale-free error measure most suitable for comparing models across series with different scales. We also use a variation of this metric, called Root Mean Squared Scaled Error (RMSSE), proposed in the M5 competition of 2020.

First, it is important to recall the definition of each of these metrics. We begin with scoring rules. As defined in [32]:

"A scoring rule is a function s that maps a predictive distribution $\hat{p}$ and a single realization y to a penalty value $s(\hat{p}, y)$. In practice, one usually reports averages of the scores over suitable pairs $(\hat{p}, y)$, e.g., over forecasts and actuals over multiple time points t,

$$S := \frac{1}{T}\sum_{t=1}^{T} s(\hat{p}_t, y_t).$$

A scoring rule is said to be *proper* if its expected value is minimal if $\hat{p}$ is the true future distribution of y,

$$E_{y\ p}(s(p, y)) \leq E_{y\ p}(s(\hat{p}, y)),$$

and *strictly proper* if its expectation is minimized only by the true future distribution,

$$E_{y\ p}(s(p, y)) < E_{y\ p}(s(\hat{p}, y)) \quad \text{if} \quad p \neq \hat{p}. \qquad "$$

By the definition above, we see that scoring rules are loss functions that have its expected values minimized by the true distribution of the data. The proper scoring rules used in this dissertation are Brier (or quadratic) and spherical scoring rules, as given by:

$$Brier(\hat{p}, y) = -2\hat{p}_{y,h} + ||\hat{p}||^2 \tag{6-1}$$

$$Spherical(\hat{p}, y) = -\frac{\hat{p}_{y,h}}{||\hat{p}||} \tag{6-2}$$

$$\text{where:} \quad ||\hat{p}|| = \sqrt{\sum_{k=1}^{\infty} \hat{p}_{k,h}^2}$$

The term $\hat{p}_{k,h}$ represents the probability mass of the h-step-ahead predictive distribution evaluated at the value k. That is, evaluating this probability at the realized value of y provides a measure of point forecast accuracy. Larger probabilities assigned to the observed value help to minimize the scoring rule, with the best achievable value being choosing the observed value of y and assigning probability one to this realization.

On the other hand, $||\hat{p}||$ assess how concentrated are the forecasts. This term is minimized when we have equal probabilities assigned for each $\hat{p}_{k,h}$ with $k \in [0, \infty)$, when we achieve $||\hat{p}|| = 0$. Assigning probability one to any forecast will make $||\hat{p}|| = 1$, its maximum value.

Brier scoring rule has values in $[-1, 1]$, while suitable spherical scoring rule values are in $[-1, 0]$. We evaluate $\hat{p}$ numerically in this study using 1000 simulations ran for each h-step-ahead forecast.

When it comes to assessing the point forecast accuracy, we analyse MASE and RMSSE, both of which we define bellow:

$$MASE = \frac{\frac{1}{h}\sum_{j=t+1}^{t+h} |y_j - \hat{y}_j|}{\frac{1}{t-8}\sum_{i=8}^{t} |y_i - y_{i-7}|} \tag{6-3}$$

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{j=t+1}^{t+h} (y_j - \hat{y}_j)^2}{\frac{1}{t-8}\sum_{i=8}^{t} (y_i - y_{i-7})^2}} \tag{6-4}$$

We call the reader's attention to the fact that the numerator of these metrics are Mean Absolute Error (MAE) and RMSE, respectively. The issue of utilizing MAE and RMSE is that both of these are dependent on the scale of the data, what makes comparisons between different series difficult. Both MASE and RMSSE scale out-of-sample errors by the in-sample forecasting errors of random walk models. In our use case, we employ a weekly seasonal random walk, since all models compared have a weekly seasonal component. The only case in which these metrics would be undefined is when all observations in-sample are equal. Values smaller than 1 for both of these metrics represent better forecasts than naive random walk models.

6.3 Descriptive statistics

We now present the descriptive statistics for the time series being analysed. Table 6.1 presents some descriptive statistics of the daily sales series. The presented statistics are the following: demand sizes, i.e. quantity sold given that we have a day with sales; demand per period, the actual observed series containing both days with and without sales; demand intervals, the time between two successive sales; and the fraction of days with zero sales.

Table 6.1: Descriptive statistics of the retail time series.

	Demand sizes	Demand per period	Demand intervals	Fraction of zeroes
Min.	1.328	0.136	1.373	0.207
1st Qu.	1.549	0.329	3.357	0.668
Median	1.665	0.427	4.221	0.738
Mean	1.777	0.512	4.625	0.723
3rd Qu.	1.866	0.592	5.288	0.795
Max.	6.089	3.711	19.120	0.945

Analysing the demand sizes column we see that when we observe positive sales, the typical value is between 1 and 2 units sold for most of the series. Demand per period, which corresponds to the mean of the series, is typically much less than half of the demand sizes. Both demand sizes and demand intervals describe the components in which Croston's method divides the observed demand. The intervals estimated are very variable, but for all series this is larger than 1. The fraction of zeroes observed is concentrated around 0.7.

We now exhibit the auto-correlation functions (ACF) at lags 1, 2 and 7 for these time series. We plot histograms of the estimated values from the series. Concerning the first lag, we note that there is a significant amount of series with values greater than 0.1. On the other hand, most of the estimated values for the second lag concentrate at around 0.05. In the auto-correlation of order 7 we observe similarities to the histograms for ACF(2), though with a little more of variability.

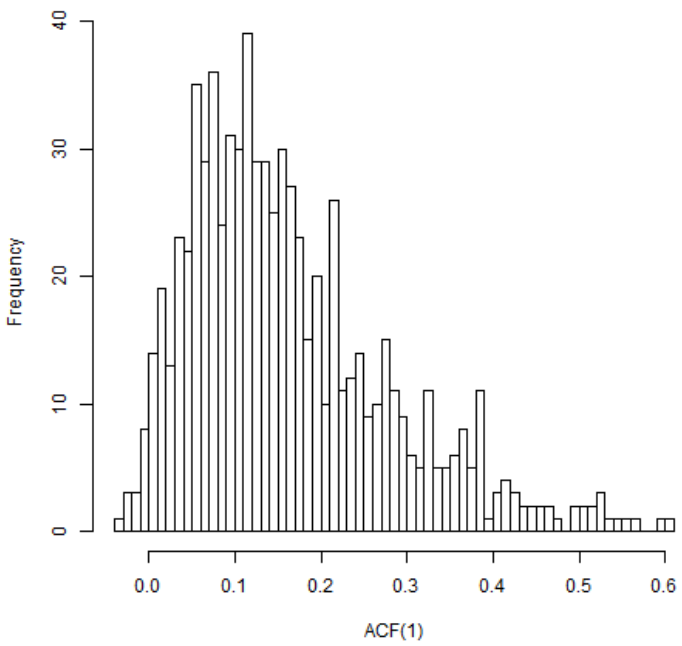

Figure 6.1: Histogram of ACF(1) for the retail time series.

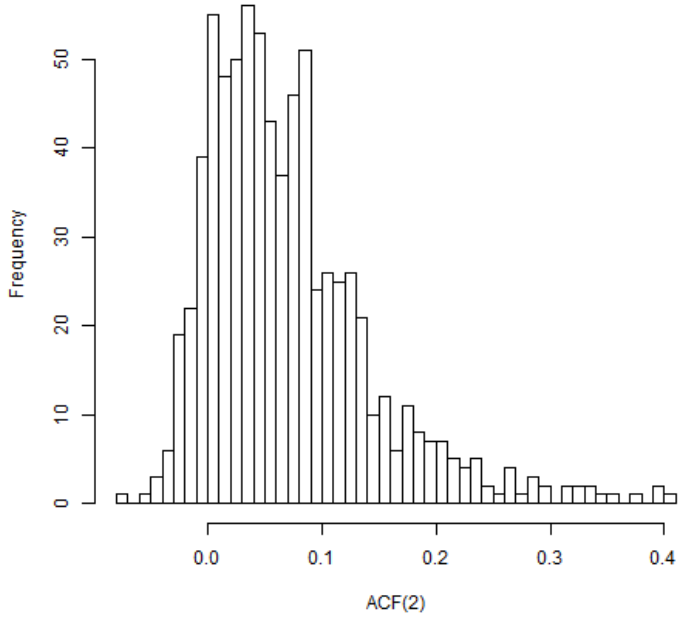

Figure 6.2: Histogram of ACF(2) for the retail time series.

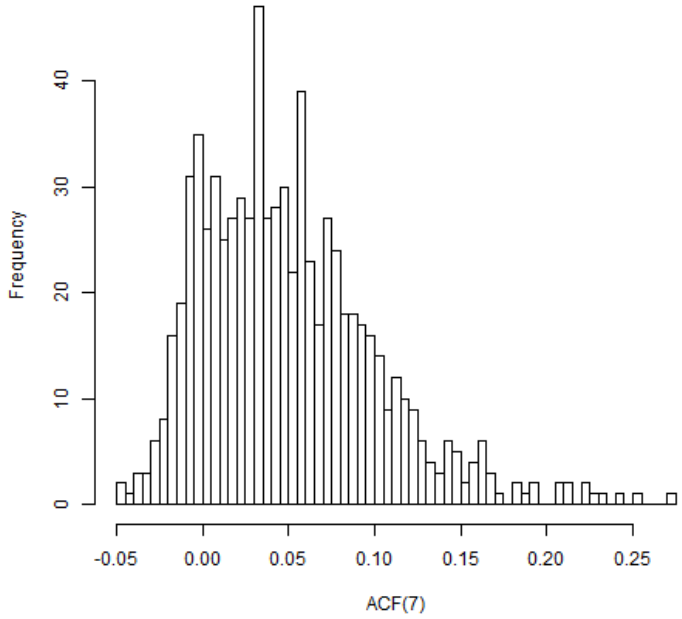

Figure 6.3: Histogram of ACF(7) for the retail time series.

Figure 6.3 raises the question of whether or not the seasonal component is necessary. To asses if this is the case, we have performed a likelihood ratio test in which we compare two different GAS Poisson models: a model containing seasonal components as detailed in equations (4-3) up to (4-6); and a model that excludes equations (4-5) and (4-6), but adds an intercept to equation (4-4). We add the intercept to account for possibly different unconditional means of the level components estimated. The former model is the alternative hypothesis of the test, while the latter details the null hypothesis.

We have used only the first estimation window of each time series to fit these models in order to avoid using pseudo-out-of-sample information to interfere in this decision. When performing the test with 5% significance level, we did not reject the null hypotheses of no seasonal effect in only 7% of the intermittent demand time series.

For the products being analysed, we typically observe that Sundays have lower average sales. Fridays, followed by Saturdays, exhibit higher average sales.

The relationship between the daily sales and price and promotions, the main explanatory variables at our disposal, is now discussed. Recall that these are only included when deemed as relevant in our variable selection procedure.

Concerning prices, we calculate its correlation with the sales for each serie. For the promotions we calculate the difference between the average sales in days with no promotions or holidays, and an average day with each one of the categorized promotions.

Estimated correlations are shown below. Most of the data exhibit a coefficient between -0.3 and -0.1. For the promotions, we plot the estimated metrics excluding the observations classified as outliers to avoid distortions in the graphs.

We see that for most of the promotions, the median stays very close to zero, with relatively low variability. Promotions 6 and 7 show a little more variation. The exception is for promotion 1, which is highly variable and presents a higher median. These numbers may seem small, but recall that the median demand per period of the series is 0.427, so these small changes represent big proportional variations.

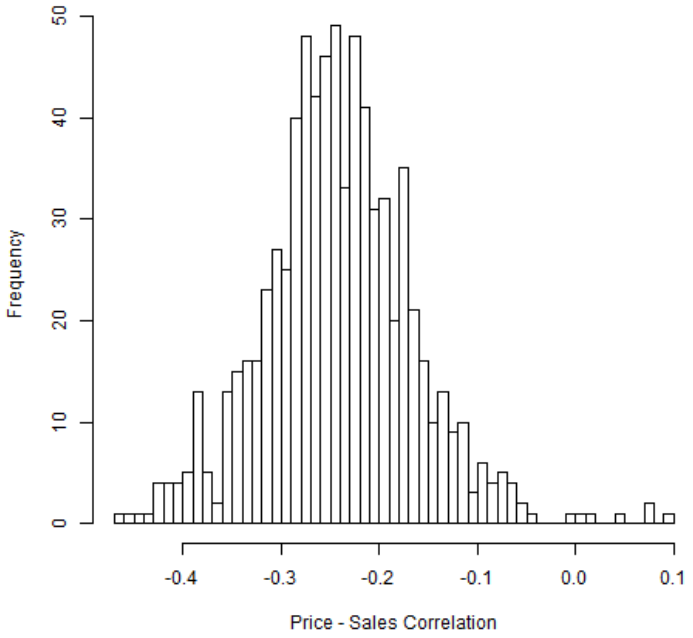

Figure 6.4: Histogram of price-sales correlation.

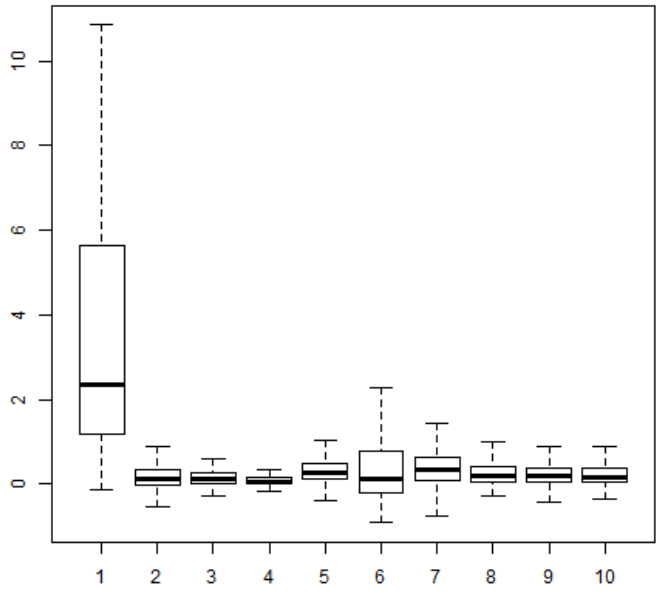

Figure 6.5: Box plot of average difference in sales between days with no promotions and each promotion category.

6.4 Models

We now examine the fitted models, detailed in Chapters 3 and 4, via estimated randomized quantile residuals. As demonstrated by the simulation study in Chapter 5, under correct specification, these residuals should be uncorrelated, homoscedastic and Gaussian. We have also verified that the Jarque-Bera test for normality and Ljung-Box test for auto-correlation often have slightly less rejections than what would be implied by the chosen significance level α, while Ljung-Box test for heteroscedasticity has approximately the right number of rejections for well specified models.

We now repeat the analysis performed in Chapter 5. We report the percentage of rejections of the null hypotheses for each of the tests mentioned, grouping results according to model label. We perform the diagnostics test in each of the estimated serie/window for every model studied. The table that follows exhibits the percentage of null hypotheses rejections for significance level $\alpha = 5\%$:

Table 6.2: Percentage of rejections for diagnostic tests with $\alpha = 5\%$.

	Normality	Auto-correlation	Heteroscedasticity
GAS			
Poisson	62.59%	5.71%	6.05%
NB	10.57%	5.68%	4.9%
ZIP	34.57%	4.93%	4.89%
ZINB	11.24%	5.74%	5.08%
HP	25.64%	8.4%	6.05%
HNB	7.12%	9.74%	4.84%
TS			
Poisson Regression	58.6%	11.46%	8.49%
NB Regression	6.32%	10.34%	6.22%
ZIP Regression	40.76%	10.16%	8.44%
ZINB Regression	19%	9.64%	7.8%
EWMA			
Damped Poisson	63.64%	19%	12.16%
Damped NB	65.15%	23.29%	13.54%
Damped Hurdle Shifted Poisson	50.18%	18.7%	5.29%
Undamped Poisson	69.41%	28.8%	24.76%
Undamped NB	65.74%	12.88%	10.5%
Undamped Hurdle Shifted Poisson	21.6%	9.07%	4%

We first discuss the results for the Jarque-Bera test. All models show higher rejections of the null hypothesis of normality than the 5% nominal value.

It seems that Poisson-type models are more prone to rejection in this test. Note that, in both TS regressions and GAS models, this result is true, for both standard and excessive zeroes distributions. For EWMA models, undamped hurdle shifted Poisson stands out as the most adequate for describing the distribution of the data.

These often bad results are caused by outliers. When we drop estimated models containing randomized quantile residuals larger than 3 in absolute values, we get results closer to those observed in Chapter 5. Outlier treatment was not pursued in this dissertation because of the computational burden that would be incurred for treatment of the total of 8285 models fitted per model class studied.

In spite of the effect that extreme observations exert in Jarque-Bera test statistics, this was less often the case in Ljung-Box tests. For both auto-correlation and heteroscedasticity, the test statistics were estimated using 14 lags, which translates to two weekly seasonal cycles.

We clearly see GAS models being more able to capture the series dynamics, with models having approximately the right percentage of null hypothesis rejections for both auto-correlation and heteroscedasticity tests. Time series count data models perform a little worse in these tests, specially for auto-correlation, but also seem adequate for most use cases. On the other hand, models based on EWMA dynamics suffer a little more. Undamped hurdle shifted Poisson again stands out as the most adequate model in this category, performing even better than time series count data benchmarks. Its damped dynamic counterpart is also one of the most well adjusted in this category, together with damped Poisson model.

Except for a few untreated observations, the models seem to adequately describe data dynamics in most cases.

Concerning only the estimated GAS models, an interesting question is whether or not it is necessary to have a stochastic seasonal component, as we could alternatively estimate seasonal factors as part of the regressors. We can also verify if the π variable should also follow a dynamic model.

To assess if this is the case, we have calculated the number of times the parameters associated with the scores in each situation are statistically significant at 5%. The results are shown in the table that follows:

Table 6.3: Fraction of models with significant estimates for the parameters associated with the scores.

	GAS Poisson	GAS NB	GAS ZIP	GAS ZINB	GAS HP	GAS HNB
κ	0.897	0.894	0.875	0.897	0.617	0.635
ρ_2	-	-	0.759	0.690	0.970	0.964

We employ the same notation of Chapter 4: κ is the parameter that drives the variation in the weekly seasonal component in equation 4-5; and ρ_2 is the parameter associated with the score of π as in equation 4-10, which is responsible for a dynamic probability of zeroes arising from Bernoulli trials.

We can see that for a large proportion of the estimated models, both time varying components are relevant.

6.5 Results

This section discusses the results of our forecasting comparison experiment. We begin the presentation with Brier score (equation 6-1). Similar findings were obtained for spherical score (equation 6-2) and, because of that, we exhibit the results for this metric at the end of the section. We then discuss our findings for point forecasts. Two illustrative examples are discussed at the end of the chapter.

We compare the percentage of times each model is the one that minimizes the metric being analysed. In all tables that follow we highlight the three models most often selected, with darker tones indicating higher percentages. We begin the discussion with the results for Brier score:

Table 6.4: Percentage of times as best model - Brier Scores.

	1 day-ahead	8 days-ahead	15 days-ahead
GAS			
Poisson	4.53%	2.53%	2%
NB	12.52%	9.99%	7.99%
ZIP	7.86%	10.12%	11.98%
ZINB	7.19%	7.59%	9.32%
HP	12.92%	14.25%	13.18%
HNB	4.26%	4.13%	4.79%
TS			
Poisson Regression	3.06%	2.4%	1.86%
NB Regression	8.66%	10.92%	10.79%
ZIP Regression	6.92%	6.79%	6.79%
ZINB Regression	7.86%	9.59%	9.45%
EWMA			
Damped Poisson	1.07%	1.46%	1.6%
Damped NB	4.26%	4.13%	5.33%
Damped Hurdle Shifted Poisson	2.13%	1.46%	1.2%
Undamped Poisson	0%	0%	0.13%
Undamped NB	5.59%	4.39%	3.86%
Undamped Hurdle Shifted Poisson	11.19%	10.25%	9.72%

The proposed GAS HP model is the most often chosen as best performer for all horizons being analysed. GAS NB and ZIP are also chosen as second in ranking for some specific horizons. In general, the models under the GAS label concentrate the most often chosen models for distribution metrics.

The second label that concentrates most often chosen models is TS label, which groups our benchmarks from time series count data literature. NB regression exhibits results similar to GAS ZIP model, being much faster to estimate.

The results for EWMA class concentrates on undamped hurdle shifted Poisson model. This finding is in line with the results obtained in model diagnostics, which points this specific model as being the one that best describes in-sample dynamics of the data analysed in its class.

Interestingly, we have two hurdle Poisson models consistently ranking as best performers for distribution forecasts. This is a good results, as it indicates that the simplest distribution among the excessive zeroes options is sufficient to characterize the data dynamics adequately.

In general, we don't see a large difference in percentage of times chosen as best model for the top three most often chosen. For these, we also don't

see a sharp increase or decrease in relative performance as we analyse further days-ahead.

We now turn our analysis to point forecasts. Both MASE (equation 6-3) and RMSSE (equation 6-4) are discussed. We get different results from the ones obtained in distribution forecasts and also find differences in relative performance among both point forecast metrics being studied. We begin our analysis with MASE, which is presented in the table that follows:

Table 6.5: Percentage of times as best model - MASE.

	1 day-ahead	8 days-ahead	15 days-ahead
GAS			
Poisson	4.26%	4.53%	3.2%
NB	4.93%	4.26%	4.79%
ZIP	5.99%	5.46%	5.06%
ZINB	3.73%	3.73%	4.13%
HP	16.11%	17.84%	15.18%
HNB	8.39%	8.79%	9.19%
TS			
Poisson Regression	2.13%	0.8%	2%
NB Regression	7.86%	4.79%	4.93%
ZIP Regression	6.13%	4.26%	3.06%
ZINB Regression	8.52%	6.79%	8.26%
EWMA			
Damped Poisson	1.73%	3.73%	3.46%
Damped NB	7.46%	9.45%	10.12%
Damped Hurdle Shifted Poisson	2%	1.2%	1.6%
Undamped Poisson	0%	0.13%	0.13%
Undamped NB	7.99%	9.05%	9.59%
Undamped Hurdle Shifted Poisson	12.78%	15.18%	15.31%

We now get results more concentrated in the two most often chosen models. Again we see GAS HP model as the best alternative for both 1 and 8-days-ahead forecasts. The second best performing model is, again, undamped hurdle shifted Poisson, with this model being slightly better than GAS HP for 15-days-ahead forecasts under MASE metric.

Damped NB model now ranks in top three for both 8 and 15-days-ahead forecasts and a similar performance is observed for its undamped dynamic counterpart. GAS HNB and the ZINB regression model show a similar performance, though this group lags behind GAS HP and undamped hurdle shifted Poisson.

As we know, the quantity that minimizes MAE (which is the numerator of MASE) is the median of the predictive distribution, which is the quantity analysed in the previous result. RMSSE, in its turn, is minimized by the mean of the predictive distribution. Count data distributions with low mean exhibit positive skew, and one of the characteristics of these distributions is that the mean is larger than the median. The table that follows repeats the same analysis for RMSSE, now using the mean of the predictive distribution as the point forecast evaluated:

Table 6.6: Percentage of times as best model - RMSSE.

	1 day-ahead	8 days-ahead	15 days-ahead
GAS			
Poisson	11.98%	7.06%	5.73%
NB	4.79%	2.26%	2.13%
ZIP	7.32%	7.06%	5.86%
ZINB	4.39%	3.33%	3.86%
HP	14.25%	15.31%	15.85%
HNB	12.25%	16.11%	18.91%
TS			
Poisson Regression	6.13%	7.32%	7.99%
NB Regression	7.99%	9.19%	5.46%
ZIP Regression	6.66%	6.52%	6.26%
ZINB Regression	5.73%	7.32%	6.39%
EWMA			
Damped Poisson	2.66%	3.6%	5.33%
Damped NB	5.06%	5.99%	5.99%
Damped Hurdle Shifted Poisson	2.26%	1.46%	1.46%
Undamped Poisson	0%	0%	0.13%
Undamped NB	5.59%	4.79%	5.99%
Undamped Hurdle Shifted Poisson	2.93%	2.66%	2.66%

As expected, we get different results from the ones exhibited above. Now, GAS HP and HNB alternate as best model and have a large margin for others.

We found that GAS HP model exhibits a good relative performance consistently across all of the metrics analysed. Concerning the benchmarks studied, we have EWMA related models showing good relative performance in different situations. Undamped hurdle shifted Poisson is the third most often chosen model for predictive distribution metrics in both 1 and 8 days-ahead forecasts, being behind GAS HP by a small margin. For MASE metric, undamped hurdle shifted Poisson places as second best model, being behind

GAS HP for 1 and 8-days-ahead forecasts and taking on the first place for 15-days-ahead predictions. Still for MASE, we see both damped and undamped dynamics NB exhibiting a similar performance.

GARMA models from time series literature behave somewhat like GAS NB model. These are sometimes ranked in the top three models, but the results are not consistent across forecast horizons nor metrics.

The GAS HP models seem specially well suited for intermittent demand forecasting problems. Curiously, although being very similar, the same is not true for the GAS ZIP model. In spite of our results in Chapter 5 demonstrating similarly good results for both HP and ZIP distributions in the examined simulation study, in more realistic settings (more regressors and short time series) it can be the case that convergence of the parameter estimates to its true values might be slower for the latter distribution. The property of zeroes arising from both the standard Poisson distribution and the Bernoulli variable is a possible explanation for slower convergence.

6.5.1 Spherical score results

Table 6.7: Percentage of times as best model - Spherical Scores.

	1 day-ahead	8 days-ahead	15 days-ahead
GAS			
Poisson	3.73%	2%	1.6%
NB	11.72%	10.25%	8.12%
ZIP	8.12%	9.85%	9.85%
ZINB	6.92%	7.59%	9.59%
HP	12.92%	13.98%	15.31%
HNB	4.53%	4.39%	3.99%
TS			
Poisson Regression	2.13%	1.6%	2.13%
NB Regression	9.19%	10.52%	9.45%
ZIP Regression	5.86%	5.99%	5.86%
ZINB Regression	8.12%	8.66%	8.66%
EWMA			
Damped Poisson	0.8%	1.07%	1.6%
Damped NB	5.73%	4.93%	5.99%
Damped Hurdle Shifted Poisson	2%	1.86%	1.46%
Undamped Poisson	0%	0%	0%
Undamped NB	6.52%	6.26%	5.86%
Undamped Hurdle Shifted Poisson	11.72%	11.05%	10.52%

6.5.2 Two illustrative examples

In this subsection we present two of the studied intermittent demand time series to discuss the differences between the GAS HP model and the undamped hurdle shifted Poisson model, which we refer to as "Undamped HSP" in Table 6.8. The criteria for the choice was to select one time series in which each one of the proposed models was selected as the best option across all of the metrics being analyzed, and also that these series reproduce some of the characteristics observed in the forecasts of both models.

For the GAS models presented, we additionally exhibit the filtered estimates for the level and seasonal component. By coincidence, both models possess a nearly constant Bernoulli variable ($\pi_{t|t-1}$). This is an infrequent result, as indicated in Table 6.3, where the vast majority (97%) of the estimates from the GAS HP model count with a statistically significant coefficient associated with the scaled score of the Bernoulli variable.

The values in Table 6.8 report the estimated performance metrics in the whole evaluation period, while the Figures 6.6 and 6.7 exhibit only one of the months in our out-of-sample period. The criterion for the choice of the month was to select the one with less observations discarded according to the criteria described in Section 6.2. Only one-day-ahead forecasts are compared below, but similar findings were obtained for further days-ahead. Labels on the vertical axis were omitted as requested by the company that provided the data. Dataset A exhibits a case in which the GAS HP model was selected as the best performer, while dataset B is better modeled with the undamped hurdle shited Poisson model.

First, concerning the estimated performance metrics exhibited in Table 6.8, we note that scoring rules are not as easily interpreted as the point forecast measures used. This being the case, we turn our analysis to the point forecast measures. As discussed during the presentation of both RMSSE and MASE, in Section 6.2, the denominator of both represents in-sample RMSE and MAE, respectively, from a (seasonal) naive forecasting model. In the absence of a structural break, it is expected that these would have approximately the same estimated values out-of-sample. In this case, values smaller than 1 indicate better forecasts than this simple benchmark. This is indeed what we observe for both models and in both datasets used, where the analysed models beat the benchmark by a large margin. Undamped HSP in dataset B and under MASE evaluation metric reduces the naive forecast errors by $\frac{2}{3}$, this represents a large gain.

Table 6.8: Forcasting metrics comparison for two time series.

	Dataset A				Dataset B			
	Brier score	Spherical score	MASE	RMSSE	Brier score	Spherical score	MASE	RMSSE
GAS HP	-0.465	-0.675	0.532	0.435	-0.592	-0.768	0.593	0.585
Undamped HSP	-0.375	-0.607	0.682	0.592	-0.750	-0.862	0.335	0.487

Figures 6.6 and 6.7 highlight the most notable feature in both models. Undamped HSP (in red) tends to exhibit more variation in the predictions, what, in some cases, leads to an excessive variation in the forecasts. This feature can be observed in both plots, as the range of the predictions is wider for undamped HSP in both cases. On the other hand, GAS HP (blue line) tends to oscillate around the same value.

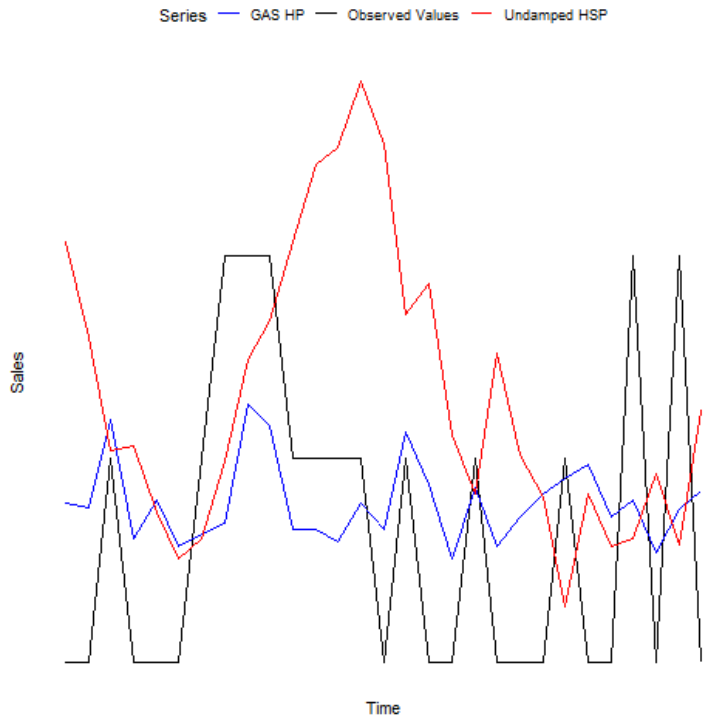

Figure 6.6: Comparison of mean 1-day-ahead forecasts - Dataset A.

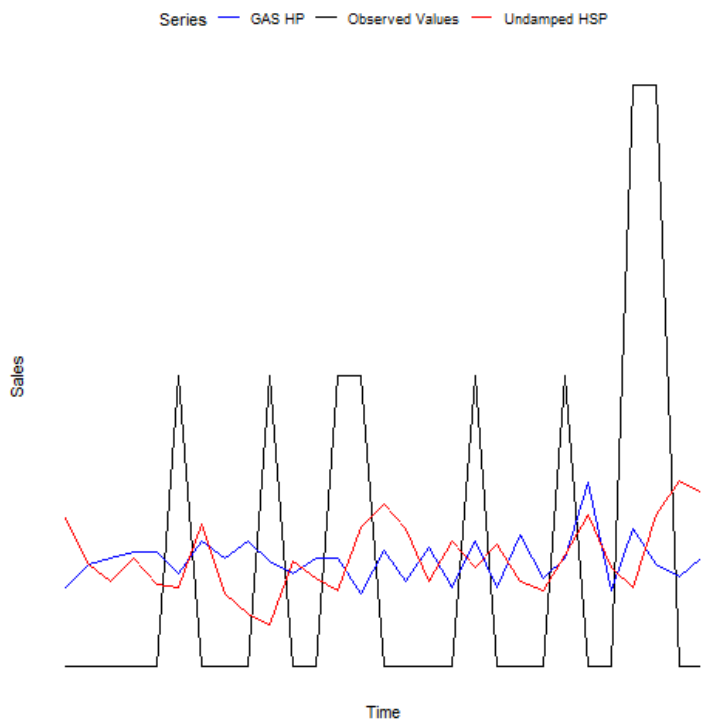

Figure 6.7: Comparison of mean 1-day-ahead forecasts - Dataset B.

The following figures exhibit the filtered level and seasonal components of both Datasets A and B. A frequent concern when modelling series with too many regressors is that the unobserved components might exhibit little variation, as the regressors might capture a big part of the variability of the series. This is specially troublesome for the seasonal component. In our present context, a situation in which this could occur would be if we have a specific promotion that takes place only on a certain day of the week and is very frequent.

This is not the case for the current examples. As we can see, the filtered seasonal components are indeed very different and have a varying amplitude, specially for Dataset B. The level component, on the other hand, captures short term shifts, more visible in Dataset A, and remaining effects not captured by the available regressors.

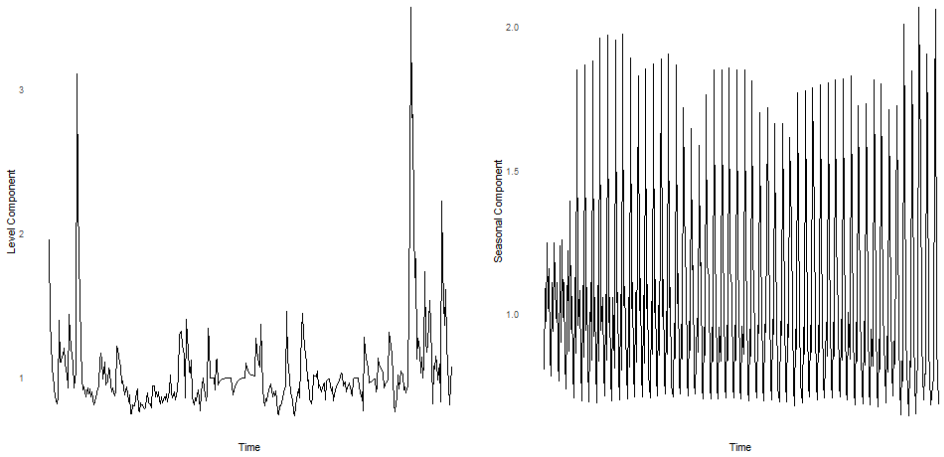

Figure 6.8: Filtered level (left) and seasonal (right) components - Dataset A

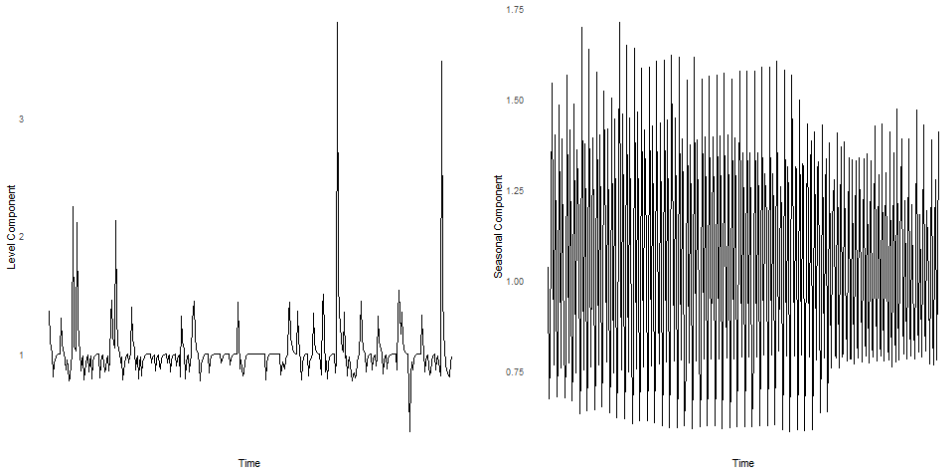

Figure 6.9: Filtered level (left) and seasonal (right) components - Dataset B

7
Conclusions

The purpose of this dissertation was to expand the GAS models framework to previously unstudied discrete distributions, namely zero-inflated Poisson, zero-inflated negative binomial, hurdle Poisson and hurdle negative binomial models. As mentioned, zero-inflated negative binomial distribution was indeed studied before in [4], but we further develop the model to make the Bernoulli variable also dynamic.

The derived models were then applied to the intermittent demand forecasting time series. We have presented an overview of the research in the area, presenting Croston´s method [8] and further works that addressed some of its shortcomings. The proposed GAS models do correct one of the mentioned issues: the necessity of observing a sale to update the forecasts. The models developed in this dissertation are updated every period because the conditional score that drives the forecasts is never null.

For each one of the discussed models, both our own and those from the literature, we have detailed model equations and hypothesised distribution, estimation procedure, inclusion of explanatory variables and forecasting algorithm. Our heuristical approaches to variable selections and parameter initialization (when necessary) were also presented.

Before entering the forecasting experiment, we have studied the finite sample properties of hurdle and zero-inflated GAS models and the properties of the randomized quantile residuals of these under correct specification through a simulation study. We have generated a simple model containing both mean and zeroes regimes and studied the behaviour of the ML estimator with variable sample sizes. The ML estimator is shown to be consistent and asymptotically normal for the excessive zeroes distributions proposed in this dissertation. We have also noted that ASE is not a good estimate of the parameter variance for the sample sizes evaluated. For randomized quantile residuals, we get the expected properties for correctly specified excessive zeroes GAS models: we have Gaussian, uncorrelated and homoscedastic residuals.

Concerning the forecasting comparison exercise, we began detailing the source of the data, one of Brazil´s largest retail companies, and specifying the variables obtained. We then described our data treatments and filters

applied to the dataset, which left us with 752 time series of variable sizes. Some issues regarding estimation and prediction were also presented, along with the accuracy measures used for our evaluation.

The obtained results for our diagnostics show that the models can adequately account for the dynamics exhibited in the series analysed, but they do struggle with Jarque-Bera test for normality. This is the case since we do not engage in model specific outlier treatment, as we have to deal with more than eight thousand estimation windows per model being evaluated. Special events that exhibit very atypical behaviour, as Black Fridays, are treated with dummies. These were the only outlier candidates known beforehand and, as such, were included when deemed important in our variable selection procedure.

The proposed GAS HP model is consistently one of the highest ranked models among the ones studied, both across metrics analysed and forecast horizons. This is a surprising result, as this is the simplest of the mixture models proposed in this dissertation. We call the reader's attention to the fact that both hurdle models don't need to resort to the EM algorithm for parameter estimation, which can take a long time to achieve convergence. Hurdle models can also be estimated by two separate (and lighter) function calls when we set $d = 0$ in the scaled score since, in this case, there is no cross dependence between zero and positive distribution parameters. We can, effectively, estimate one GAS Bernoulli model and another GAS zero-truncated Poisson (or NB) independently and then combine the estimated models for forecasting. Both of this facts contribute for a faster model estimation routine, which can be better suited for the large amount of time series that need to be predicted in a typical retail setting.

Concerning our benchmark models, we note that undamped hudle shifted Poisson model has consistently performed well in all metrics analysed, performing similarly to the proposed GAS HP model for most cases, except when evaluating RMSSE metric. Of the remaining EWMA-like recursions, damped and undamped NB models were only relevant under MASE metric, but lagging behind the most often chosen models. Count time series models, on the other hand, don't excel in neither distribution forecast nor point forecasts metrics, presenting performance comparable to some of the GAS models analysed, but being much faster to estimate.

There are some possible streams of research to be followed from this dissertation. The most obvious is to replicate our findings in Chapter 6 with other datasets and setups. We took advantage of some features of our dataset that are not always available: we used the most disaggregated level of

observations, while in some cases only weekly or monthly data are available; and also we had access to explanatory variables that are relevant to the problem being analysed. The M5 competition dataset provides very large amount of time series with a similar structure in which we could further validate this findings. On the other hand, the asymptotic properties of the developed models are not formally demonstrated. Our findings indicate that the parameters estimates are indeed consistent and asymptotic normal, but a theoretical proof is needed.

Bibliography

[1] MA, S.; KOLASSA, S. ; FILDES, R.. **Retail forecasting: Research and practice**. International Journal of Forecasting, 2019.

[2] JOHNSTON, F. R.; BOYLAN, J. E. ; SHALE, E. A.. **An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items**. Journal of the Operational Research Society, 54(8):833–837, 2003.

[3] NIKOLOPOULOS, K.; SYNTETOS, A.; BOYLAN, J.; PETROPOULOS, F. ; ASSIMAKOPOULOS, V.. **An aggregate-disaggregate intermittent demand approach (adida) to forecasting: An empirical proposition and analysis**. JORS, 62:544–554, 03 2011.

[4] BLASQUES, F.; HOLY, V. ; TOMANOVA, P.. **Zero-inflated autoregressive conditional duration model for discrete trade durations with excessive zeros**. Tinbergen Institute Discussion Papers, (19-004/III), Jan 2019.

[5] BLAZSEK, S.; ESCRIBANO, A.. **Score-driven dynamic patent count panel data models**. Economics Letters, 149:116–119, 10 2016.

[6] CREAL, D.; KOOPMAN, S. J. ; LUCAS, A.. **Generalized autoregressive score models with applications**. Journal of Applied Econometrics, 28, 08 2013.

[7] HARVEY, A.. **Dynamic models for volatility and heavy tails: With applications to financial and economic time series**. Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series, p. 1–262, 01 2011.

[8] CROSTON, J. D.. **Forecasting and stock control for intermittent demands**. Operational Research Quarterly (1970-1977), 23(3):289–303, 1972.

[9] COX, D. R.; GUDMUNDSSON, G.; LINDGREN, G.; BONDESSON, L.; HARSAAE, E.; LAAKE, P.; JUSELIUS, K. ; LAURITZEN, S. L.. **Statistical analysis of time series: Some recent developments [with**

discussion and reply]. Scandinavian Journal of Statistics, 8(2):93–115, 1981.

[10] BLAZSEK, S.; LICHT, A.. **Dynamic conditional score models: a review of their applications**. Applied Economics, p. 1–19, 09 2019.

[11] HARVEY, A.; SUCARRAT, G.. **Egarch models with fat tails, skewness and leverage**. Computational Statistics Data Analysis, 76, 11 2012.

[12] BLAZSEK, S.; LICHT, A. ; ESCRIBANO, . **Score-driven non-linear multivariate dynamic location models**. UC3M Working papers. Economics, (25739), Oct 2017.

[13] BLAZSEK, S.; ESCRIBANO, A. ; LICHT, A.. **Co-integration and common trends analysis with score-driven models: an application to the federal funds effective rate and us inflation rate**. UC3M Working papers. Economics, (28451), 07 2019.

[14] CREAL, D.; KOOPMAN, S. J. ; LUCAS, A.. **A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations**. Tinbergen Institute Discussion Papers, (10-032/2), 2010.

[15] SCHWAAB, B.; KOOPMAN, S. J.; LUCAS, A. ; CREAL, D.. **Observation driven mixed-measurement dynamic factor models with an application to credit risk**. European Central Bank Working Paper Series, (1626), Dec. 2013.

[16] OPSCHOOR, A.; LUCAS, A.; BARRA, I. ; VAN DIJK, D.. **Closed-form multi-factor copula models with observation-driven dynamic factor loadings**. Tinbergen Institute Discussion Papers, (19-013/IV), Feb 2019.

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

[18] HARVEY, A.; LUATI, A.. **Filtering with heavy tails**. Journal of the American Statistical Association, 109, 07 2014.

[19] CAIVANO, M.; HARVEY, A. ; LUATI, A.. **Robust time series models with trend and seasonal components**. SERIEs, 7, 12 2015.

[20] HARVEY, A.; ITO, R.. **Modeling time series when some observations are zero**. Journal of Econometrics, 214, 08 2019.

[21] SYNTETOS, A. A.; BOYLAN, J. E.. **The accuracy of intermittent demand estimates**. International Journal of Forecasting, 21(2):303–314, 2005.

[22] TEUNTER, R.; SYNTETOS, A. ; BABAI, M. Z.. **Intermittent demand: Linking forecasting to inventory obsolescence**. European Journal of Operational Research, 214:606–615, 11 2011.

[23] GUTIERREZ, R.; SOLIS, A. ; MUKHOPADHYAY, S.. **Lumpy demand forecasting using neural networks**. International Journal of Production Economics, 111:409–420, 01 2001.

[24] WILLEMAIN, T.; SMART, C. ; SCHWARZ, H.. **A new approach to forecasting intermittent demand for service parts inventories**. International Journal of Forecasting, 20:375–387, 02 2004.

[25] JOHNSTON, F. R.; BOYLAN, J. E.. **Forecasting for items with intermittent demand**. The Journal of the Operational Research Society, 47(1):113–121, 1996.

[26] SYNTETOS, M.; BOYLAN, J. ; CROSTON, J.. **On the categorization of demand patterns**. Journal of the Operational Research Society, 56, 05 2005.

[27] KOSTENKO, A.; HYNDMAN, R.. **A note on the categorization of demand patterns**. Journal of the Operational Research Society, 57:1256–1257, 10 2006.

[28] HEINECKE, G.; SYNTETOS, A. ; WANG, W.. **Forecasting-based sku classification**. International Journal of Production Economics, 143:–, 06 2013.

[29] SYNTETOS, A.; BABAI, M. Z. ; ALTAY, N.. **On the demand distributions of spare parts**. International Journal of Production Research, 50:2101–2117, 04 2012.

[30] SNYDER, R.; ORD, K. ; BEAUMONT, A.. **Forecasting the intermittent demand for slow-moving inventories: A modelling approach**. International Journal of Forecasting - INT J FORECASTING, 28, 04 2012.

[31] HYNDMAN, R.; KOEHLER, A.; ORD, K. ; SNYDER, R.. **Forecasting with exponential smoothing. The state space approach**. 01 2008.

[32] KOLASSA, S.. **Evaluating predictive count data distributions in retail sales forecasting**. International Journal of Forecasting, 32:788–803, 07 2016.

[33] CZADO, C.; GNEITING, T. ; HELD, L.. **Predictive model assessment for count data.** Biometrics, 65 4:1254–61, 2009.

[34] HILBE, J.. **Negative Binomial Regression**. Cambridge University Press, 2011.

[35] GREENE, W.. **Accounting for excess zeros and sample selection in poisson and negative binomial regression models**. NYU Working Paper, (EC-94-10), 02 1994.

[36] BENJAMIN, M.; RIGBY, R. ; STASINOPOULOS, D.. **Generalized autoregressive moving average models**. Journal of the American Statistical Association, 98:214–223, 02 2003.

[37] DAVIS, R. A.. **Observation-driven models for poisson counts**. Biometrika, 90(4):777–790, 2003.

[38] FOKIANOS, K.; TJØSTHEIM, D.. **Log-linear poisson autoregression**. Journal of Multivariate Analysis, p. 563–578, 03 2011.

[39] DAVIS, R.; WU, R.. **A negative binomial model for time series of counts**. Biometrika, 96:735–749, 08 2009.

[40] BLASQUES, F.; KOOPMAN, S. J. ; LUCAS, A.. **Maximum likelihood estimation for score-driven models**. Tinbergen Institute Discussion Papers, (14-029/III), Mar. 2014.

[41] LAMBERT, D.. **Zero-inflated poisson regression, with an application to defects in manufacturing**. Technometrics, 34:1–14, 02 1992.

[42] DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, D. B.. **Maximum likelihood from incomplete data via the em algorithm**. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977.

[43] HYNDMAN, R.; KOEHLER, A.; SNYDER, R. ; GROSE, S.. **A state space framework for automatic forecasting using exponential smoothing methods**. International Journal of Forecasting, 18:439–454, 02 2002.

[44] ZOU, H.. **The adaptive lasso and its oracle properties**. Journal of the American Statistical Association, 101(476):1418–1429, 2006.

[45] TIBSHIRANI, R.. **Regression shrinkage and selection via the lasso**. Journal of the Royal Statistical Society (Series B), 58:267–288, 1996.

[46] DUNN, P. K.; SMYTH, G. K.. **Randomized quantile residuals**. Journal of Computational and Graphical Statistics, 5(3):236–244, 1996.

[47] HYNDMAN, R. J.; KOEHLER, A. B.. **Another look at measures of forecast accuracy**. International Journal of Forecasting, 22(4):679–688, 2006.