Applications of Bayesian mixture models and self-exciting processes to retail analytics

James Pitkin

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of the

University of London.

Department of Statistical Science

University College London

April 13, 2020

Declaration of authorship

I, James Pitkin, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Retail analytics has been transformed by big data, which has led to many retailers using detailed analytics to improve performance at a range of operational levels. This is the case with the collaborator of this research, dunnhumby, who have large amounts of retailer data derived from the numerous activities that retailers operate at. This thesis focuses on two challenges retailers face; the analysis of products through their price elasticity coefficients and demand forecasting of products known as *slow-moving inventory.*

The analysis of products in terms of their price elasticity coefficients is well studied. Existing approaches are hampered by the challenging nature of cross-elasticity data, as cross-elasticity coefficients typically vary in dimension and exhibit an inherent censoring. We address these problems by developing a systematic model-based approach by reinterpreting the cross-elasticity coefficients as realisations of *variable length order statistics sequences*, and develop a nonparametric Bayesian methodology to cluster these sequences. Our approach uses the Dirichlet process mixture model that allows data to dictate the appropriate number of clusters and provides interpretable parameters characterising the decay of the leading entries.

Slow-moving inventory are characterised by having intermittent demand, in that the demand is populated with an abundance of zero sales and that, when a sale does a occur, it is often followed by a quick succession of sales. This demand intermittency inhibits the use of traditional analytics which crucially affects optimal inventory management. To combat this, we represent intermittent demand as a structured multivariate point process which allows for auto- and cross-correlation frequently observed in sparse sales data. Our approach uses a hurdle component to cope with zero sales inflation, the Hawkes process to capture the temporal clustering and a hierarchal structure to pool information across products.

We illustrate our methods on real retailer data, from access granted by dunnhumby.

Impact statement

The UK retail sector is a significant one; during 2017, consumers in the UK spent around £406 billion on retail purchases, with 39p of every £1 being spent in food stores [Rhodes and Brien, 2018]. This scale, along with the proliferation of data sources derived from the services and products that retailers offer to consumers, has meant responsible data science is increasingly being used to identify inefficiencies and opportunities, as well as helping to provide a higher degree of personalisation to consumers than ever before. A company at the forefront of consumer data science is the collaborator of this research, dunnhumby.

This research, supported by the EPSRC, dunnhumby and the Alan Turing Institute, looks to explore the applications that Bayesian nonparametric mixture modelling, excitation processes and hierarchical modelling have to retail analytics. We focused on two specific problems that retailers face: product clustering and intermittent demand forecasting. The first output of this research was a product clustering methodology that used a Dirichlet process mixture model to capture the nuanced structure exhibited by the elasticity coefficients outputted from demand models traditionally used by retailers. The second output was a forecasting methodology, where we demonstrated the effectiveness that information pooling, a discretised Hawkes process and regression covariates had on the issue of time series forecasting of intermittent demand. These methodologies provide refreshing reference points from which other product clustering and intermittent demand forecasting models could be benchmarked. In addition, each of these approaches may have fruitful applications to fields beyond retail analytics that strive to cluster strictly decreasing or increasing censored data, or to time series forecasting where intermittency inhibits the use of traditional methodologies. These investigations led to two paper submissions to leading statistics journals: the Annals of Applied Statistics and Journal of the Royal Statistical Society: Series C.

The insights of this research provide important implications to professional analysts within the retail analytics industry. Our product clustering methodology casts light over the structural

differences among products' sales sensitivities as exhibited through their cross-elasticity coefficients, and how these sales sensitivity differences are split across brands and food categories. These findings are valuable to retailers, as information on product differences can be used to improve promotional activities and help retailers to differentiate themselves from competitors, by using such features to improve customer loyalty campaigns. The forecasting methodology developed during this research afforded greater transparency into the challenging dynamics exhibited by intermittent demand. Our approach allows retailers to more clearly understand the predictive benefits that hierarchical modelling, seasonality, price and temporal excitation have in intermittent demand. This benefits retailers by allowing them to manage their supply of inventory more optimally by improving the short-term demand forecasts of products, which in turn enables retailers to reduce operational costs associated with stockpiling. Furthermore, it gives retailers the ability to make accurate assessments of the effects that promotions, price changes and marketing campaigns would have on aspects such as profit and revenue, that inaccurate forecasting methodologies are unable to do.

Acknowledgements

I would like to thank the EPSRC, the Centre for Doctoral Training in Financial Computing & Analytics and the Alan Turing Institute for the financial support that made my PhD possible. I am especially grateful to Professor Treleaven, who believed in the project from the outset and whose advice and contacts were valuable throughout my work. I am also grateful to our industrial collaborator dunnhumby, both for the further financial support received and for the range of exciting projects made available to us through this partnership. Particular mention goes to Rosie Prior, Dr. Marco Brambilla, Dr. David Hoyle, Piers Stobbs, and Giles Pavey, all of whom provided valuable insights into our work and who facilitated an innovative research environment. Furthermore I would like to thank my supervisors, Dr. Ioanna Manolopoulou and Dr. Gordon Ross, who have been hugely influential on this work and who have both been a pleasure to work with. I have learned so much from you both.

On the social front, I want to thank Tristan, Phil and Hugh for the lively debates and, importantly, people to play squash with. Special thanks goes to Anthony & Jean Levy, who were both a home away from home and on numerous occasions listened to me ramble at length about many topics. Needless to say, so much thanks goes to Emma, who has been so supportive, caring, and amusing throughout my endeavours. I only hope I can be the person who you were for me one day. Finally, Mum & Dad - thanks for everything.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The key aim of this work is to develop a range of novel statistical methodologies that have applications to problems arising in retail analytics. The retail sector, as defined by the UK Office of National statistics, is an industry where 'sales of products by retailers are made directly to end consumers, including spending on goods (in store and online) and spending on services'. The British retail industrial sector is a significant one; during 2016 the retail sector contributed £192 billion to UK economic output (11.4% of the total) as measured by Gross Value Added [Rhodes, 2017] and the value of retail sales at current prices (non-seasonally adjusted) on predominantly food stores grew from 151 billion to 154 billion (1.5% increase) [ONS]. Within 2016, for every pound spent in the retail industry, 40 pence was spent in food stores.

Many companies comprise the British retail industry as a whole, but especially in the food sector, a large percentage of the supply of food to British households is through a small number of companies who operate numerous supermarket chains throughout multiple regions in the United Kingdom. To illustrate this point, 69% of the market share of grocery stores in the UK between January 2015 to March 2017 was composed of the 'big four': Tesco, Sainsbury's, Asda and Morrisons [Statista, 2017]. To put this scale into context: Tesco during the 2015-2016 financial year reported approximately 79 million shopping trips per week across their over 6700 stores globally [Tesco PLC, 2017]. Unsurprisingly, as a consequence of operating at such a transactional and regional scale, the volume of data these companies now collect is large, and increasing. Whether from the increase of retail purchases done online in the UK [Rhodes, 2017], or the fact that globally 66% of shoppers report to be a member of one or more loyalty programs [Nielsen Company, 2016], it generally illustrates the vast resource of information to which retailers now have access.

In recent years, this growth of data that British retailers now collect is part of a larger trend known as ‘big data’. For retailers, the value and insights that data can provide is well established, as retailers are increasingly deploying teams of analysts and modelling specialists who devote themselves to better understanding how data can be used to find solutions and insights that previously were unobtainable without such a mass of available information. On the cutting edge of this development of retail analytics is the collaborator of this work, dunnhumby, who have used analytics to save £16 million in one year by optimally managing their supply chain during seasonal periods and to improve repeat business by building effective loyalty schemes [Swabey, 2013, Turner and Wilson, 2006]. In the increasingly competitive landscape of retail, innovative companies are now using ‘big data’ to identify inefficiencies and create a distinct advantage that otherwise would be difficult to achieve.

During this work, we focus on the supermarket sector of the retail industry and explore how clustering and forecasting methodologies can be used to create a competitive advantage. Although it should be noted, that all data this work is based on has been fully anonymised for general research purposes so that no individual shoppers, or any other sensitive data could be identified.

1.1 Clustering methodologies

The scale and complexity of data that retailers now manage has, in instances, led to many conventional methodologies of processing, analysing and interpreting data needing to adapt. One such class of approach used to better understand and interpret data is clustering methodologies, which have now been widely applied to the field of retail analytics. Clustering approaches have been extensively used to improve retailer processes; for example, boost revenue by improving product recommendations [Lawrence, Almasi, Kotlyar, Viveros, and Duri, 2001], improve the stock replenishment of inventory systems [Stefanovic, Stefanovic, and Radenkovic, 2008, Bala, 2012], reduce computational runtime [Sarwar, Karypis, Konstan, and Riedl, 2002], target marketing efforts more effectively [Kashwan and Velu, 2013] and represent a neat and automated approach of classifying large and complex retailer data that allows for clear, interpretable and actionable results [Ghosh and Strehl, 2005]. In short, clustering approaches are widely applied in retail, where they demonstrate utility in terms of predictive performance and allowing data to be more easily interpreted, both of which create an advantage to retailers who have the data to support such initiatives.

1.1.1 Clustering cross-elasticity coefficients

One such area in which retail analytics could benefit from a clustering methodology is the analysis of cross-elasticity coefficients. Retailers frequently implement demand modelling methodologies for a range of purposes, one reason being to better understand how the underlying phenomena at play in the environment, such as a product's price, its competitors' prices and seasonal dynamics, impact a product's demand. One set of phenomena retailers are particularly interested in is how a product's demand is affected by changes of its own price as well as changes of its competitors' prices. This relationship between change in demand and price changes is often expressed through direct and cross-elasticity coefficients. These coefficients link the rate of change in the demand of a product to a change of its own price, and the rate of change in the demand of a product to changes of its competitor's price, respectively. Broadly speaking, the motivation of analysing products in terms of their direct and cross-elasticity coefficients is two-fold. Firstly, these coefficients are analysed to derive possible insights aimed at aiding public policy with findings often striving to improve dietary trends. Secondly, they are analysed from the retailers' perspective, where these coefficients are used to form a strategic understanding of how their products demand is affected by competitors' prices, measure the responsiveness to marketing efforts or generally inform profit maximisation strategies. However, in spite of the analysis of cross-elasticity coefficients being widespread in the literature, there is surprisingly no formal methodology that automates the comparison of the cross-elasticity coefficients in a systematic way. Instead, economists and retail analytics professionals manually analyse the cross-elasticity output in an ad-hoc fashion, where they assess similarity between products by studying the absolute differences between cross-elasticity coefficients, and ignore much of the structurally interesting aspects that cross-elasticity coefficients can exhibit.

The first motivation of this work is the following: we look to develop a methodology that aims to non-parametrically cluster products in terms of their cross-elasticity coefficients in an automated fashion, requiring minimal experimenter intervention as possible. In particular, we use novel Bayesian nonparametric methods that are able to accommodate the complex heterogeneity and inherent structure exhibited in cross elasticity data.

1.2 Demand forecasting

Forecasting the demand of products that retailers offer to their market is well established. The aim of demand forecasting is rooted in retailers trying to balance supply and demand. By being able to anticipate demand inflows, organisations are able to appropriately manage their supply of a product, which allows retailers to avoid opportunity costs of understocking products (loss of potential revenue and customer dissatisfaction), or overstocking products (incurring inventory costs, stock depreciation and the likelihood of making sharp price reductions to shift stock). Consequently, accurate demand forecasting is closely related to effective supply chain and inventory management and consequently, retailers often make concerted efforts to develop methodological approaches that reduce the downsides of inaccurate demand forecasts. Finally, retailers' interest in demand forecasting stems from a desire to understand the effect that numerous drivers have on the demand of their products. By understanding what drives a product's demand, retailers can better ascertain the motivations behind consumer choices, marketing effects and seasonal aspects which crucially allows them to make confident and assured decisions in light of the available information, and invest resources accordingly [Steenburgh, 2007, Rudin, Letham, and Madigan, 2013, Ferreira, Lee, and Simchi-Levi, 2015]. Unsurprisingly, demand forecasting is especially important for large retailers who operate many large regional outlets, as the impact of incremental improvements of efficiencies translates to significant gains or losses in revenue. Thus, retailers that use their data to anticipate demand place themselves at distinct advantage compared to retailers who do not.

1.2.1 Forecasting intermittent demand of slow-moving-inventory

Many demand forecasting approaches are successful at achieving their objectives of profit maximisation, improving inventory management or gaining a clearer understanding of the factors impacting a product's demand. However, forecasting the demand of some products is more difficult than it is for others. One such class of products that are traditionally difficult to forecast the demand for are known as *slow-moving-inventory.*

The demand patterns of slow-moving-inventory products are generally characterised by being very intermittent, in that there is typically an inflation of zero sales and that, when a sale does occur, it is often followed by a quick succession of sales. The inflation of zeros and burstiness in the intermittent demand of slow-moving inventory make demand forecasting challenging, as it can obfuscate an understanding how the environment, and more specifically the covariate information, impacts the demand signal. Consequently, forecasting models not fully incorporating the nuanced covariate and temporal dynamics of intermittent demand of slow-moving inventory frequently lead to inaccurate forecasts, which in turn inhibits a retailer's ability to balance supply with demand. Methodologies from fields such as machine learning and statistics have sought to handle the issues arising from intermittent demand forecasting. However, there does not seem to be a unified forecasting methodology that handles the zero-inflation, temporal burstiness and provides an explanation of the underlying phenomena existing in the intermittent demand of slow-moving inventory in a simultaneous fashion.

The second motivation of this work is the following: we strive to develop a forecasting methodology for the intermittent demand of slow-moving inventory that unifies the structural artefacts of hierarchy, auto-correlation, cross-correlation and temporal clustering across multiple intermittent demand series whilst still offering explanatory power. We do this by representing the intermittent demand as a structured multivariate point process which includes a hurdle component for the abundance of zero demand, a Hawkes process to cope with temporal clustering within and across products, and a hierarchal structure to pool information across a large number of products.

The rest of this work is structured as follows: Chapter 2 provides a background of the Bayesian method and outlines the inferential procedures relevant to this work. Chapter 3 presents the work of Bayesian nonparametric (BNP) models, focusing on Dirichlet process mixture models. Chapter 4 describes in deeper detail retailers' interest in the analysis of cross-elasticity coefficients and outlines existing work in cross-elasticity analysis. Chapter 5 presents a clustering methodology for the coefficients of cross-elasticity demand models which allows retail analysts to characterise supermarket products in terms of their sales sensitivity. Chapter 6 introduces the challenges related to forecasting the intermittent demand of slow-moving inventory. This chapter goes on to outline the existing work in intermittent demand forecasting, and further describes hurdle regression models and a class of point process known as the Hawkes process. Chapter 7 presents a novel regression model able to forecast the intermittent demand for slow-moving-inventory that uses a Bayesian hierarchical hurdle model with excitation components described by a Hawkes process that captures the temporal dynamics exhibited in the intermittent demand of slow-moving-inventory data. Chapter 8 concludes by reiterating the contribution of this work and further describes the scope for future applications of the Bayesian approach to retail analytics.

Chapter 2

Bayesian inference

This Chapter gives a brief overview of the statistical modelling paradigm known as Bayesian statistics. Section 2.1 describes a general statistical framework relevant to this work along with descriptions and benefits of the Bayesian method. Section 2.2 provides a background on Markov chain Monte Carlo (MCMC) methods as the primary route to statistical inference under the Bayesian framework. The objective of this chapter is to give the reader a practical understanding of the framework underpinning Bayesian statistics and MCMC theory as well as providing a description of the relevant MCMC algorithms employed in this work.

2.1 The statistical model paradigm

The broad aim of statistical inference is to describe a set of observations from some process as draws from some probability model that is itself a representation of the original process. More concretely, suppose we observe a sequence $y_1, y_2, ..., y_n$ of instances which are assumed to be drawn randomly from a sequence of random variables $Y_1, Y_2, ..., Y_n$ with respect to the sample set Ω. We then introduce the notion of probability models, P_θ (i.e. processes of identical distributional form with respect to some θ), indexed by $\theta \in \Theta$ - where θ are *parameters* with respect to some parameter space Θ. Assuming that these $y_1, y_2, ..., y_n$ observations were generated independently and identically from the models P_θ, we can then write the following:

$$Y_i \overset{iid}{\sim} P_\theta, \text{ for } i = 1, ..., n \tag{2.1}$$

for some $\theta \in \Theta$ [Cai, 2014]. The key objective to statistical modelling is making inferences based on the observed data $\mathcal{D} = \{y_1, y_2, ..., y_n\}$ with the constraints of the probability models specified by P_θ. There are two main approaches to statistical inference; the *classical* and *Bayesian* approaches. Broadly speaking, the *classical* approach treats all parameter values θ as unobserved fixed constants, whereas the *Bayesian* approach treats θ as another *random variable*. During this work, we operate under the Bayesian approach.

2.1.1 The Bayesian method

Under the Bayesian paradigm, the model of (2.1) can be re-expressed as the following:

$$\begin{aligned} y_i &\overset{iid}{\sim} F(y \mid \theta), \text{ for } i = 1, ..., n \\ \theta &\sim \pi \end{aligned} \tag{2.2}$$

where $F(y \mid \theta)$ is the distributional form of P_θ and π is the prior distribution of θ. The key to the *Bayesian approach* is the specification of the prior π over parameter space of Θ. The prior π is aimed to represent all prior knowledge of the θ values, and supposed to reflect expertise before the outset of an experiment. Consequently, Bayesian statistics is interested in characterising the entire distribution of $p(\theta \mid \mathcal{D})$, referred to as the *posterior distribution*, and then deriving relevant statistics from this distribution. The derivation of relevant statistics from $p(\theta \mid \mathcal{D})$ is known as *posterior inference*.

Bayesian methodologies have been widely applied to a range of disciplines from spatial weather modelling to the temporal modelling of coal mining disasters [Reich and Fuentes, 2007, Taddy, Kottas, et al., 2012], and are perceived to have useful properties when used to model data. However, for the purposes of this work, the benefits of the Bayesian method are the following:

- **Expressing prior beliefs:** Through the specification of prior distribution π, experimenters are able to express their prior beliefs of the likely values of θ. This allows experimenters to penalise the complexity of fitted statistical models and therefore, can be used as a mechanism to reduce *overfitting* data [Simpson, Rue, Riebler, Martins, Sørbye, et al., 2017, Murray and Ghahramani, 2005, MacKay, 1992]. The process of *overfitting* is the situation where a statistical model too closely fits a limited set of data. In addition to penalising complexity, priors can be used to handle issues related to parameter estimation in instances of small sample sizes [Sahu and Smith, 2006].

- **Hierarchical borrowing:** In the situation of hierarchical or multilevel modelling, the Bayesian framework naturally allows information pooling between parameters across the various levels of the model hierarchy. Under the *i.i.d* assumption, a Bayesian hierarchical model can be expressed as:

$$\begin{aligned} y_i &\overset{iid}{\sim} F(y \mid \theta), \text{ for } i = 1, ..., n \\ \theta &\sim \pi(\omega) \\ \omega &\sim \Pi \end{aligned} \tag{2.3}$$

where Π is the prior of the hyper-parameters ω that parametrise π. In this instance, the hyper-parameters ω of the prior distribution π are themselves random quantities. This information pooling across parameters has been found to offer improvements to model fit and predictive performance in instances where a hierarchical structure exists [Gelman, 2006, Jensen, Shirley, and Wyner, 2009].

- **Expressing uncertainty of experimental quantities:** As quantities in the Bayesian framework are themselves random processes, expressing the uncertainty around parameter inferences is automatically inherited from the Bayesian approach [Kass and Raftery, 1995, Berger and Pericchi, 1996]. Consequently, the posterior predictive distribution $p(y^* \mid \mathcal{D})$, where y^* is new predictive data point, is naturally accompanied with uncertainty which crucially allows prediction intervals to be constructed around quantities of interest. This is particularly valuable in contexts such as ours, where experimenters are interested in the likely distribution of outcomes. In these situations, the Bayesian approach has been shown to offer particular utility [Tu and Zhou, 2010, Kalyanam, 1996, Lee, Boatwright, and Kamakura, 2003].

Although these points are by no means exhaustive, they are a stylised list of benefits relevant to this work. However, analytical expressions and direct sampling from $P(\theta \mid \mathcal{D})$ is often intractable, which therefore presents challenges. A class of sampling algorithms known as *Markov chain Monte Carlo* (MCMC) have been developed that can accommodate posterior inference.

2.2 MCMC algorithms

The key idea behind MCMC is to simulate a sequence of random variables $\{\theta_0, \theta_1, \theta_2, \ldots\}$ such that these samples are samples drawn from $P(\theta \mid \mathcal{D})$. One route to generating a sequence of random variables $\{\theta_0, \theta_1, \theta_2, \ldots\}$ equivalent to samples drawn from $P(\theta \mid \mathcal{D})$ is by constructing a Markov chain whose stationary distribution π is precisely the target distribution $P(\theta \mid \mathcal{D})$. We now denote the distribution $P(\theta \mid \mathcal{D})$ as $\phi(\cdot)$, and refer to this as the *target distribution*. Before constructing the appropriate Markov chain whose stationary distribution π is $\phi(\cdot)$, we outline the sufficient conditions that such a Markov chain T needs to satisfy. These conditions are as follows:

Definition 1. *T is irreducible*

A Markov chain T_{ij} is irreducible if for every i, j there exists a finite integer m_{ij} such that $p(T_{m_{ij}} = j \mid T_0 = i) > 0$.

Definition 2. *T is aperiodic*

A Markov chain T_{ij} is aperiodic if for every i, $\gcd\{n > 0 : p(T_n = i \mid T_0 = i) > 0\} = 1$ is satisfied.

Definition 3. *T is positive recurrent*

Let $t_i = \min\{n \geq 0 \mid T_n = i\}$, then a Markov chain T_{ij} is positive recurrent if for every i, $\mathbb{E}(t_i \mid T_0 = i) < \infty$.

Given such a Markov chain T, it will have a unique stationary distribution $\pi(\cdot)$, and is such that for every $i, j \in S$

$$\lim_{n\to\infty} T^n_{i,j} = \pi_j$$

where $T^n_{i,j} = p(T_n = j \mid T_0 = i)$, and π_j is the density of state j under the distribution π [Gilks, Richardson, and Spiegelhalter, 1995]. Importantly, the right hand side of this limit is independent of the initial state i, which therefore indicates the unique stationary distribution is independent of the starting state of the Markov chain.

Having generated $\{\theta_0, \theta_1, \theta_2, \ldots\}$ samples from the target distribution $\phi(\cdot)$, quantities of interest such as posterior mean and variance associated with $\phi(\cdot)$ can be estimated using Monte Carlo integration. Monte Carlo integration evaluates integrals of the form $\mathbb{E}\left[f\left(\theta\right)\right] = \int f(\theta)\phi(\theta)d\theta$ by drawing samples $\{\theta_t, t = 1, \ldots, n\}$ from ϕ, and approximating $\mathbb{E}\left[f\left(\theta\right)\right]$ as:

$$\mathbb{E}\left[f\left(\theta\right)\right] \approx \frac{1}{n}\sum_{t=1}^{n} f\left(\theta_t\right).$$

We now present a few MCMC methodologies that allows one to generate samples from the target distribution ϕ.

2.2.1 Metropolis Hastings

The Metropolis-Hastings (MH) algorithm [Metropolis et al., 1953] is a method used to generate samples from some target distribution $\phi(\cdot)$, and does so by constructing a Markov chain whose stationary distribution $\pi(\cdot)$ is exactly $\phi(\cdot)$. The algorithm is as follows:

Metropolis-Hastings algorithm:

Given a current state θ_t at iteration t, the next state θ_{t+1} is selected by first sampling a candidate point $\theta^* \sim q(\cdot \mid \theta_t)$, where $q(\cdot \mid \theta_t)$ is some proposal distribution. The candidate point θ^* is accepted as the new state θ_{t+1} with probability $\alpha(\theta_t, \theta^*)$ which is given by:

$$\alpha(\theta_t, \theta^*) = \min\left(1, \frac{\phi(\theta^*)q(\theta_t \mid \theta^*)}{\phi(\theta_t)q(\theta^* \mid \theta_t)}\right)$$

If the candidate point θ^* is accepted, then $\theta_{t+1} = \theta^*$ and otherwise $\theta_{t+1} = \theta_t$. The MH algorithm is guaranteed to converge to the stationary distribution $\pi(\cdot)$ subject to conditions on $q(\cdot \mid \cdot)$ [Roberts and Smith, 1994]. This algorithm produces a sequence of $\{\theta_0, \theta_1, \theta_2, \ldots\}$ samples such that, once the first instance θ_t is sampled from $\pi(\cdot)$, then all subsequent samples θ_k for $k > t$ will also be from $\pi(\cdot)$, as guaranteed by the positive recurrent condition [Gilks, Richardson, and Spiegelhalter, 1995].

The MH algorithm is equivalent to constructing a Markov chain with transition matrix $T_{t,t+1} = p(\theta_{t+1} \mid \theta_t)$ given by:

$$T_{t,t+1} = q(\theta_{t+1} \mid \theta_t)\alpha(\theta_t, \theta_{t+1}) + \mathbb{1}_{(\theta_{t+1}=\theta_t)} \left[1 - \int q(\theta \mid \theta_t)\alpha(\theta_t, \theta) d\theta\right].$$

This Markov transition matrix satisfies the conditions of irreducibility, aperiodicity and positive recurrence and has stationary distribution $\phi(\cdot)$. In other words, the Metropolis Hasting algorithm converges to our target distribution $\phi(\cdot)$.

2.2.2 Gibbs sampling

An alternative route to posterior inference used in instances when the target distribution $\phi(\cdot)$ is multivariate and has closed-form conditional posterior distributions is Gibbs sampling [Geman and Geman, 1984]. Gibbs sampling generates θ_t samples by sampling each component-wise element of the vector of θ_t from its conditional distribution, subject to keeping the other components of the vector θ_t fixed to their current values. The algorithm is as follows:

Gibbs sampling algorithm:

Take θ_t^i to be the i^{th} coordinate of vector θ_t (of dimension D) and θ_t^{-i} be the vector of all coordinates of θ_t excluding the i^{th} component, i.e. $\theta_t^{-i} = \left(\theta_t^1, \ldots, \theta_t^{i-1}, \theta_t^{i+1}, \ldots, \theta_t^D\right)$. At the t^{th} iteration, the following sequence of samples are taken:

$$\begin{aligned}
\theta_{t+1}^1 &\sim \phi(\theta_t^1 \mid \theta_t^2, \theta_t^3, \ldots, \theta_t^D) \\
\theta_{t+1}^2 &\sim \phi(\theta_t^2 \mid \theta_t^1, \theta_t^3, \ldots, \theta_t^D) \\
&\vdots \\
\theta_{t+1}^D &\sim \phi(\theta_t^D \mid \theta_t^1, \theta_t^2, \ldots, \theta_t^{D-1}).
\end{aligned}$$

This produces a sequence of $\{\theta_0, \theta_1, \theta_2, \ldots\}$ samples that converges to the stationary distribution $\phi(\cdot)$. Gibbs sampling is a popular approach to posterior inference in instances when sampling from the conditional $\phi(\theta_t^1 \mid \theta_t^2, \theta_t^3, \ldots, \theta_t^D)$ is manageable. The Gibbs sampler is a special case of the MH algorithm with a proposal distribution $q(\theta_{t+1} \mid \theta_t) = \phi(\theta_t^i \mid \theta_t^{-i})$ [Geyer, 1998], and thus guarantees algorithm convergence by satisfying the necessary Markov chain conditions of irreducibility, aperiodicity and positive recurrence.

2.2.3 Hamiltonian Monte Carlo

Another MCMC methodology aiming to efficiently sample $\phi(\cdot)$ is the Hamiltonian Monte Carlo algorithm (HMC) [Duane, Kennedy, Pendleton, and Roweth, 1987]. HMC uses the principles of Hamiltonian dynamics to describe the evolution of a physical system as a function of its state pair (q, p), where q is the position and p is the momentum of the system. The system is then defined by:

$$\frac{dp}{dt} = -\frac{\partial H}{\partial q}, \quad \frac{dq}{dt} = -\frac{\partial H}{\partial p} \tag{2.4}$$

where $H(q, p)$ is the *Hamiltonian function.* Frequently, $H(q, p)$ coincides with the total energy of the system and assumed to take the form $H(q, p) = U(q) + K(p)$, where $U(q) =$ *potential energy* and $K(q) =$ *kinetic energy.* Crucially, (2.4) fully describes the trajectory of motion such that, for time t', the $H(q, p)$ defines a mapping from any state (q, p) at time t to state (q', p') at time $t' + t$.

The key to connecting Hamiltonian dynamics with MCMC is to construct a Hamiltonian function $H(q, p)$ in terms of the target distribution $\phi(\cdot)$. In particular, HMC assumes $H(q, p)$ takes the form $U(\theta_t) = -\log(\phi(\theta_t))$, and typically assumes $K(p) = -p^T M^{-1} p/2$ where M is a symmetric, positive-definite matrix. There are many functional forms that $K(p)$ can take, for further discussion refer to [Betancourt, 2017]. This produces the following dynamics:

$$\frac{dp}{dt} = -\frac{\partial U}{\partial \theta}, \quad \frac{d\theta}{dt} = M^{-1}p.$$

To simulate the evolution of this system it is necessary to discretise time. The dynamics can be approximated with arbitrary precision by solving these differential equations using Euler's or Leapfrog methods. The HMC algorithm is as follows:

Hamiltonian Monte Carlo algorithm:

At the t^{th} iteration, sample a momentum variable $p' \sim N(0, M)$. Then with p' and q_t, simulate Hamiltonian dynamics from:

$$\frac{dp}{dt} = -\frac{\partial U}{\partial \theta}, \quad \frac{d\theta}{dt} = M^{-1}p$$

with the leapfrog method or some equivalent method to solve a set of differential equations to produce a final state (θ', p'), and take $(\theta^*, p^*) = (\theta', -p')$. With the proposed state (θ^*, p^*), a

MH step is performed with acceptance probability:

$$\alpha(\theta,\theta^*) = \min\left(1, \exp\left(-H(\theta^*,p^*) + H(\theta_t,p_t)\right)\right) = \min\left(1, -U(\theta^*) + U(\theta_t) - K(p^*) + K(p_t)\right).$$

If the candidate point is accepted, set $\theta_{t+1} = \theta^*$ otherwise $\theta_{t+1} = \theta_t$. This produces a Markov chain that converges to $\phi(\cdot)$. For further HMC details and review, refer to [Neal et al., 2011].

2.3 Implementation details

The algorithms covered in subsections 2.2.1, 2.2.2 and 2.2.3 all produce sequences of random samples $\{\theta_0, \theta_1, \theta_2, \ldots\}$ eventually converging to the target distribution $\phi(\cdot)$. We now discuss approaches used to assess convergence and factors affecting the rate of convergence. There are many diagnostic measures used to indicate MCMC convergence, such as the Geweke diagnostic, Heidelberger and Welch statistic and Augmented Dicker-Fuller [Geweke et al., 1991, Heidelberger and Welch, 1981, Godfrey, 1978] to mention a few. It is important to note however, that MCMC diagnostic tools are only indicators of MCMC convergence, rather than 'proving' convergence. For a deeper review of these MCMC convergence assessments, refer to [Cowles and Carlin, 1996]. During this work, we use the following diagnostic approaches to assess MCMC convergence:

MCMC diagnostics:

The core interest is in establishing whether the samples $\{\theta_0, \theta_1, \theta_2, \ldots\}$ have converged to the stationary distribution $\phi(\cdot)$, and thus, originate from a single distribution. Throughout this work, we use the following criteria to assess MCMC convergence:

- **Trace plots**: These plot the sequence of sampled θ_t values against t (MCMC iteration index). Once an MCMC algorithm has converged to the stationary distribution, samples should look like they originate from a single distribution. Furthermore, multiple MCMC chains from differing starting values of θ_0 should eventually converge to the same stationary distribution $\phi(\cdot)$.

- **Heidelberger and Welch statistic**: This test is used to determine whether MCMC samples come from a stationary distribution (the null hypothesis). There are two phases to this test:

  1) Phase one: the MCMC samples are iteratively tested for stationarity, i.e. we initially test the whole chain for stationarity, if it passes, then we infer the chain has converged (and move on to the second phase). Otherwise, we successively omit the first $10\%, 20\%, \ldots$

of the samples and test for stationarity until there are less than 50% of the original samples remaining or the chain has at some point passed one of the stationarity tests. If the chain does not pass any of these iterative stationarity tests, we infer non-convergence and do not progress on to the second phase.

2) Phase two: Using the proportion of the chain which passed the previous phase, we calculate a 95% credible interval around the mean using the half-width test.

MCMC tuning:

Producing chains that efficiently converge often requires tuning of the parameters associated with the chosen MCMC methodology. In terms of the previously outlined algorithms, this relates to the chosen proposal distribution used in the MH algorithm, burn-in and chain length. In particular, we focus on the following:

- **MH $q(\cdot \mid \theta_t)$ proposal:** The choice of $q(\cdot \mid \theta_t)$ in the MH algorithm significantly impacts convergence efficiency. Much work has been done in investigating the link between $q(\cdot \mid \theta_t)$ and convergence performance [Neal et al., 2006, Rosenthal et al., 2011, Sherlock et al., 2010], but generally speaking, optimal performance is achieved when the transition kernel is similar to the target distribution $\phi(\cdot)$ [Pasarica and Gelman, 2010, Sun, 2008]. Work has further looked at how incorporating information about the distribution of $\phi(\cdot)$ into $q(\cdot \mid \theta_t)$ can be used to help produce accelerated convergence and reduce correlation between MCMC samples [Roberts and Stramer, 2002, Cai et al., 2008, Turner et al., 2013].

- **Burn-in:** As discussed earlier, once the first instance θ_t is sampled from the stationary distribution $\pi(\cdot)$, all subsequent samples are also from the stationary distribution. It is however, not always straight-forward to know when this first sample from the stationary distribution is obtained. In theory, depending on the desired level of similarity between $T^n_{t,t+1}$ and $\phi(\cdot)$, the first sample can be analytically calculated [Gilks et al., 1995, Roberts, 1996], but this in practice is computationally infeasible. To circumvent this, a burn-in period is taken, i.e. the initial $t = 1, \ldots, m$ samples are discarded and the remaining samples are assumed to originate from the stationary distribution.

- **Chain length:** Once a chain has converged to its stationary distribution, a decision on how many samples are needed for adequate precision is required. Often, Monte Carlo variance or calculations of the effective sample size based on MCMC chains are performed to assess whether the appropriate number of samples have been obtained.

Throughout this work, Bayesian inference is performed by implementations of 2.2.1, 2.2.2 and 2.2.3. Although the focus of this research is not the inferential procedures used during this work's retail analytics modelling, particular attention is paid to tuning and MCMC chain convergence, which is assessed in line with the previously discussed points.

Chapter 3

Bayesian Nonparametrics

This Chapter gives an overview of the Bayesian nonparametric mixture modelling paradigm. We provide definitions of parametric and nonparametric models, with particular reference to the Dirichlet process, and discuss the motivations of using nonparametric models when handling complex and non-linear datasets. We further provide various relevant expositions of the Dirichlet process, and briefly mention approaches used for posterior inference for the Dirichlet process.

3.1 Bayesian Nonparametric mixture models

During section 2.1, we introduced the notion of probability models and the broad framework that statisticians interpret data through. To refresh the reader, the model (2.2) specified the Bayesian paradigm of statistical modelling which supposed a sequence $y_1, y_2, ..., y_n$ of instances drawn independently and identically from the probability models P_θ (with distribution $F(\cdot \mid \theta)$), indexed by parameters $\theta \in \Theta$ (with parameter space Θ), can be expressed as the following:

$$\begin{aligned} &y_i \overset{iid}{\sim} F(y \mid \theta), \text{ for } i = 1, ..., n \\ &\theta \sim \pi \end{aligned} \tag{3.1}$$

where π is the distribution over parameter space Θ.

We now introduce the notions of parametric and nonparametric models in relation to the model (3.1). We say the model has a *parametric prior* if $\Theta \subset \mathcal{L}$, where $\mathcal{L}$ is a *finite linear spanning set.* A set $\mathcal{L}$ is a *finite linear spanning set* if there exists a $k \in \mathbb{N}$ and elements $l_1, \ldots, l_k$ of $\mathcal{L}$, such that, all elements of $\mathcal{L}$ can be expressed as a linear combination of elements $l_1, \ldots, l_k$ [Vallejos, 2008]. Consequently, parametric models bound the dimension of the solution space, irrespective of the number of samples being modelled. This can lead to model misfit when data is characterised by heterogeneity beyond probability models with a finite parameter space.

Motivated by the shortcomings of parametric models at handling complex data, nonparametric models can be devised to circumvent these aforementioned issues. We define a nonparametric model as one with an infinite dimensional parameter space [Orbanz and Teh, 2011, Müller et al., 2004], and for such a parameter space Θ, there exists no k and corresponding elements of Θ such that these elements span Θ. Thus, in order for model (2.2) to be Bayesian nonparametric, it is necessary to specify an infinite dimensional prior over the space Θ. One of the main strengths of nonparametric models over parametric models is their capability of capturing any distribution of the data. The caveat of this flexibility is the computational cost of requiring more data to adequately infer the correct model structure. This can be intuited from the following example; suppose one fits a simple linear regression to data. In this situation, inference for such a model requires less data as one borrows strength from the assumption the data lies on a straight line. In the case of a nonparametric model, no such linearity assumption is made and consequently, linearity is inferred from the data. This results in nonparametric methods requiring more data.

A popular branch of Bayesian nonparametrics (BNP) models are Bayesian nonparametric priors. Bayesian nonparametric priors can be thought of as the prior over probability measures, i.e. the distribution over distributions. Such nonparametric priors have useful applications to mixture modelling. Extending (2.2) to a mixture model can be expressed hierarchically as:

$$\begin{aligned} y_i|\theta_i &\overset{ind.}{\sim} F(y|\theta_i), \text{ for } i = 1, ..., n \\ \theta_i|G &\overset{i.i.d.}{\sim} G \\ G &\sim G_0 \end{aligned} \tag{3.2}$$

where G_0 is some nonparametric prior over countable measures, and G is an instance of such a measure. The *Dirichlet process* is an example of such a nonparametric prior over countable measures, and is widely used in fields such as finance, medicine and survival analysis [Kottas, 2013, De Iorio et al., 2009, Wade et al., 2014].

The forthcoming sections of this chapter introduce the *Dirichlet distribution* along with its basic properties. The Dirichlet distribution can be considered a semi-parametric prior, and will provide the reader with the prerequisite material to understand subsequent content. Latter sections will move on to defining the Dirichlet distribution's nonparametric extension, the *Dirichlet process*, along with its properties and applications relevant to this work.

3.2 Notation & mathematical background

We introduce some prerequisite terminology and mathematical definitions needed to formally define the Dirichlet Process. In particular, we define the key concepts of a *σ-algebra*, a *measurable space*, a *probability measure* and the measurable function known as the *Dirac measure*.

Definition 4. *σ-algebra*

We say $\mathcal{B}$ is a *σ-algebra* on the set $\mathcal{X}$ if it satisfies the following:

1. $\mathcal{X} \in \mathcal{B}$.

2. If S is in $\mathcal{B}$, then $S^c \in \mathcal{B}$.

3. For all countable collections $\{E_i\}_{i\in\mathbb{N}} \in \mathcal{B}$, then $\cup_{i\in\mathbb{N}} E_i \in \mathcal{B}$.

Definition 5. *Measurable space*

We say the pair $(\mathcal{X}, \mathcal{B})$ is a *measurable space* if $\mathcal{B}$ defines a *σ-algebra* on the set $\mathcal{X}$.

Definition 6. *Probability measure*

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. We say a function $\mu : \mathcal{B} \to [0, 1]$ is a probability measure if the following are satisfied:

1. $\forall E \in \mathcal{B},\ \mu(E) \geq 0$.

2. $\mu(\varnothing) = 0$.

3. $\mu(\mathcal{B}) = 1$.

4. For all countable collections $\{E_i\}_{i\in\mathbb{N}}$ of pairwise disjoint sets in $\mathcal{B}$, we have $\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{k=1}^{\infty} \mu(E_i)$.

Definition 7. *Dirac measure δ_x*

The Dirac measure δ_x on a measurable space $(\mathcal{X}, \mathcal{B})$ is defined such that, for any measurable set $A \subset \mathcal{B}$ and $x \in \mathcal{X}$:

$$\delta_x(A) = \begin{cases} 0,\ x \notin A \\ 1,\ x \in A \end{cases}$$

Crucially, a probability distribution function can be thought of as a probability measure, and a *σ-algebra* can be informally thought of as the sensible, non-paradoxical sets upon which traditional probability distributions are defined.

3.3 Dirichlet distribution

The Dirichlet distribution is a popular distribution under the Bayesian formulation, and is frequently used as a semi-parametric prior π over a parameter space Θ. Informally, the Dirichlet distribution can be thought of as a prior over finite probability mass functions, and is defined as follows:

Definition 8. *Dirichlet distribution*

Let $P = (P_1, ..., P_k)$ be the random components of a probability mass function, i.e. $P_i \geq 0$ for $i = 1, ..., k$ and $\sum_i^k P_i = 1$. We say P is distributed according to a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$ where $\alpha_i \geq 0$ for $i = 1, ..., k$, denoted as $P \sim \text{Dir}(\boldsymbol{\alpha})$, if its probability mass function $f(P_1, \ldots, P_k \mid \boldsymbol{\alpha})$ satisfies:

$$f(P_1, \ldots, P_k \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{k} P_i^{\alpha_i - 1} \tag{3.3}$$

where $\Gamma(t) = \int_0^\infty x^{t-1} e^x dx$. For $k = 2$, the Dirichlet distribution reduces to the $\text{Beta}(\alpha_1, \alpha_2)$ distribution.

An important observation of the Dirichlet distribution is its domain space; since $P = (P_1, \ldots, P_k) \sim \text{Dir}(\boldsymbol{\alpha})$ where $\sum_i^k P_i = 1$ such that $P_i \geq 0$ for each i, it follows that P is a probability mass function. Thus, the Dirichlet distribution can be thought of as a prior distribution over the space of finite probability mass functions. The Dirichlet distribution has the following properties:

Dir($\boldsymbol{\alpha}$) moments:

The mean and covariance of $\text{Dir}(\boldsymbol{\alpha})$ are given by:

$$\mathbb{E}(P_i) = \frac{\alpha_i}{\sum_{i=1}^k \alpha_i} \tag{3.4}$$

$$\text{Cov}[P_i, P_j] = \frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \ (i \neq j) \tag{3.5}$$

respectively, where $\alpha_0 = \sum_{i=1}^k \alpha_k$. These expressions provide insight into how the parameters $\boldsymbol{\alpha}$ contribute to samples generated from $\text{Dir}(\boldsymbol{\alpha})$ in that, the α_i control the relative likelihood of the i^{th} component and further contribute to the correlation between components.

Multinomial conjugacy:

A key property of the Dirichlet distribution is its conjugacy with the multinomial distribution. The multinomial distribution is defined as follows; suppose we take n independent samples of k mutually exclusive categories with the i^{th} category having success probability P_i ($i = 1, \dots, k$), then the multinomial distribution is the distribution over the number of occurrences x_i of category i for $i = 1, \dots, k$, and has probability mass function given by:

$$f(x_1, \dots, x_k | n, P_1, \dots, P_k) = \frac{n!}{x_1! x_2! \dots x_k!} \prod_{i=1}^{k} P_i^{x_i}. \tag{3.6}$$

We denote this distribution as Multinomial(k, n, P). Importantly, suppose $P \sim \text{Dir}(\boldsymbol{\alpha})$ and $X \mid P \sim$ Multinomial(k, n, P) where $X = (x_1, \dots, x_k)$, then the posterior distribution is given by $P|X \sim \text{Dir}(\boldsymbol{\alpha} + X)$. This is a significant result, as it means the Dirichlet distribution is the conjugate distribution over the space of finite probability mass functions.

Aggregation property of the Dirichlet distribution:

Our final Dirichlet distribution property is the aggregation property. This property can be informally thought of as the effect of clumping together different parts of probability space, i.e. if one combines non-intersecting sectors of the space, then one has a Dirichlet distribution over the new augmented space. More concretely, if one has a partition $\{A_1, ..., A_r\}$ of the set $\{1, ..., k\}$, then:

$$\left(\sum_{i \in A_1} Q_i, \sum_{i \in A_2} Q_i, \dots, \sum_{i \in A_r} Q_i \right) \sim Dir \left(\sum_{i \in A_1} \alpha_i, \sum_{i \in A_2} \alpha_i, \dots, \sum_{i \in A_r} \alpha_i \right).$$

This property offers an intuition into the formulation of the Dirichlet distribution, and also provides utility with respect to inferential methods which we will allude to in latter sections.

3.4 Dirichlet process

Ferguson [1973], motivated by the difficulty of specifying nonparametric priors in the Bayesian framework, introduced the Dirichlet process as a class of nonparametric prior distributions. The Dirichlet process can be thought of as the prior distribution over countable measures, or the distribution over distributions. Ferguson argued that any Bayesian nonparametric prior should exhibit the following two appealing characteristics:

1. The support for the prior is large.
2. The posterior distributions given a sample of observations from the true probability distribution should be analytically manageable.

Definition 9. *Dirichlet process*

We say G is distributed according to a Dirichlet process with parameters G_0 (base distribution) and ν (scale), denoted as $G \sim \mathrm{DP}(\nu G_0)$, if for any *$\sigma$-algebra* $\mathcal{B}$ of the space Θ, and given any finite partition $A_1, ..., A_k \subset \mathcal{B}$, we have the following property:

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}(\nu G_0(A_1), \ldots, \nu G_0(A_k))$$

for any $k \in \mathbb{N}$.

Ferguson [1973] proved the existence of such a stochastic process sharing these DP criteria by verifying the Kolmogorov consistency theorem [Kolmogorov, 1933]. The Dirichlet process has the following key properties:

Discreteness:
Samples of $\mathrm{DP}(\nu G_0)$ are discrete random measures. Ferguson [1973] proved their discreteness by using an involved analysis of Gamma processes. Crucially, this discreteness allows for mixture modelling, as the probability of two samples coinciding from a G measure is non-zero, thus allowing a cluster interpretation.

Realisations of $G \sim \mathrm{DP}(\nu G_0)$ are random probability measures:
This is verified by an alternative definition provided by Ferguson [1973], which expresses a Dirichlet process G as:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$$

with $p_i = j_i / \sum_{l=1}^{\infty} j_l$, where j_l are random variables constructed from the distribution functions $P(j_1 \leq x_1) = e^{N(x_1)}$, $N(x) = -\alpha \int_x^\infty e^{-y} y^{-1} dy$ where $\alpha > 0$, and $P(j_k \leq x_j | j_{k-1} = x_{k-1}, \ldots, j_1 = x_1) = e^{N(x_k) - N(x_{k-1})}$ for $0 < x_k < x_{k-1}$ and $j \geq 2$. They showed that this construction is a random probability measure and is in fact, a Dirichlet process. Ferguson further showed that for DP(νG_0) defined over measurable space $(\Theta, \mathcal{B})$, and Q a fixed probability measure on $(\Theta, \mathcal{B})$ with $Q \ll \nu G_0$, then for any m and measurable sets $A_1, \ldots, A_m$, and $\epsilon > 0$:

$$P\left(\left|G\left(A_j\right) - Q\left(A_j\right)\right| < \epsilon \text{ for } j = 1, \ldots, k\right) > 0.$$

These two results have the following important implications; firstly, it verifies that G is a probability measure on the measurable space $(\Theta, \mathcal{B})$ and secondly, it demonstrates the fulfilment of the objective that the nonparametric prior support is sufficiently large. However, it is important to note that this alternative definition is not constructive, i.e. we are unable to generate samples from this definition due to the normalisation constant $\sum_{l=1}^{\infty} j_l$.

Moments of DP(νG_0):

The mean and variance of DP(νG_0) are given by:

$$\mathbb{E}(G(A)) = G_0(A)$$

$$\text{Var}(G(A)) = \frac{G_0(A)(1 - G_0(A))}{1 + \nu}$$

respectively. These moments give us a sense of how the parameters of DP(νG_0) contribute to $G \sim$ DP(νG_0) realisations; i.e. G_0 controls where realisations are centred, and ν controls the degree to which realisations are *close* to G_0.

DP conjugacy:

A key property of the DP is its conjugacy with the multinomial. Since G is a random measure, we can sample $\theta_i \overset{iid}{\sim} G$ for $i = 1, \ldots, n$. The posterior of G given observed values $\theta_1, ..., \theta_n$ is then given by:

$$G|\theta_1, \ldots, \theta_n \sim \text{DP}\left(\nu + n, \frac{\nu}{\nu + n} G_0 + \frac{n}{\nu + n} \frac{\sum_{i=1}^n \delta_{\theta_i}}{n}\right).$$

This means the DP is the conjugate prior over distributions that are closed under posterior updates given observations [Teh, 2011]. This has a theoretical significance, since it meets Ferguson's second criterion, that posterior distributions given a sample from the true random measure from the DP should be tractable. This has practical significance, as it gives one routes into developing straight-forward procedures for posterior inference.

3.4.1 Blackwell-MacQueen urn scheme

Blackwell and MacQueen [1973] used a Polya urn scheme to provide an alternative definition of the Dirichlet process. This provided a generative model for producing random measure instances as specified by Ferguson's Dirichlet process. More concretely, the DP is re-expressed as the following; given samples $\theta_1, \dots, \theta_n$ of G where $G \sim \mathrm{DP}(\nu G_0)$, G is integrated out of the joint distribution of $\theta_1, \dots, \theta_{n+1}$. The posterior predictive distribution then becomes:

$$\begin{aligned} \theta_1 &\sim G_0 \\ \theta_{n+1}|\theta_1, \dots, \theta_n &\sim \frac{\nu}{\nu+n} G_0(\theta_{n+1}) + \frac{1}{\nu+n}\sum_{i=1}^{n} \delta_{\theta_i}(\theta_{n+1}). \end{aligned} \tag{3.7}$$

Crucially, Blackwell and MacQueen [1973] establish two key properties of the limiting distribution of G. Firstly, that as $n \to \infty$, the urn scheme converges with probability 1 to a discrete random measure G, and secondly, that this G is identical to a sample from $\mathrm{DP}(\nu G_0)$. From the joint distribution induced from $p(\theta_1, \dots, \theta_{n+1})$, and along with the definition of exchangeability and De Finetti's theorem, Blackwell and MacQueen [1973] establish critically that $\theta_1, \dots, \theta_n|G \sim G$ is equivalent to the DP and that these $\theta_1, \dots, \theta_{n+1}$ are i.i.d draws from G.

There is a clustering property implied from (3.7). By denoting $\theta_1^*, \dots, \theta_{n^*}^*$ as the set of unique values of $\theta_1, \dots, \theta_n$, we then can rewrite (3.7) as:

$$\begin{aligned} \theta_1 &\sim G \\ \theta_{n+1}|\theta_1, \dots, \theta_n &\sim \frac{\nu}{\nu+n} G(\theta_{n+1}) + \frac{1}{\nu+n}\sum_{i=1}^{n^*} n_i \delta_{\theta_i^*}(\theta_{n+1}) \end{aligned} \tag{3.8}$$

where n_i is the number occurrences of θ_i^*, and n^* is the number of unique atoms of $\theta_1, \dots, \theta_n$. (3.8) demonstrates the clustering structure and preferential attachment process at play, as the probability of the sample θ_{n+1} being assigned to the atom θ_i^* is $\frac{n_i}{\nu+n}$. This illustrates that there is a non-zero probability of being assigned to an existing cluster, as well as an increased probability for atoms to be added to larger clusters.

Antoniak [1974] investigated the expected number of unique mixture components that a sample $\theta_1, \dots, \theta_n \sim G$ induces. They derived:

$$\begin{aligned} \mathbb{E}(n^*|\nu) =& \sum_{i=1}^{n} \frac{\nu}{\nu+i-1} \\ \approx& \nu \log(\frac{\nu+n}{\nu}). \end{aligned} \tag{3.9}$$

This shows the number of unique clusters n^* grows logarithmically in n, which indicates the clustering is well defined, and allows straightforward inference to be developed around the ν parameter. The DP shares connections with the Chinese Restaurant Process (CRP) [Aldous, 1985]. The CRP is a stochastic process which defines a distribution over partitions of the integers $\{1, \dots, n\}$. The CRP can be intuited by the process of customers attending a restaurant and selecting tables to sit at. This is as follows; the first customer sits at the first table, and n^{th} customer chooses table c_n, where probability of selecting the k^{th} table is probability $\frac{n_k}{\nu+n-1}$, with n_k being the number of customers at the k^{th} table, and selects a new table with probability $\frac{\nu}{\nu+n-1}$, i.e.:

$$p(c_n \text{ table choice} \mid K \text{ partitions }) = \begin{cases} k^{th} \text{ table with probability } \frac{n_k}{\nu+n-1}, \\ \text{new table with probability } \frac{\nu}{\nu+n-1}, \end{cases}$$

This is exchangeable with respect to the customer labels and the tables they sit at. This possesses a close resemblance with (3.8), which becomes clearer from the relabelled expression (3.7):

$$p(c_{n+1}|c_1, \dots, c_n) = \frac{\nu}{\nu+n}\delta(c_{n+1} = K+1) + \sum_{k=1}^{K} \frac{n_k}{\nu+n}\delta(c_{n+1} = k). \tag{3.10}$$

As with the DP, the CRP exhibits a preferential attachment process, as well as the distribution over the table partitions being the same as the distribution over the cluster sizes [Teh, 2011].

3.4.2 Sethuraman stick-breaking construction

Sethuraman [1994] provided an alternative constructive definition of the DP, known as the stick-breaking construction. This is as follows; given G satisfying:

$$\begin{aligned} G &= \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}, \\ \beta_i &\overset{i.i.d.}{\sim} Beta(1, \nu), \\ \pi_i &= \beta_i \prod_{l=1}^{i-1} (1 - \beta_l), \end{aligned} \tag{3.11}$$

where $\theta_i \overset{i.i.d.}{\sim} G$, then $G \sim \text{DP}(\nu G_0)$. This construction has connections with (3.7), as it provides an intuitive distribution over the cluster partitions. More precisely, Sethuraman [1994] verified that the probability of the draw θ_1 of (3.7)'s urn scheme as the same probability of the partition denoted by π_1 in (3.11). This construction has useful applications to density estimation and mixture modelling.

3.5 Dirichlet process applications

Having formally introduced the DP, its various expositions and theoretical properties, we now describe the applications of the DP. In-particular, we discuss DP's applications to mixture modelling and density estimation.

3.5.1 Dirichlet process mixture model

The Dirichlet process mixture model (DPMM) was proposed by Antoniak [1974] as a mixture model with a DP prior over the random mixing distribution. The DPMM can be hierarchically expressed as:

$$\begin{aligned} y_i|\theta_i \sim& F(\theta_i) \\ \theta_i|G \sim& G \\ G \sim& \mathrm{DP}(\nu G_0). \end{aligned} \tag{3.12}$$

As discussed in section 3.4.1, the samples $G \sim \mathrm{DP}(\nu G_0)$ induce a clustering structure over the $\theta_1, ..., \theta_n \sim G$, which in turn induce a partition over the responses $y_1, ..., y_n$. An example of (3.12) is given by (3.13), with $\mathrm{N}(\cdot, \sigma_1)$ as the kernel, and base distribution $G_0 = \mathrm{N}(0, \sigma_2)$:

$$\begin{aligned} y_i|\theta_i &\sim \mathrm{N}(\theta_i, \sigma_1^2) \\ \theta_i|G &\sim G \\ G &\sim \mathrm{DP}(\nu G_0) \end{aligned} \tag{3.13}$$

where σ_1, σ_2 are fixed constants. The DPMM of (3.12) can be further re-expressed using allocation variables as:

$$\begin{aligned} y_i|\theta_i &\sim F(y_i|\theta_{Z_i}) \\ Z_i|G_0 &\sim \sum_{j=1}^{\infty} \pi_j \delta_j(.) \\ G_0 &= \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \\ \beta_i &\overset{i.i.d.}{\sim} Beta(1, \nu) \\ \pi_i &= \beta_i \prod_{l=1}^{i-1} (1 - \beta_l) \\ \theta_i &\sim G_0 \end{aligned}$$

where Z_i denotes the index that the i^{th} data point has been allocated to. Though the observations $y_i, ..., y_n$ are continuous, by discreteness of G, we interpret two observations y_i, y_j being clustered together if they share the same allocation variable, i.e. $Z_i = Z_j$, and therefore the same atom $\theta_i = \theta_j$. Furthermore, if $Z_i \neq Z_j$ for $i \neq j$, it follows almost surely that $\theta_i \neq \theta_j$, since the θ_i realisations are unique for continuous G_0. This guarantees different allocations have different atoms. We derive the probability of the responses y_i, y_j for $i \neq j$, belonging to the same cluster as follows; from the exchangeability of labels of i, j, we can therefore relabel them as $i = 1, j = 2$. Thus, from (3.7) we obtain:

$$\theta_2 | \theta_1 \sim \frac{\nu}{\nu+1} G(\theta_1) + \frac{1}{\nu+1} \delta_{\theta_1}.$$

Hence, the probability of responses y_1, y_2 belonging to the same cluster as $\frac{1}{\nu+1}$.

The parameters G_0, ν of DP play a significant role in mixture modelling. For fixed G_0, increasing ν decreases the variance of $G \sim \text{DP}(\nu G_0)$ samples. Consequently, ν ought to reflect the strength of an experimenter's belief around which samples G are centred around the measure G_0.

3.5.2 Density estimation

The DP has further applications to density estimation [Lo et al., 1984, Neal, 1992, Escobar and West, 1995, Rasmussen, 1999]. Density estimation essentially estimates some unknown distribution F, such that:

$$x_i \sim F(\cdot) \quad \text{for } i = 1, \ldots, n$$

given the samples $x_1, \ldots, x_n$. The Bayesian framework requires placing a prior distribution over the functional space of F. Parametric methods traditionally restrict the functional space by assuming the space can be expressed as a finite combination of mixing densities (of a known family). Bayesian nonparametric approaches however, assume an infinite dimensional prior over this functional space, examples of which include Polya trees, Berstein polynomials and other variants of the Dirichlet process. For the purposes of this work, we will focus on the DP's utility to density estimation.

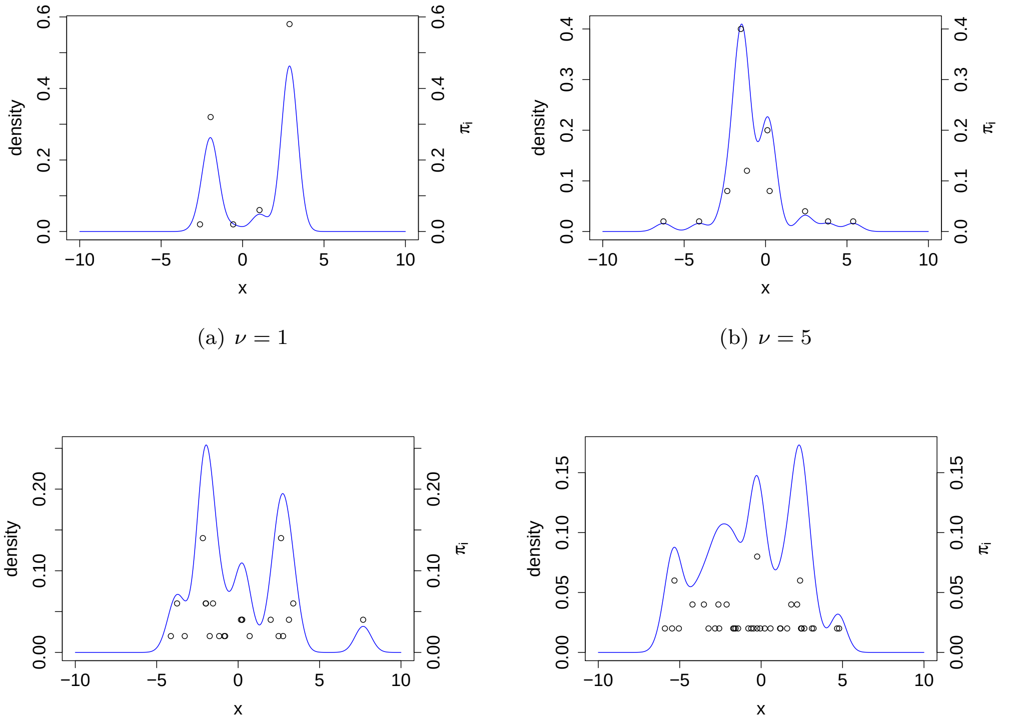
The DP's approach to estimating a density $p(x)$ involves convolving a family of kernels $f(x|\theta)$ (parametrised by θ) with a DP prior over the mixing proportions to produce a countably infinite mixture of smooth kernel functions. More concretely, suppose we aim to estimate a density $p(x)$ as a DP mixture of densities $f(x|\theta)$, we then write $p(x)$ as:

$$p(x) = \int f(x|\theta) dG_0(\theta) \tag{3.14}$$

where $G \sim \mathrm{DP}(\nu G_0)$. An alternative way of expressing (3.14) is by using the stick-breaking construction of (3.11) where given $G \sim \mathrm{DP}(\nu G_0)$, (3.14) can be rewritten as:

$$\begin{aligned} p(x) &= \int f(x|\theta) dG_0(\theta) \\ &= \int f(x|\theta) d(\sum_i^\infty \pi_i \delta_{\theta_i})) \\ &= \sum_{i=1}^\infty \pi_i f(x|\theta_i). \end{aligned} \tag{3.15}$$

Thus, we can smoothly estimate the density of $p(x)$ as countably infinite mixture of $f(x|\theta_i)$ densities, where ν contributes to the degree of smoothness; with small ν producing smoother estimates and lumpier estimates otherwise. Figure 3.1 provides examples of measures generated from urn process of (3.7), along with the corresponding densities produced from a mixtures of normal kernels $f(\cdot \mid \theta, 0.5)$ when these measures are priors over θ. Here we use the base distribution $G_0 = \mathrm{N}(0,\ 3.0)$. We notice increasing ν in turn increases the number of unique atoms produced from the DP, as well as increasing the multi-modal behaviour of the mixtures. The R code producing these simulations are included in appendix A.1.



(a) $\nu = 1$

(b) $\nu = 5$

(c) $\nu = 15$

(d) $\nu = 40$

Figure 3.1: Simulated DP measures produced from the urn process of (3.7) for 50 iterations, along with the corresponding densities produced from a mixtures of normal kernels $f(\cdot \mid \theta, 0.5)$ when these measures are used as priors over θ. These measures and mixture of densities are produced over various ν values. The R code producing these plots is included in A.1.

3.6 Posterior inference

Broadly speaking, there are two routes for posterior inference used in DPMMs, marginal and conditional approaches, each relying on the differing representations of a DP. Marginal approaches involve integrating out the infinite dimensional measure of G, and then by utilising the property of (3.7) and the exchangeability of atoms θ_i, then takes Gibbs samples from the distributions of $\theta_i \mid \theta_{-i}$ (where θ_{-i} is the vector of locations excluding the i^{th} atom) for each i. This inferential procedure was made possible by the initial work of Escobar and West [1995], and thus, represents the first step made towards enabling DPs to be an applicable methodology. Neal [2000] provides an excellent summary of marginal approaches. Conditional samplers, first proposed by Ishwaran and Zarepour [2000], do not marginalise out the infinite random measure of G, but instead involve imputing G, then sampling the cluster assignments for each of the location atoms θ_i from their posteriors. One of the initial challenges conditional approaches had was handling the infinite dimensional nature of G. Initial approaches relied on making finite approximations of a DP [Ishwaran and James, 2001, 2003], but extensions of conditional samplers have included innovations around slice and retrospective samplers, which have offered marked improvements in many more challenging DP inferential contexts [Walker, 2007, Papaspiliopoulos and Roberts, 2008, Kalli et al., 2011, Hastie et al., 2015]. Although marginal and conditional methodologies are amongst the more popular approaches to DP inference, other methodological procedures exist, including sequential greedy search algorithms and variational approaches to name a few [Blei, Jordan, et al., 2006, Wang and Dunson, 2011].

For the purposes of this work, we will focus on marginal approaches for posterior inference of DPMMs. Our motivations for orienting our inferential procedure around a marginal method, over other conditional approaches, are the following two closely related reasons. Firstly, marginal approaches circumvent the issues of the infinite dimensional nature the DP by marginalising over the random measure G. By exploiting this property of (3.7), and thus maintaining the infinite dimensional nature of the DP, we maintain a particularly favourable feature of the DP that is worth keeping. This is especially valuable in cases when the number of unique clusters characterising the data is uncertain, such as is with our retail analytics context where we devising a DPMM around unfamiliar and novel data. Our second, closely related reason, is the computational straightforwardness of the implementation of marginal inferential procedures, whilst still maintaining the trait of possessing the infinite dimensional nature of the DP. An important caveat of this is that, although more advanced conditional samplers exist that make it possible to maintain the infinite dimensional property of the DP, for example, Papaspiliopoulos and Roberts [2008]'s retrospective sampler approach or Walker [2007]'s slice sampler (and other related methods), they often involve nuanced and complicated updating strategies that make these procedures non-trivial to implement. Thus, although these approaches, and other related methods, are demonstrated to work effectively in complex modelling scenario's, such as profile regression and large data contexts, there implementational complexity for the scale of data we plan to model (at most 1500 data-points) may be unnecessary.

Our particular marginal method our inferential procedure will be largely based on Neal [2000]'s algorithm 8, where G_0 is a non-conjugate prior with respect to the likelihood function of $f(\cdot \mid \theta)$. Their approach uses the DP exposition of (3.8), which then iteratively takes Gibbs samples of the θ_i conditionals, and then updates each of the unique θ_i atoms using a Metropolis-Hasting step. We then use Escobar and West [1995] methodology of updating the scale ν of the DP. More concretely, the algorithm is as follows:

1. Neal [2000]'s approach iteratively samples the locations for each data-point by sampling the multinomial distribution of order n^*+c (where c is the chosen number of auxiliary components). More concretely, the resultant sample of θ_i equates to sampling a multinomial with probabilities:

$$P\left(\theta_i = \theta_k^* \mid \theta_{-i}, y_i, \theta_1^*, \ldots, \theta_{n^*+c}^*\right) \propto \begin{cases} \frac{n_k^*}{n-1+\nu} f\left(y_i \mid \theta_k^*\right) & \text{for } 1 \leq k \leq n^* \\ \frac{\nu/k}{n-1+\nu} f\left(y_i \mid \theta_k^*\right) & n^* < k \leq n^* + c \end{cases}$$

where:

$$\theta_k^* \overset{iid}{\sim} G_0 \text{ for } k = n^* + 1, \ldots, n^* + c$$

2. The θ_k atoms are then updated for each of the unique clusters $k = 1, \ldots, n^*$. This avoids inefficiencies associated with having to pass through extremely low probability states to get to a higher probability states. This can be done by Metropolis Hastings updates.

3. Finally, by specifying $\nu \sim G\left(\tau_1, \tau_2\right)$ and introducing an auxiliary variable γ, enables ν to be Gibbs sampled. Specifically, we take the following samples:

$$\left(\gamma \mid \nu, n^*\right) \sim Beta\left(\nu + 1, n\right)$$

$$\left(\nu \mid \gamma, n^*\right) \sim \pi_\gamma Gamma\left(\tau_1 + n^*, \tau_2 - \log\left(\gamma\right)\right) + \left(1 - \pi_\gamma\right) Gamma\left(\tau_1 + n^* - 1, \tau_2 - \log\left(\gamma\right)\right)$$

where the weights π_γ are defined by $\pi_\gamma / \left(1 - \pi_\gamma\right) = \left(\theta + n^* - 1\right) / \left(n\left(\tau_2 - \log\left(\gamma\right)\right)\right)$. The detailed steps of this facet of the DP inferential procedure will be elaborated on during section 5.3.3.

This DPMM inference methodology is simple and intuitive, and will be the primary methodology for posterior inference of DPMMs during this work. The implementation details of this approach are explained in more depth in subsequent Chapters.

Chapter 4

Elasticity clustering & related methodologies

This Chapter introduces and defines the interest around analysing cross-elasticity coefficients. In particular we discuss the relationship between these cross-elasticity coefficients and the sensitivity of sales of a product with respect to changes in its own price and the prices of competing products. As referenced in section 1.1, retailers are becoming increasingly interested in classifying and segmenting many of their processes, as it allows them to mitigate storage and computational costs as well as providing valuable insights with which can create a competitive advantage. One way this interest manifests itself is in retailers' analysis of the sensitivity of their products' sales to the price changes across competing products.

The subsequent sections are structured as follows: section 4.1 introduces the concept of a product's sales sensitivity in the context of its cross-elasticity coefficients and outlines retailers' motivation in clustering products in terms of their sales sensitivities. In particular we articulate how a product's sales sensitivity is exhibited through these cross-coefficients and specify a class of regression models that such cross-elasticity coefficients can be generated from. The section continues on to outline the data by which our sales sensitivities analysis is motivated. Section 4.2 describes and reviews the traditional approaches used to investigate and analyse these cross-elasticity coefficients in the fields of retail analytics and econometrics, and outlines the methods used to interpret the differences between cross-elasticity coefficients. This section concludes by highlighting the shortcomings of these current approaches.

4.1 Elasticity clustering background

Characterising products by how sensitive their sales are to their competitor prices is of particular interest to supermarkets. The price at which a product is offered to the consumer is arguably one of the most important controls a retailer has. Unsurprisingly, retailers' analytics teams are constantly striving to develop models and inferential methods that provide insights into how price fluctuations that propagate throughout stores will impact the sales of products whose prices have not changed [Persson, 1995, Ferreira et al., 2015, McGill and Van Ryzin, 1999, Joho et al., 2009]. Consequently, retailers are invested in understanding how these price sensitivities are manifested, as they view it as core to their business operations. More concretely, such a price sensitivity analysis and segmentation provides the following benefits to retailers:

1. Understanding how a product's sales are sensitive to its competitor's price changes allows store planners to decide the value of a given display combination, as it provides information on how a product's sales are likely to react to the deviations of prices of other products. Retailers are increasingly interested in fully understanding the triggers around consumers purchasing decisions when faced with various display combinations and multiple product choices [Burke and Leykin, 2014, de Wijk et al., 2016, Bezawada et al., 2009], and retailers are aware that prices are a key factor that drives these decisions. For instance, a poor display combination could be one that consists entirely of products characterised by their sales being primarily driven by the prices of its competition. This would lead to margin cannibalisation - where profit made on one product is offset by the loss of profit of another product. A characterisation of products in terms of their sales sensitivities would allow store planners to circumvent such pitfalls, and generally empower them to make better pricing and display decisions.

2. The exercise of segmenting products in terms of their sales sensitivity can reveal hidden structure and an informative narrative of data that could provide valuable insights. For example, it is common practice of retailers to cluster consumers according to their product preferences; this allows retailers to efficiently summarise consumers' tastes, which can be used to create personalised recommender systems, improve sales forecasts and consumer loyalty [Lawrence et al., 2001, Kashwan and Velu, 2013, Shih and Liu, 2005]. Such a price sensitivity segmentation could allow retailers to improve their personalised services by understanding consumers' purchasing patterns in terms of their preferences to a particular price sensitivity segmentation. Such improvements could provide a significant advantage in the competitive landscape of consumer retail.

3. In addition to the product display and possible consumer personalisation benefits that a sales sensitivity segmentation would provide, a sales sensitivity segmentation would reduce storage costs and would improve the efficiency of this analysis compared to any manual method that compares these sale sensitivities product by product. Theoretically, a supermarket with N products could want to understand all of the sale sensitivities between all products, i.e. how do the sales of a given product change with respect to price changes of every other product. Such an analysis would lead to needing to store $N \times N$ quantities, which can be impractical for large N. Segmentation or clustering methodologies capable of reducing the dimension of such an analysis to a more simple generating process that neatly characterise products into groups or clustering is appealing to retailers. Consequently, where possible, retailers are interested in developing efficient summaries of their data, to mitigate storage and computational costs [Akcay, 2013, Intel, 2014, Sarwar et al., 2002].

Although these aforementioned benefits are by no means exhaustive, it provides the reader an overview of the motivations behind why retailers are interested in product segmentation.

4.1.0.1 Cross-elasticity demand models

Having discussed the desire to categorise products in terms of their sales sensitivities, we now introduce one approach of quantitatively deriving a product's sensitivity in sales with respect to changes to its competitors prices. We do this by introducing the notion of a product's *price elasticity of demand*, which we define as the rate of change of the quantity demanded of a product with respect to changes in its own price. More formally, by defining P as the price of the product and $S(P, \boldsymbol{x})$ as the sales of the product as some function of its price P and other variables $\boldsymbol{x} = (x_1, \ldots, x_k)$, we then define ψ, the *price elasticity of demand*, as:

$$\psi = \frac{\partial S(P, \boldsymbol{x})}{\partial P}. \tag{4.1}$$

This describes the nature in which a product's sales changes with respect to changes in its own price. The majority of products that retailers offer to their markets have elasticity of demands such that $\psi \leq 0$, i.e. increases in prices lead to decreases in sales. Importantly, more negative ψ implies that small increases in the price leads to large reductions in sales, hence indicating a product that is highly price sensitive.

We can extend this definition to the concept of a product's sensitivity in sales with respect to changes in price of another product. We define a *product i's cross elasticity of demand with* *respect to product* j, as the rate of change between the quantity demanded of product i with respect to a change in the price of product j. More concretely, by defining P_i as the price of product i, P_j as the price of product j, $S_i(P_j, \boldsymbol{x})$ as the sales of product i as some function of P_j and other variables $\boldsymbol{x}$, we then define the χ_{ij}, the *product i's cross elasticity of demand with respect to product* j, as:

$$\chi_{ij} = \frac{\partial S_i(P_j, \boldsymbol{x})}{\partial P_j}. \tag{4.2}$$

Similarly as before, this encapsulates the effect that changes in a product's prices affects another product's sales. The majority of products that retailers offer to their markets have cross-elasticity of demands such that increases/decreases in another product's prices lead to increase/decreases in sales of another product. As before, the larger χ_{ij} is, the more sensitive product i's sales are to changes in product j's price.

Having defined (4.1) and (4.2) as a concept of a sales sensitivity measures, we now introduce a method of calculating these quantities. Importantly, we need to assume a functional form for $S_i(P_j, \boldsymbol{x})$. There are many approaches that link the relationship between price elasticity of demand with sales, such as market share models, attraction models, structural equation and consumer utility modelling [Walters and MacKenzie, 1988, Kim et al., 1999, Leeflang and Parreño-Selva, 2012, Chidmi and Lopez, 2007, Erdem et al., 2008]. Market share and attraction models generally interpret consumers' demand for particular products as having a multiplicative structure with normally distributed errors. The proportion of demand exhibited for each of these products relative to other competitor products is then described as a function of the original attraction of the products [Fok et al., 2002, Cooper and Nakanishi, 1989]. Structural equation modelling approaches involves using factor analysis and multivariate regression techniques to analyse the graphical structure that describes the relationships between the variables of interest [Walters and MacKenzie, 1988, Kim, Srinivasan, and Wilcox, 1999]. Random utility modelling assumes that consumer preferences between two or more options are discrete decisions that are made with respect to random utility functions that describes an individual's underlying objective in which they strive to maximise [Manski, 1977, Kim, Blattberg, and Rossi, 1995, Richards, Hamilton, Yonezawa, et al., 2015, Rossi, 2014]. All of these models have been successful at describing how the direct and cross-elasticity quantities impact demand at numerous levels of aggregation. However, for the purposes of our analysis, we focus on the functional form known as the Working-Leser equations. The Working-Leser regression models are parametric models that predict the demand of a product given covariate data. More precisely, for a category of N different products (also referred to as items), we estimate the set of coefficients

$\{c_i, \chi_{ij}, \psi_i \mid 1 \leq j \leq m_j,\ 1 \leq i \leq N\}$ derived from the system of N regression models:

$$y_{it} = \log(S_{it}) = c_i + \psi_i \log(P_{it}) - \sum_{j=1}^{m_i} \chi_{ij} \log(P_{ijt}) + \epsilon_{it}, \tag{4.3}$$

where:

$$\begin{array}{ll}
S_{it} = \text{sales of item } i \text{ at time } t, & \psi_i = \text{item } i\text{'s direct elasticity,} \\
P_{it} = \text{item } i\text{'s price at time } t, & \chi_{ij} = \text{item } j\text{'s cross elasticity with item } i, \\
P_{ijt} = \text{price of item } i\text{'s } j^{th} \text{ cross item at time } t, & c_i = \text{item } i\text{'s additive constant,} \\
m_i = \text{number of cross competitors of item } i, & \epsilon_{it} \sim N(0, \sigma_i^2.)
\end{array}$$

These and similar models are widely used in a range of econometrics and retail analytics settings [Andreyeva et al., 2010], but in practice companies often use much more sophisticated versions of this model, taking into account a larger range of covariates and that include autoregressive terms, time-dependencies via smoothing and seasonality modelling. Crucially, this model assumes that the log of sales of each product are conditionally independent, conditioned on the aforementioned covariates. The popularity of these models is that the output of fitted models is highly transparent as the coefficients are straight-forward to interpret and are computationally efficient to implement. This is of key importance, as companies can use models not only to predict demand but also gain a greater insight into the underlying phenomena at play.

Importantly, the cross-elasticity output of models such as (4.3) allows us to quantify the sales sensitivities of products with respect to the price changes of its relevant competition. In particular, the vector of a product's direct- and cross-elasticity coefficients, $(\psi_i, \chi_{i1}, \ldots, \chi_{im_i})$, are the quantities conveying a product's sale sensitivities with respect to changes in a product's own and competitor prices. Our aim is to cluster products as a function of their direct- and cross-elasticity coefficients vectors $(\psi_i, \chi_{i1}, \ldots, \chi_{im_i})$.

4.1.1 PriceStrat and cross-elasticity output

We now describe the dataset our sales sensitivities analysis and elasticity clustering is motivated by. Access to this dataset was permitted by dunnhumby ltd, and comprises the *relative cross-elasticity vectors* for a set of products from a large UK supermarket retailer. These relative cross-elasticity vectors have been generated from a cross-elasticity regression model known as *PriceStrat*, which is closely related to (4.3). Although the precise mechanics of how these estimates are obtained are highly engineered, the general form of the model is given by the following regression:

$$\log\left(S_{i,t}\right) = c_i - \varphi_i \log\left(Q_{i,t}\right) + \sum_{j=1}^{n_i} \varphi_i \eta_{ij} \log\left(P_{i,j,t}\right) + f\left(Q_{i,1:T}, P_{i,1:n_i,1:T}\right) + \epsilon_{i,t}, \tag{4.4}$$

where, for each product i and time t, S_{it} denotes its sales, Q_{it} its price, P_{ijt} the price of its j^{th} competitor product, φ_i its direct elasticity and η_{ij} product j's relative cross-elasticity with product i (as a multiple of the direct elasticity) and c_i is some additive constant. We use the notation $1:n$ to denote the set $1,\ldots,n$. The map $f(\cdot)$ involves nuanced data aggregation and smoothing, seasonality patterns relevant to retail sales, as well as additional information on display combinations and promotions specifically engineered to induce $\epsilon_{i,t} \sim N(0,\sigma_i^2)$. Here n_i is the number of competitor products of product i, which are pre-selected using expert knowledge. The regression coefficients are estimated using shrinkage methods, so that only l_i of the η_{ij}'s are non-zero, with the remaining exactly equal to 0. To ease notation and terminology in latter sections, we assume that competitor products are labelled such that product i's relative cross-elasticity coefficients η_{ij} are decreasing in magnitude with increasing j and that all products have the same potential number of competitors, i.e. $n_i := n \; i = 1, \ldots N$, and from now onwards refer to these relative cross-elasticity coefficients as the *cross-elasticity coefficients*. Table 4.1 provides some toy examples of the cross-elasticity vectors typically observed from dunnhumby's implementation of (4.4). These examples are provided to allow the reader to intuit and visualise the data in question. Although a clustering approach of the regression coefficients can be performed alongside the regression, this is often computationally prohibitive in any context where the original predictive sales model is highly tailored and engineered, such as with model (4.4). Consequently, any clustering methodology that clusters the cross-elasticity vectors separately from the regression analysis is often the preferred route to any sales sensitivity analysis.

The cross-elasticity data is such that, for each product i, we observe a decreasing set of entries of a larger vector, censored to only the top few entries with the remaining values set to 0. Thus, the data are in the form:

$$\mathbf{X} = \{\boldsymbol{\eta}_{i,1:n} : \eta_{i,n-l_i+1} \leq \eta_{i,n-l_i+2} \leq \ldots \leq \eta_{i,n}, \text{with } \eta_{i,j} \text{ censored to 0 for } 1 \leq j \leq n - l_i\} \tag{4.5}$$

where $\boldsymbol{\eta}_i$ is the cross-elasticity vector of dimension n for product i, which has l_i uncensored ordered entries, with the remaining being censored.

4.1.1.1 Challenges

We now describe some of the pertinent artefacts typical of the data (4.5) and the associated challenges.

1. **Preselection of significant cross competitors:** Although companies could store the entire data of (4.5) as a matrix containing all the cross-elasticities for each pair of products, this in practice would be computationally prohibitive. Consequently, companies induce sparsity through expert preselection of competitor products and shrinkage. They often do this by the use of highly tailored black-box sparse regression sales models [Liu, Ren, Choi, Hui, and Ng, 2013, Beheshti-Kashi, Karimi, Thoben, Lütjen, and Teucke, 2015] and only measure the cross-elasticities for a small number of competitors for each product, with the remaining entries being treated as missing or negligible. This resultant cross-elasticity coefficient data implicitly reflects only the top competitors within the market and thus induces an inherent informative missingness that means a global interpretation of the behaviour of the entire market may not be directly available. Thus, any proposed clustering strategy ought to reflect that the data of (4.5) is indeed pre-selected using expert knowledge to represent the top competitor products across a market (and therefore subject to error), with omitted entries being treated as zero or missing minor competitors and furthermore, should accommodate instances where relevant competitors have been omitted from the original regression that should have been included.

2. **Varying dimensions:** Due to the shrinkage previously, many of the remaining η_{ij}'s are exactly equal to 0. This effectively leads to the $\boldsymbol{\eta}_{i,1:n}$ having different dimensions for differing i as some products may have more or less zero shrinkage than others due the amount of competition with the market that they encounter. Consequently, any clustering methodology has to support this variation in dimensions exhibited in the cross-elasticities coefficients. Figure 4.2(a) shows the histograms of the varying dimensions (l_i) of cross-elasticity vectors $\boldsymbol{\eta}_{i,1:n}$.

Table 4.1: Ordered elasticity output φ and η for two fictional products, *Bobby's puffs* and *Lucan's Salted crisps.* For each product we have columns of order elasticity coefficients φ_i, $\varphi_i\eta_{ij}$ along with the respective sequences of η_{ij}, which demonstrates the decreasing nature of data from model (4.4). The number of potential cross competitors is set to $n_i = 6$, although the number of terms censored to 0 differs. Importantly, the set of competitors can differ for each of the products and in instances where there is a shared competitor (as with *Supermarket puffs* in this case), the value of $\varphi_i\eta_{ij}$, as well as its position in the ordering, need not be consistent across products.

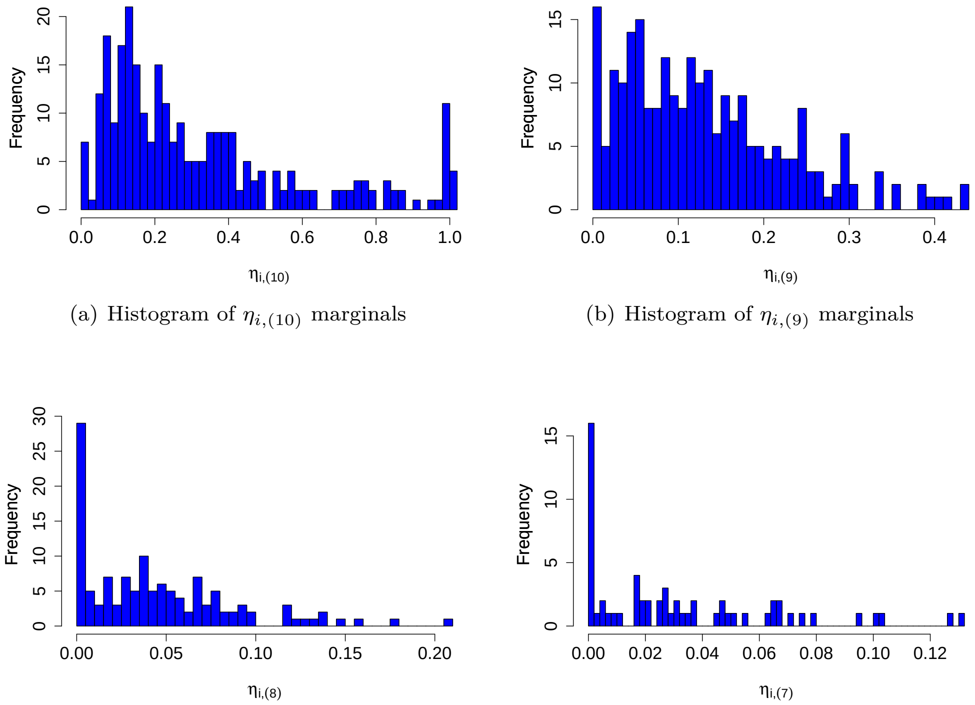
	Bobby's Cheesy puffs			**Lucan's Salted crisps**		
	Relevant competitors	$\varphi_1, \varphi_1\eta_{1j}$	η_{1j}	$\varphi_2, \varphi_2\eta_{2j}$	η_{2j}	Relevant competitors
φ_i	Bobby's puffs	-1.41		-1.86		Lucan's Salted crisps
$\varphi_i\eta_{i6}$	Supermarket puffs	-1.12	0.79	-0.8	0.43	Sussex's Chives crisps
$\varphi_i\eta_{i5}$	Harry's puffs	-1.10	0.78	-0.44	0.23	Chef's Paprika crisps
$\varphi_i\eta_{i4}$	Supermarket Nuts	-0.80	0.57	-0.10	0.05	Supermarket puffs
$\varphi_i\eta_{i3}$	Bobby's Tortillas	-0.48	0.34	-0.04	0.02	Lucan's nuts
$\varphi_i\eta_{i2}$	Tommy's chips	-0.35	0.25	0	0	Harry's Popcorn
$\varphi_i\eta_{i1}$	Tommy's puffs	-0.05	0.04	0	0	Chef's BBQ crisps

3. **Strictly decreasing sequences:** As a consequence of the η_{ij} reordering, the individual entries of cross-elasticity vectors $\boldsymbol{\eta}_{i,1:n}$ are strictly decreasing in magnitude, i.e. $\eta_{i,j} \leq \eta_{i,j+1}$. One of the key aspects in the analysis of cross-elasticity coefficients is the relative decay between successive values of the cross-elasticity coefficients, as this conveys the degree of competition a product encounters with the market. Figure 4.1 plots the histograms of the marginal elasticity entries (η_{ij}) of $\boldsymbol{x}_{i,1:n}$ which demonstrates the varying decay rates across the cross-elasticity vectors. Figure 4.2(b) plots some real data examples of decreasing cross-elasticity coefficients. We observe the entries of the cross-elasticity vectors $\boldsymbol{x}_{i,1:n}$ are decaying at differing rates and further notice $\boldsymbol{\eta}_{i,1:n}$ have differing number of entries (and thus differing dimensions).

Our goal is to summarise products' sensitivity in sales by clustering them by their cross-elasticities coefficient vectors. We however want to do this in way that handles the three aforementioned challenges. Namely, we want to cluster these cross-elasticity vectors according to the distribution of their competition in the market that accommodates not only clustering products with similar decay rates and cross-elasticity dimension, but also reflects that these entries are assumed to represent a product's most significant competitors (and further deals with possible omitted competitors).

4.2 Analysis of cross-elasticity coefficients

We now introduce the existing work done in the analysis of direct- and cross-elasticity coefficients generated from models closely related to (4.3). This section aims to give the reader an understanding of the current methodologies employed to investigate the differences in magnitudes between the direct- and cross-elasticity coefficients of differing products. We then explore



(a) Histogram of $\eta_{i,(10)}$ marginals

(b) Histogram of $\eta_{i,(9)}$ marginals

(c) Histogram of $\eta_{i,(8)}$ marginals

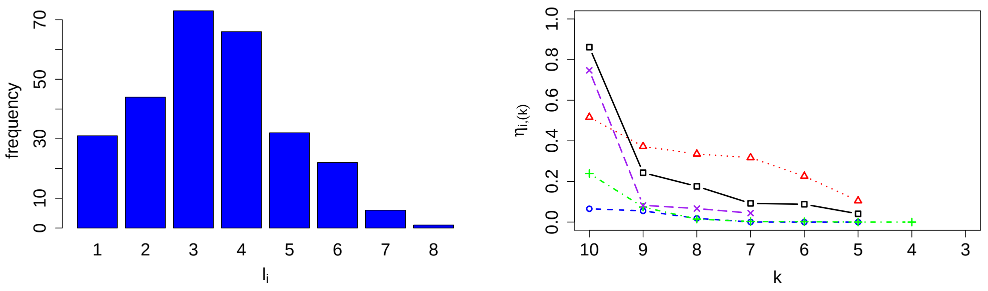
(d) Histogram of $\eta_{i,(7)}$ marginals

Figure 4.1: Various plots of η_i data

the narrative of what these differences convey. The section then moves on to highlight some of the shortcomings of the current analytical approaches.

4.2.1 Current approaches

We now outline the main contributions and describe the broad pattern employed among research into the analysis of direct and cross-elasticity coefficients generated from cross-elasticity regression models. Generally speaking, the body of research into direct- and cross-elasticity coefficients can be largely split into two fields. The first body of work is from the public health perspective which strives to understand how the public's consumption of specific food categories are sensitive to competing prices of other relevant categories. These papers are ultimately interested in finding insights aimed to aid public policy [Leeflang and Parreño-Selva, 2012, Guerrero-López et al., 2017, Andreyeva et al., 2010, Mhurchu et al., 2013]. In such studies, researchers broadly investigate how factors such as ethnicity and income are related to the sales sensitivities of food groups, with findings being used help reduce obesity and other diet related disorders. The second body of work into the analyses of direct- and cross-elasticity coefficients is from a retailers' perspective, in which they try to better understand and improve their business operations. These motivations are varied, but examples range from investigating the differing store-wise pricing policies and quantifying the effect of promotional strategies to



(a) Histogram of l_i

(b) Order statistics of few products

Figure 4.2: Various plots of η_i data

profit maximisation across product categories [Walters, 1991, Mulhern and Leone, 1991, Zellner, 1962]. Retailers ultimately aim to understand the substitutional relationship between products that these cross-elasticity coefficients convey, whether in terms of measuring brand loyalty or supermarket preference, and use this substitutional knowledge to help improve business practices.

Both bodies of work generally employ very similar post-processing methodologies when interpreting the direct- and cross-elasticity output exhibited from cross-elasticity demand models. As stated earlier, cross-elasticity regressions models for product i are generally of the form:

$$f\left(S_i\right) = c + \psi_i g\left(P_i\right) + \sum_{j=1}^{n} \psi_i \eta_{ij} g\left(P_{ij}\right) + h(x) + \epsilon_i$$

where c is some additive constant, x are some relevant covariate information and $f(\cdot)$, $g(\cdot)$ and $h(\cdot)$ are functional forms depending on the context and ϵ_i is additive noise. Very often the functional forms of $f(\cdot)$ and $g(\cdot)$ are such that $f(x) = x$ or $f(x) = \log(x)$ and similarly for $g(\cdot)$. Variations of this generalised model are frequently used to generate direct- and cross-elasticity coefficients, examples of which include Almost-Ideal-Demand models [Deaton and Muellbauer, 1980] or for the Working-Leser demand system [Working, 1943]. However, independent of the original regression details, the direct- and cross-elasticity output is essentially the same as it ultimately gives rise to data of the form (4.5), i.e.:

$$\mathbf{X} = \{\boldsymbol{\eta}_{i,1:n} : \eta_{i,n-l_i+1} \leq \eta_{i,n-l_i+2} \leq \ldots \leq \eta_{i,n}, \text{with } \eta_{i,j} \text{ censored to } 0 \text{ for } 1 \leq j \leq n - l_i\}$$

where $\boldsymbol{\eta}_i$ is the cross-elasticity vector of dimension n for product i, which has l_i uncensored ordered entries, with the remaining being censored. The regression output from various implements of cross-elasticity models is of this form as long as there is a single coefficient that summarises the relationship between the change of demand with respect to the change of a product's price, which is the case in our cross-elasticity analysis. There are then systematic trends in the post-processing analysis of data (4.5), which are as follows.

Firstly, the majority of work focuses on direct-elasticities, and in instances where cross-elasticity effects are considered, the possible competitors are usually constrained to a preselected group of products (either through expert pre-selection or a set of products determined by some variable selection technique) that are chosen to assess the substitutability of some product category of interest [Oliveira, Foxall, and Schrezenmaier, 2007, Guerrero-López, Unar-Munguía, and Colchero, 2017, Andreyeva, Long, and Brownell, 2010, Walters, 1991]. In such studies, the relative substitutability of products and general sensitivity to sales are assessed by raw comparisons of the direct- and cross-elasticities across the products. However, the relative sizes between the direct- and cross-elasticity coefficients are not assessed, rather the cross elasticity effects are simply compared to one another when they exist.

Secondly, the decision about how many competitors are selected in the cross-elasticity regression is generally considered secondary in such analyses. The relative importance and relevance of the inclusion (or omission) of cross competitors is determined from consideration of statistical significance, by either looking at p-values or some other equivalent variable selection method. The potential values of the omitted competitors is then ignored in the corresponding direct- and cross-elasticity analysis.

Often, much research is interested in understanding how an entire category or set of brands are affected by the price deviations of their competition or other relevant group of products [Leeflang and Parreño-Selva, 2012]. In such analyses, the method of segmenting and categorising the direct- and cross-elasticity coefficients is done a priori before the analysis begins. In particular, the models are implemented across the relevant products and the direct- and cross-elasticity coefficients are aggregated across the chosen categories. The corresponding category-wise analysis of this output is generally done by studying summary statistics across these predefined segments. More concretely, given category-wise coefficient data $\mathbf{X} = \{(\psi_i^{(k)}, \eta_{i1}^{(k)}, \dots, \eta_{im_i}^{(k)}) \mid \text{ where } (\psi_i^{(k)}, \eta_{i1}^{(k)}, \dots, \eta_{im_i}^{(k)}) \text{ are coefficients from category } k\}$, summaries for the direct- and cross-elasticities are often taken to be $\frac{1}{N_k}\sum_{i=1}^{N_k}\psi_i^{(k)}$ and $\frac{1}{N_k}\sum_{i=1}^{N_k}\eta_{ij}^{(k)}$ for some cross effect product j and where N_k is the number products in the category k. These summaries are often calculated as the mechanism of conveying the aggregate price elasticity across some predefined set of products. These category-wise summary statistics are then compared relative to one another and heterogeneity, where it exists, is discussed along with the relevant implications within the given modelling context [Hoch, Kim, Montgomery, and Rossi, 1995, Gordon, Goldfarb, and Li, 2013]. Existing approaches generally demonstrate that differences between these elasticity summary statistics across different categories of $\mathbf{X}_k$ do exist, which supports the hypothesis that products exhibit fundamentally different sales sensitivities.

4.2.2 Shortcomings of existing analyses

As the previous subsection illustrates, there has been much research in investigating the differences in sales sensitivities between differing products and the implications such heterogeneity has at strategic pricing and policy decision making levels. However, there are arguably some systematic weaknesses of how these direct- and cross-elasticities are analysed. Here are some of the pertinent shortcomings that we see in the current approaches of analysing these direct- and cross-elasticity coefficients that could be improved upon:

1. **Direct elasticities**: Much of the research in the price sensitivity analysis of products looks only as far as a product's direct elasticities, i.e. a product's sales sensitivity is summarised in terms of its own price changes. Although the primary driver of a product's sales is its own price, there may be additional information in how a product's sales changes with respect to changes in its relevant competitor's price changes. Typically, these effects are frequently overlooked and it is our hypothesis that product's may not only be characterised by their direct-elasticities, but also by the nature of their cross-elasticities.

2. **Relative cross elasticity magnitudes**: In the cases when cross-elasticities are considered, information about the magnitudes of leading cross-elasticity coefficients relative to the direct elasticity is not considered. Information on this decay between the direct-elasticity and leading cross-elasticity coefficients could be another way of segmenting and interpreting the differences between these sales sensitivities.

3. **Top competitor assumption**: In the majority of direct- and cross-elasticity coefficient analysis, the decision of which products qualify as significant in implementations of models such as (4.3) is often fixed a priori. Thus, any interpretation of cross-elasticity coefficients with respect to one another ought to reflect that the number and particular competitors are selected to represent the most significant competitors a product has across the entire market. This is crucially important, as the relative decay of the cross-elasticity coefficients may convey information about the values of possibly omitted regression coefficients had they been included in the original regression. This is an additional aspect that we be-

lieve products could differentiate themselves from one another with respect to their sales sensitivities.

4. **Summary statistics**: The current approaches only consider summary statistics of the elasticity coefficients. Although high-level summary statistics capture headline information on the cross-elasticity coefficients, characterising the entire distribution of these elasticity coefficients would allow experimenters to compare the similarity or dissimilarity between elasticity coefficients in a mixture modelling framework. Currently however, elasticity coefficients are simply eye-balled and little thought is given to a possible generating process describing the structure of the observed elasticity coefficients. The lack of such a distributional characterisation of elasticity coefficients makes it difficult to cluster and segment the sales sensitivities of products in systematic way.

Chapter 5

Elasticity clustering using Dirichlet process mixtures

This Chapter is largely based on a paper due to be published in JRSSC titled "Dirichlet Process Mixtures of Order Statistics with Applications to Retail Analytics". arXiv:1805.05671

This Chapter presents a novel methodology that clusters products in terms of their cross-elasticity coefficients, and thus allows us to segment the universe of supermarket products in terms of their relative sales sensitivities. We achieve this by developing a Bayesian non-parametric modelling framework and interpreting our observed data of (4.5) as realisations of *variable length order statistics sequences.* Crucially, by reframing the data of (4.5) as *variable length order statistics sequences*, it allows us to specify a distributional form that characterises the cross-elasticity coefficient data. This in turn, allows us to define a mixing kernel that quantifies the degree of similarity between different cross-elasticity coefficients and therefore accommodates a mixture modelling setting. We will show this succinctly handles the partial censoring and allows for computationally straight-forward inference on the unobserved entries of the cross-elasticity matrix. Our approach uses tools from survival analysis to address inherent censoring mechanisms, together with a Dirichlet Process mixture model that allows products to be clustered into distinct groups. By using the Exponentiated Weibull distribution as a mixture kernel [Mudholkar and Srivastava, 1993], we are able to account for both light and heavy tail behaviour apparent in the data. As we will discuss later, the Exponentiated Weibull distribution has several unique properties which makes it ideal for modelling order statistics. We develop efficient sampling mechanisms by using Neal [2000]'s algorithm 8 and provide interpretations and visualisations of the fitted output. Our approach fully characterises the entire cross-elasticity vector, offering two distinct benefits. Firstly, by interpreting these elasticity vectors as order statistic sequences, we can directly cluster products by all of their cross-elasticity coefficients and conveniently handle their varying dimensions. Secondly, it provides a framework for predicting censored entries which can shed light on potentially important competitors which have been omitted from the original regression. Hence, our approach neatly handles the challenges outlined in 4.1.1.1. We implement our proposed methodology on three datasets, two simulated examples and one from real cross-elasticity data generated from a large UK supermarket retailer's regression model output which we were given access to through dunnhumby's secured servers. We show that our proposal successfully partitions the space of products in terms of their sales sensitivities.

The rest of the Chapter is organised as follows: Section 5.1 introduces the concept of *variable length order statistic sequences* as a reinterpretation of data (4.5), and provides a background of the pertinent characteristics of the Exponentiated Weibull distribution and its relevance as a kernel to *variable length order statistic sequences.* Section 5.2 introduces our proposal of a Dirichlet process mixture model of *variable length order statistic sequences* along with specifications of the prior distributions used during this analysis. Section 5.3 outlines the algorithm used for posterior inference of our proposed mixture model. Section 5.4 illustrates the results of our methods on three datasets: two simulated examples and one real dataset. Section 5.5 finally summaries our contribution and further discusses some potential extensions our model and applications of our approach to other fields.

5.1 Relevant distributions

As discussed in the previous Chapter, the cross-elasticity data at hand is such that, for each product i, we observe a decreasing set of entries (i.e. observed order statistics) of a larger vector, that have been censored for sparsity purposes to only the top few entries. Mathematically speaking, the data are in the form:

$$\mathbf{X} = \{\boldsymbol{x}_{i,1:n} : x_{i,n-l_i+1} \leq x_{i,n-l_i+2} \leq \ldots \leq x_{i,n}, \text{with } x_{i,j} \text{ censored to 0 for } 1 \leq j \leq n - l_i\},$$

where $\boldsymbol{x}_{\mathrm{i}}$ is the cross-elasticity vector of dimension n for product i, which has l_i uncensored ordered entries, with the remaining being censored. We recast these decreasing sequences of varying length as *variable length order statistics sequences*, which we will go onto define along with a distribution known as Exponentiated Weibull which we propose as a suitable mixing kernel for such sequences.

5.1.1 Order statistics of continuous distributions

The order statistics of a random sample are the *reordered* observations in terms of increasing size. More concretely, given a continuous distrbution variable X and observations $x_{1:n} \overset{i.i.d.}{\sim} X$, the order statistics $x_{(1)}, \dots, x_{(n)}$ are given by:

$$x_{(1)} < x_{(2)} < \dots < x_{(n)}. \tag{5.1}$$

The j^{th} order statistic of (5.1) is denoted as $x_{(j)}$ and thus, $x_{(1)}$ and $x_{(n)}$ are the smallest and largest observations respectively. Given a density function $f(x)$ of a continuous random variable X, the density of the j^{th} order statistic $x_{(j)}$, denoted by $f_{(j)}(x)$ is given by [Arnold et al., 1992]:

$$f_{(j)}(x) = nf(x)\binom{n-1}{j-1}F(x)^{j-1}(1-F(x))^{n-j}. \tag{5.2}$$

An implicit assumption of the cross-elasticity coefficient data is that the elasticities represent the top competitors a product encounters throughout the entire market, i.e. these cross-elasticity coefficients are the largest in magnitude a product will encounter across all of its competitors. Hence, we then make the following assumption that the partially observed cross-elasticity vector $\boldsymbol{x}_{i,1:n}$, of length n with l_i non-zero entries in fact corresponds to the top l_i order statistics of a random sample of size n. We term each of these vectors of the top l_i order statistics as *variable length order statistics sequences*, and denote them as $\boldsymbol{x}_{i,1:n} = \left(x_{i,(n)}, \dots, x_{i,(n-(l_i-1))}\right)$. We also denote the j^{th} order statistic of sequence $\boldsymbol{x}_{i,1:n}$ by $x_{i,(j)}$. For notational ease, we drop the i index for the remaining of this section. The density of $\boldsymbol{x} \mid l$ denoted as $f_{(n):(n-l+1)}$ is given by:

$$\begin{aligned} f_{(n):(n-l+1)}(\boldsymbol{x} \mid l) =& f_{(n):(n-l+1)}\left(x_{(n)}, \dots, x_{(n-l+1)} \mid l\right) \\ =& \frac{n!}{(n-l)!}F\left(x_{(n-(l-1))}\right)^{n-l}\prod_{j=1}^{l} f\left(x_{(n+j-l)}\right). \end{aligned} \tag{5.3}$$

By the independence of $x_{(n-j)} \mid x_{(n-j+1)} \perp\!\!\!\perp x_{(n)}, x_{(n-1)}, \dots, x_{(n-j+2)}$ and by (5.3), the density of the conditional distribution of $x_{(n-j)} \mid x_{(n-j+1)}, l$ for $j < l$ (denoted as $f_{(n-j)|(n-j+1)}$) is given by:

$$f_{(n-j)|(n-j+1)}\left(x_{(n-j)} \mid x_{(n-j+1)}, l\right) = (n-j)\, f\left(x_{(n-j)}\right)\frac{F\left(x_{(n-j)}\right)^{n-(j+1)}}{F\left(x_{(n-j+1)}\right)^{n-j}} \tag{5.4}$$

and thus the density of the joint sample $\boldsymbol{x} \mid l$ can also be expressed in hierarchical format:

$$f_{(n):(n-l+1)}(\boldsymbol{x} \mid l) = f\left(x_{(n)}\right)\prod_{j=1}^{l-1} f_{(n-j)|(n-j+1)}\left(x_{(n-j)} \mid x_{(n-j+1)}, l\right) \tag{5.5}$$

Finally, the joint density of a *variable length order statistics sequence* $\boldsymbol{x}$, denoted as f_{vloss}, can therefore be expressed as:

$$\begin{aligned} f_{vloss}\left(\boldsymbol{x}, l\right) =& f_{vloss}\left(l\right) \times f_{vloss}\left(\boldsymbol{x} \mid l\right) \\ =& p(l) \times f_{(n):(n-l+1)}\left(\boldsymbol{x} \mid l\right) \end{aligned} \tag{5.6}$$

since $f_{vloss}\left(\boldsymbol{x} \mid l\right)$ (the density of an observed vector of order statistics $\boldsymbol{x} \mid l$) is precisely $f_{(n):(n-l+1)}\left(\boldsymbol{x} \mid l\right)$, and where $f_{vloss}\left(l\right)$ is simply the probability mass function over the length of the sequence, which we denote as $p(l)$. Here we assume that l and the non-zero entries of $\boldsymbol{x}$ are independent.

Much work has been done in the study of the theoretical properties of order statistics [Beutner and Kamps, 2009], from which they have been applied to areas such as modelling the reliability of software and to the modelling of recommender systems [Wilson and Samaniego, 2007, Caron and Teh, 2012]. A relevant field of order statistics which bears resemblance to our problem set-up lies in the field of reliability analysis, known as k-out-of-n systems. A k-out-of-n system models the failure of k out of n components within a finite time horizon. The set of k ordered values of the time until failure (censored or not) can then be modelled as the observed order statistics of a base distribution. Much of the relevant non-parametric work has focused on flexibly learning the underlying base distributions [Wilson and Samaniego, 2007, Barghout et al., 1998] and building hierarchical versions of these models [Ghosh and Tiwari, 2007].

In the current context, we observe the top few order statistics of the cross-elasticity vector, with the remaining entries treated as missing. This type of data is akin to the format of models in survival analysis, where the probability of survival decreases over time and may be right-censored. One aspect important to the success of Bayesian non-parametric models in survival analysis is the choice of kernel, as it impacts whether the relevant statistics and survival functions are recoverable. As a consequence, much attention is paid to the choice of kernel. Notably, a hierarchical structure in the base measure was introduced by De Iorio et al. [2004], whereas Hanson et al. [2006] and Kottas [2006] used Gamma and Weibull kernels within a Dirichlet process mixture model framework respectively. The Exponentiated Weibull distribution was shown to be a distribution that could model non-monotone hazards [Mudholkar and Srivastava, 1993], which in our context correspond to order statistics terms whose modes exist but are not necessarily light-tailed.

5.1.2 Exponentiated Weibull distribution

Following the formulation of our observations as order statistics of random samples, the choice of the underlying distribution of X will determine the behaviour of the corresponding order statistics. Here we are interested in a distribution which can allow for a range of light and heavy tail behaviour and provide interpretable analytical expressions for the distribution of its order statistics. We thus assume that these random samples are distributed according to the Exponentiated Weibull distribution. A random variable X is distributed according to the Exponentiated Weibull (EW) distribution, denoted as $X \sim EW\left(\alpha,\beta,\lambda\right)$, if its probability density and distribution function are given by

$$f\left(x\right) = \alpha\beta\lambda^{\beta}x^{\beta-1}\left(1-e^{-(\lambda x)^{\beta}}\right)^{\alpha-1}e^{-(\lambda x)^{\beta}} \tag{5.7}$$

and

$$F\left(x\right) = \left(1-e^{-(\lambda x)^{\beta}}\right)^{\alpha} \tag{5.8}$$

respectively, where $x > 0, \lambda > 0, \beta > 0, \alpha > 0$. The Exponentiated Weibull is an extension to the standard Weibull distribution through the inclusion of the additional parameter α, which allows the distribution to have a wide range of tail behaviours. Similarly to the Weibull distribution, λ is a scale parameter whereas β controls the tail behaviour of the distribution; distributions are heavy tailed for $\beta < 1$ and light-tailed otherwise. Furthermore, decreasing β monotonically increases the mean and variance, kurtosis and skew of the EW distribution. The impact of α depends on both the value $\alpha\beta$ and whether $\alpha < 1$; increasing α increases symmetry around the mean and mode. These different modal, asymptotic and tail behaviours [Nassar and Eissa, 2003] are summarised in Table 5.1. Figure 5.1 demonstrates various density plots for differing combinations of (α, β, λ), various asymptotic, modal and tail behaviours are observed.

Table 5.1: EW density behaviours for various combinations of (α, β, λ)

Ranges of α, β	$x \to 0$	Mode	Order statistic marginal tails
$\alpha > 1, \beta > 1, \alpha\beta > 1$	$f\left(x\right) \to 0$	$\approx \frac{1}{\lambda}\left[\frac{2(\alpha\beta-1)}{\beta(\alpha+1)}\right]^{1/\beta}$	Light
$\alpha > 1, \beta < 1, \alpha\beta > 1$	$f\left(x\right) \to 0$	$\approx \frac{1}{\lambda}\left[\frac{2(\alpha\beta-1)}{\beta(\alpha+1)}\right]^{1/\beta}$	Heavy
$\alpha > 1, \beta < 1, \alpha\beta < 1$	$f\left(x\right) \to \infty$	none	Heavy
$\alpha < 1, \beta > 1, \alpha\beta < 1$	$f\left(x\right) \to \infty$	none	Light
$\alpha < 1, \beta > 1, \alpha\beta = 1$	$f\left(x\right) \to \lambda$	0	Light

5.1.3 EW distribution application to order statistics

There are some key properties of the EW distribution that lead to useful applications to order statistics and *variable length order statistics sequences.* The joint density of (5.3) under the EW

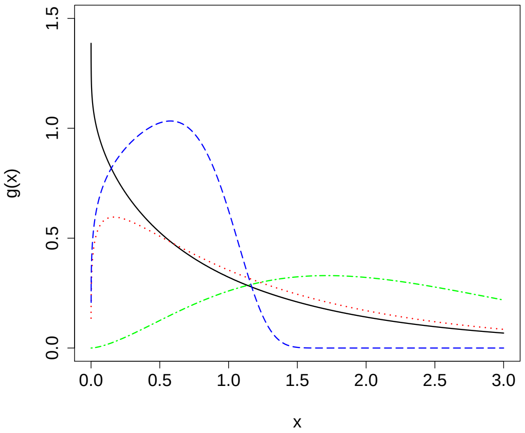


Figure 5.1: EW density for $(\alpha, \beta, \lambda) = (1.2, 0.8, 1.0)$ [black solid], $(1.55, 0.8, 1.0)$ [blue dashed], $(0.24, 5.0, 1.0)$ [red dotted] and $(1.8, 1.4, 0.5)$ [green dashed-dotted lines] respectively.

distribution for fixed order sequences of lengths l is given by:

$$f_{(n):(n-l+1)}\left(\boldsymbol{x} \mid l\right) = \frac{n!}{(n-l)!}\left(1 - e^{-\left(\lambda x_{(n-(l-1))}\right)^{\beta}}\right)^{\alpha(n-l)} \prod_{j=1}^{l} f\left(x_{(n+j-l)}\right) \tag{5.9}$$

where f is the EW density function of (5.7). The EW distribution handles censoring naturally, since the censored, joint and conditional densities under the EW distribution belong to the same family, i.e. $x_{(n-j)} \mid x_{(n-j+1)} \sim EW_{x_{(n-j)}<x_{(n-j+1)}}\left((n-j)\,\alpha, \beta, \lambda\right), 1 \leq j \leq n-1$ are also readily available. This means that the properties and interpretability of the EW distribution transparently carry over to its order statistics. Finally, the EW can account for both light and heavy tails, allowing us to capture different types of decay behaviours of the elasticity vectors. Figure 5.2 provides some examples of order statistics sequences, which demonstrate various decay behaviours and tail behaviours that can be produced under the EW kernel.

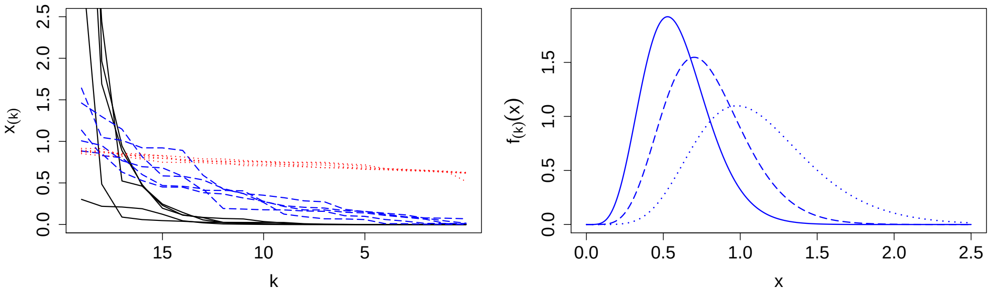


Figure 5.2: Left panel: realisations of order statistics sequences with $EW\left(\alpha, \beta, \lambda\right)$ kernel for combinations $(\alpha, \beta, \lambda) = (0.2, 0.6, 0.7)$ [black solid], $(0.5, 1.5, 1.5)$ [blue dashed], $(4, 5, 1.5)$ [red dotted] respectively. Right panel: Density plots of $f_{(k)}\left(x\right)$ with EW(0.5, 1.5, 1.5) kernel for orders k=10 [dotted], 9 [dashed] and 8 [solid].

5.2 Model

Our ultimate goal is to characterise the behaviour of different products in terms of their cross-elasticity coefficients. To this end, we use the EW distribution as a representation of cross-elasticity decay behaviour. However, in order to account for different behaviour across products, we additionally cluster products that potentially correspond to the same EW distribution. We thus model the entire set of cross-elasticity vectors non-parametrically as a Dirichlet Process Mixture Model [Antoniak, 1974] as outlined in the section 3.5.1.

5.2.1 Nonparametric mixture model of variable length order statistic sequences

We now propose a DPMM of *variable length order statistics sequences* on mixtures of distributions satisfying (5.9). Placing a $\mathrm{DP}(\nu G_0)$ on the distributions of (5.9) is an attractive approach to handling the complex multi-modalities, decay rates and variable lengths that order statistics sequences can exhibit as discussed in Section 5.1.2. Thus, the DPMM of *variable length order statistics sequences* expressed in hierarchical format of (5.5) by:

$$\begin{aligned}
\nu &\sim Gamma\left(\tau_1, \tau_2\right), \\
G \mid \nu &\sim \mathrm{DP}\left(\nu G_0\right), \\
\left(\alpha_i, \beta_i, \lambda_i, w_i\right) \mid G &\sim G, \\
l_i &\sim 1 + Binomial\left(n-1, w_i\right), \\
x_{i,j} &\sim EW\left(\alpha_i, \beta_i, \lambda_i\right),\ j = 1, \ldots, n,
\end{aligned} \tag{5.10}$$

where $i = 1, 2, \ldots, N$ are the number of observations and for each observation vector i, with all but the top l_i entries being censored. The final line of (5.10) can also be expressed through the iterative formulation:

$$\begin{aligned}
x_{i,(n-j)} \mid x_{i,(n-j+1)} &\sim EW_{x_{i,(n-j)} < x_{i,(n-j+1)}}\left(\left(n-j\right)\alpha_i, \beta_i, \lambda_i\right), 1 \leq j \leq l_i - 1 \\
x_{i,(n)} &\sim EW\left(n\alpha_i, \beta_i, \lambda_i\right).
\end{aligned} \tag{5.11}$$

which follows from equation (5.5). We treat the lengths l and observations $x_{i,(j)}$ of $\boldsymbol{x}$ as independent to allow detection of competitor omissions and to ease computation. Since cross-elasticity coefficients are identically distributed a priori, each individual coefficient has the same probability of being censored, leading to a Binomial prior on l_i; to avoid the degenerate case of empty cross-elasticity vectors, we force one of the Bernoulli trials to be 1. It important to note, that w plays an important role in (5.10) that is particularly relevant to our modelling setup. The inclusion of w as a cluster level parameter, rather than being a global parameter, is especially important for the following two reasons. Firstly, a characterising feature of cross-elasticity coefficient vectors are their lengths, not just the relative decay between entries (which is captured by α, β, λ parameters). These lengths of the cross-elasticity coefficient vectors have an important retail analytics interpretation, in that they convey the number of significant competitors a product has throughout the market - which is typically seen as one of the defining features of a product (and group of products). Secondly, w conveys the level of truncation, and thus implicitly the number of censored observations. For clusters with few censored coefficients, we expect larger values of cluster-wise w values. Consequently, the cluster-wise variable w has important ramifications in section of 5.4.4.1 (and subsequent sections), where we define two statistics relevant to our retail analytics analysis; namely the *omitted competitors* (OC) and *aggregate competition* (AC) statistics. These statistics strive to respectively describe the relative magnitude of possible omitted cross-elasticity coefficients and the total effect of competition a product receives throughout the market (detailed definitions of these are provided in section 5.4.4.1). Both definitions crucially rely on an accurate summarisation of the number of non-censored coefficients a cluster typically observes. Consequently, a cluster-wise variable of w is key for this accurate summarisation, with possible misfit being likely if w were a global parameter.

The base distribution G_0 is a key aspect of the $\mathrm{DP}(\nu G_0)$ as it specifies the prior over $(\alpha, \beta, \lambda, \omega)$ atoms which defines the cluster structure of the model; here we specify G_0 as:

$$\begin{aligned} G_0\left(\alpha, \beta, \lambda, w\right) &= \mathit{Gamma}\left(\alpha \mid \alpha^1, \alpha^2\right) \times \mathit{Gamma}\left(\beta \mid \beta^1, \beta^2\right) \times \\ &\qquad \times \mathit{Gamma}\left(\lambda \mid \lambda^1, \lambda^2\right) \times \mathit{Beta}\left(w \mid a, b\right). \quad (5.12) \end{aligned}$$

The hyperparameters $\left(a, b, \alpha^1, \alpha^2, \beta^1, \beta^2, \lambda^1, \lambda^2\right)$ are treated as fixed, chosen depending on the modelling context and reflecting prior expertise. The prior for ν is assumed to be $\mathit{Gamma}\left(\tau_1, \tau_2\right)$, allowing the relation $\mathbb{E}\left[N^* \mid \nu\right] = \nu \log\left(\frac{\nu+N}{\nu}\right)$ [Escobar and West, 1995] (where N^* is the number of occupied clusters) to inform our prior expectation of the number of clusters. As discussed during section 3.5.2, and illustrated by the simulation study of figure 3.1, the ν parameter of $\mathrm{DP}(\nu G_0)$ acts as smoothing parameter controlling the degree of 'smoothness' of density estimates, and equivalently, the number of unique mixture components induced by the $\mathrm{DP}(\nu G_0)$ prior. Consequently, it is important to place uncertainty over the ν parameter, especially in the context of *variable length order statistic sequences* where it is not clear how to fix ν, as it not obvious how many unique clusters will exist in the retail analytics dataset. Although it should be noted, in some modelling contexts, the number of unique clusters is often driven by the data.

5.3 Posterior inference

We now present an efficient Markov Chain Monte Carlo (MCMC) procedure for obtaining samples from the posterior of $p\left(\alpha, \beta, \lambda, w, \nu \mid \mathbf{X}\right)$ according to the model proposed by (5.10) with:

$$\mathbf{X} = \{\boldsymbol{x}_{i,1:n} : x_{i,n-l_i+1} \leq x_{i,n-l_i+2} \leq \ldots \leq x_{i,n}, \text{with } x_{i,j} \text{ censored to 0 for } 1 \leq j \leq n - l_i\},$$

where $\boldsymbol{x_i}$ includes the *variable length order statistics sequence* of length l_i (uncensored ordered entries), with the remaining $(n - l_i)$ being censored. Our posterior inference methodology as outlined in section 3.6, consists of three steps to obtaining samples from $p\left(\alpha, \beta, \lambda, w, \nu \mid \mathbf{X}\right)$ for each MCMC iteration: sampling the atoms $(\alpha, \beta, \lambda, w)$ of the $DP\left(\nu G_0\right)$ for each order statistics sequence; sampling the cluster-wise atoms for each of the unique clusters (as induced by $DP\left(\nu G_0\right)$), and finally, sampling the ν scale parameter. As discussed during section 3.6, this DPMM inference methodology is simple, intuitive and circumvents issues relating to the infinite dimensional nature of the DP by marginalising over the random measure G, and then updating the cluster allocations of the data-points. These three steps are manifest as follows:

5.3.1 Sample from $p\left(\theta_i \mid \theta_{-i}, \nu, x_i\right)$

We initiate by using the Polya urn exposition of a DP [Blackwell and MacQueen, 1973] by taking a Gibbs sample of $\boldsymbol{\theta}_i = (\alpha_i, \beta_i, \lambda_i, w_i)$ atoms associated to observation $\boldsymbol{x}_i$ using:

$$p\left(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}, \nu, \mathbf{X}\right) = q_0^* H_i + \sum_{k=1}^{N^*} q_k^* \delta_{\boldsymbol{\theta}_k^*} \tag{5.13}$$

where $q_0^* \propto \nu \int f\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}\right) G_0\left(d\boldsymbol{\theta}\right)$ and $q_k^* \propto N_k^* f\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k^*, \nu\right)$ subject to $\sum_{k=0}^{N^*} q_k^* = 1$. Here $f\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}\right) = f_{(n):(n-l_i+1)}\left(\boldsymbol{x}_i \mid l_i, \alpha, \beta, \lambda\right) p\left(l_i \mid w\right)$, where $f_{(n):(n-l_i+1)}$ is specified in (5.9) and the conditional distribution $p\left(l_i \mid w\right) = \binom{n-1}{l_i-1} w^{(l_i-1)} \left(1-w\right)^{(n-l_i)}$. H_i is the posterior distribution for $\boldsymbol{\theta}$ based on the prior distribution G_0 of (5.12) with likelihood $f\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}, \nu\right)$. Here $\boldsymbol{\theta}_{-i}$ denotes the vectorised atoms of $\boldsymbol{\theta}$ excluding the i^{th} atom $\boldsymbol{\theta}_i$, $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{N^*}^*\}$ denotes the unique values of $\boldsymbol{\theta}_i$, N^* the number of unique clusters induced by the DP and N_k^* the number of points assigned to atom $\boldsymbol{\theta}_k^*$.

As calculating the integral q_0^* is intractable, we use algorithm 8 [Neal, 2000] to approximate q_0^* by a weighted mixture of likelihoods by taking c auxiliary components sampled from the prior distribution G_0. Concretely, we sample $\boldsymbol{\theta}_i = (\alpha_i, \beta_i, \lambda_i, w_i)$ by sampling from the multinomial distribution of degrees of freedom of order $N^* + c$ with entries

$$\boldsymbol{\theta}_k^* \overset{iid}{\sim} G_0 \text{ for } k = N^* + 1, \ldots, N^* + c$$

$$G_0 = Beta\left(w \mid a, b\right) \times Gamma\left(\alpha \mid \alpha^1, \alpha^2\right) \times Gamma\left(\beta \mid \beta^1, \beta^2\right) \times Gamma\left(\lambda \mid \lambda^1, \lambda^2\right)$$

with probabilities

$$P\left(\boldsymbol{\theta}_i = \boldsymbol{\theta}_k^* \mid \boldsymbol{\theta}_{-i}, \boldsymbol{x}_i, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{N^*+c}^*\right) \propto \begin{cases} \frac{N_k^*}{N-1+\nu} f\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k^*\right) \text{ for } 1 \leq k \leq N^* \\ \frac{\nu/k}{N-1+\nu} f\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k^*\right) \; N^* < k \leq N^* + c \end{cases}$$

where

$$\begin{aligned} f\left(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k^*\right) &= \binom{n-1}{l_i - 1} \left(w_k^*\right)^{(l_i - 1)} \left(1 - w_k^*\right)^{(n - l_i)} \times \\ &\times F\left(x_{i,(n-(l_i-1))} \mid \alpha_k^*, \beta_k^*, \lambda_k^*\right)^{n-l_i} \prod_{j=1}^{l_i} f\left(x_{i,(n+j-l_i)} \mid \alpha_k^*, \beta_k^*, \lambda_k^*\right) \quad (5.14) \end{aligned}$$

where $F\left(x \mid \alpha, \beta, \lambda\right) = \left(1 - e^{-(\lambda x)^{\beta}}\right)^{\alpha}$ and $f\left(x \mid \alpha, \beta, \lambda\right) = \alpha\beta\lambda^{\beta}x^{\beta-1}\left(1 - e^{-(\lambda x)^{\beta_i}}\right)^{\alpha-1} e^{-(\lambda x)^{\beta}}$. The number auxiliary components c chosen determines the level q_0^* is approximated to.

5.3.2 Sample from $p\left(\alpha^*, \beta^*, \lambda^*, w^* \mid \nu, x_{\{i:\, C_i = k\}}\right)$

The $\boldsymbol{\theta}_k$ atoms are then updated for each of the unique clusters $k = 1, \ldots, N^*$ to avoid inefficiencies associated with having to pass through extremely low probability states to get to a higher probability states. This is achieved by updating $\boldsymbol{\theta}_k$ to be a single sample generated from the posterior $p\left(\boldsymbol{\theta}_k \mid \nu, \boldsymbol{x}_{\{i:\, C_i = k\}}\right)$, for each $k = 1, \ldots, N^*$. As taking exact samples from $p\left(\boldsymbol{\theta}_k \mid \nu, \boldsymbol{x}_{\{i:\, C_i = k\}}\right)$ is intractable for our choice of kernel (5.9) and prior G_0 (5.12), the Metropolis Hastings algorithm is used to sample from $p\left(\boldsymbol{\theta}_k \mid \nu, \boldsymbol{x}_{\{i:\, C_i = k\}}\right)$, for $k = 1, \ldots, N^*$. This involves following t_{MH} iterations of the Metropolis Hasting procedure, and saving the final t_{MH}^{th} sample as $\boldsymbol{\theta}_k$, for each $k = 1, \ldots, N^*$. The primary motivation of these Metropolis Hastings updates is to perturb the unique locations of $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{N^*}^*\}$, for $k = 1, \ldots, N^*$, and thus avoiding the Polya Urn sweeps getting stuck, rather than obtaining a complete representation of the entire posterior distribution specified by $p\left(\boldsymbol{\theta}_k \mid \nu, \boldsymbol{x}_{\{i:\, C_i = k\}}\right)$.

The precise Metropolis Hasting procedure is as follows: to ease notation, we suppress the asterisks from the exponents in this subsection. For $t_{MH} = 1, \ldots, T_{MH}$ iterations, we draw new parameters using a Normal proposal for $\left(\alpha_k', \beta_k', \lambda_k', w_k'\right)$ centred at the current points

$\left(\alpha_k^{t_{MH}}, \beta_k^{t_{MH}}, \lambda_k^{t_{MH}}, w_k^{t_{MH}}\right)$ with standard deviations $\sigma_\alpha, \sigma_\beta, \sigma_\lambda, \sigma_w$ respectively:

$$(\alpha_k', \beta_k', \lambda_k', w_k') \sim N\left(\alpha_k^{t_{MH}}, \sigma_\alpha^2\right) \times N\left(\beta_k^{t_{MH}}, \sigma_\beta^2\right) \times N\left(\lambda_k^{t_{MH}}, \sigma_\lambda^2\right) \times N\left(w_k^{t_{MH}}, \sigma_w^2\right) \qquad (5.15)$$

Then, with probability

$$a = min\left(1, \frac{\pi\left({\alpha_k}', {\beta_k}', \lambda_k', w_k' \mid \boldsymbol{x}_{\{i:\, C_i=k\}}\right)}{\pi\left(\alpha_k^{t_{MH}}, \beta_k^{t_{MH}}, \lambda_k^{t_{MH}}, p_k^{t_{MH}} \mid \boldsymbol{x}_{\{i:\, C_i=k\}}\right)}\right),$$

$$\begin{aligned}
&\text{set } \left(\alpha_k^{t_{MH}+1}, \beta_k^{t_{MH}+1}, \lambda_k^{t_{MH}+1}, w_k^{t_{MH}+1}\right) = (\alpha_k', \beta_k', \lambda_k', w_k') \\
&\text{otherwise } \left(\alpha_k^{t_{MH}+1}, \beta_k^{t_{MH}+1}, \lambda_k^{t_{MH}+1}, w_k^{t_{MH}+1}\right) = \left(\alpha_k^{t_{MH}}, \beta_k^{t_{MH}}, \lambda_k^{t_{MH}}, w_k^{t_{MH}}\right).
\end{aligned}$$

Here

$$\begin{aligned}
\pi\left(\alpha_k, \beta_k, \lambda_k, w_k \mid \boldsymbol{x}_{\{i:\, C_i=k\}}\right) \;&\propto \alpha_k^{\alpha^1-1} e^{-\alpha^2 \alpha_k} \times \beta_k^{\beta^1-1} e^{-\beta^2 \beta_k} \times \lambda_k^{\lambda^1-1} e^{-\lambda^2 \lambda_k} \times w_k^{a-1}\left(1-w_k\right)^{b-1} \\
&\times \textstyle\prod_{\boldsymbol{x}_i : C_i=k} \Bigg[w_k^{(l_i-1)}\left(1-w_k\right)^{(n-l_i)} F\left(x_{i,(n-(l_i-1))} \mid \alpha_k, \beta_k, \lambda_k\right)^{n-l_i} \\
&\qquad\qquad \times \textstyle\prod_{j=1}^{l} f\left(x_{i,(n+j-l_i)} \mid \alpha_k, \beta_k, \lambda_k\right) \Bigg].
\end{aligned}$$

This is performed for each of the unique clusters $k = 1, \ldots, N^*$, and initiated after t_{init} iterations of step 5.3.1. This is to allow the $\boldsymbol{\theta}_i = (\alpha_i, \beta_i, \lambda_i, w_i)$ atoms produced from the Polya urn sampling to settle into the appropriate number of unique clusters. If the Metropolis Hastings iterations are initiated immediately after the first sweep of step 5.3.1 (i.e. $t_{init} = 0$), then the sampler can spend unnecessary time updating the unique locations of many $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{N^*}^*\}$ induced by the DP, which is typically large in N^* during the initial phases of the Polya urn sweeps of 5.3.1. This is ultimately unnecessary, as typically these unique locations of $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{N^*}^*\}$ eventually get merged into larger grouping of locations during the latter phases of completed cycles of step 5.3.1. The scales of the proposal normal distributions $(\sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda)$ should be tuned depending on the dataset. The details of this tuning will likely involve paying attention to the acceptance rate during the iterations of the Metropolis Hasting algorithm across each of the unique clusters $k = 1, \ldots, N^*$. The primary mechanism at optimising this acceptance rate during this research is via $(\sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda)$ tuning parameters, as these parameters control the jump size proposed during the random walk outlined in (5.15). If these standard deviations $(\sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda)$ are too small, then the acceptance rate will be to high and thus likely will result in highly autocorrelated samples. Alternatively, if these standard deviations $(\sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda)$ are to large, then the acceptance will be to low as the sampler will be 'stuck' at its current position. Much work has been done into the performance of the Metropolis Hasting algorithm around effective proposal distributions and parameter tuning, but a good guide for an optimal acceptance rate was proposed by Neal et al. [2006], Rosenthal et al. [2011], Sherlock et al. [2010], and this was the approach adopted during this research. Other more sophisticated Metropolis Hasting approaches can be adopted over the (5.15) proposals, for example, an adaptive Metropolis Hastings algorithm that adapts the standard deviations of $(\sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda)$ during the iterations according to the acceptance rate as proposed by Haario et al. [2001] could be beneficial depending on the modelling circumstances.

5.3.3 Sample from $p\left(\nu \mid \theta_1, \ldots, \theta_N, N^*, \mathbf{X}\right)$

As discussed during section 3.5.2, and illustrated by the simulation study of figure 3.1, the ν parameter is key in contributing to the extent of smoothness of density estimates as well as the number of unique clusters induced from $G \sim \mathrm{DP}\left(\nu G_0\right)$ samples. Consequently, we place uncertainty over the ν parameter, and specify $\nu \sim G\left(\tau_1, \tau_2\right)$. This prior specification, when augmented with an additional auxiliary variable γ (the definition of which will be detailed in the preceding discussion), induces a convenient conjugacy property that allows straight-forward Gibbs sampling for ν.

Assuming a continuous prior on ν, Escobar and West [1995] showed:

$$\begin{aligned} p\left(\nu \mid \theta_1, \ldots, \theta_N, N^*, \mathbf{X}\right) &= p(\nu \mid N^*) \\ &\propto p(\nu) p(N^* \mid \nu) \end{aligned} \tag{5.16}$$

since the data $\mathbf{X}$ is conditionally independent of ν, given N^* and locations $\theta_1, \ldots, \theta_N$ (and thus partition proportions), and furthermore, since the locations $\theta_1, \ldots, \theta_N$ are conditionally independent of ν, given N^* and data $\mathbf{X}$. From (5.16), we notice a simple Gibbs procedure can be derived, i.e. given ν, we resample the $\theta_1, \ldots, \theta_N$ parameters (according the procedural steps outlined as in sections 5.3.1 and 5.3.2) and hence N^*. Then at each iteration, we sample from the condition posterior of $p(\nu \mid N^*)$. Escobar and West [1995] further showed when $\nu \sim G\left(\tau_1, \tau_2\right)$, (5.16) can be re-expressed as:

$$\begin{aligned} p\left(\nu \mid N^*, \mathbf{X}, \theta_1, \ldots, \theta_N\right) &\propto p(\nu) p(N^* \mid \nu) \\ &\propto p(\nu) \nu^{N^*} (\nu + N) \int_0^1 x^\nu (1 - \nu)^{N-1} dx. \end{aligned} \tag{5.17}$$

The key observation about (5.17) is that the distribution of $p\left(\nu \mid N^*\right)$ is in fact the marginal distribution of a joint distribution for ν and another continuous variable γ (the auxiliary variable) such that $p\left(\nu, \gamma \mid N^*\right) \propto p(\nu) \nu^{N^*-1} (\nu + N) \gamma^\nu (1 - \gamma)^{N-1}$ for $\nu > 0$ and $\gamma \in (0, 1)$. Escobar and

West [1995] finally showed the conditional distributions of $p\left(\nu \mid \gamma, N^*\right)$ and $p\left(\gamma \mid \nu, N^*\right)$ were given by:

$$\begin{aligned} p\left(\gamma \mid \nu, N^*\right) &\propto \gamma^{\nu}(1-\gamma)^{N-1} \\ p\left(\nu \mid \gamma, N^*\right) &\propto \nu^{\tau_1+N^*-1} e^{\nu(\tau_2-\log(\gamma))+N} \\ &\quad + \nu^{\tau_1+N^*-2} e^{\nu(\tau_2-\log(\gamma))+N} \end{aligned} \tag{5.18}$$

which can easily be recognised as a beta density and a mixture of gamma densities respectively. Thus, to sample ν at each Gibbs iteration with current values of N^* and ν, we initially sample the auxiliary variable γ from the beta distribution specified in (5.18), then conditioned on this γ and N^*, we sample a new ν from the mixture of Gamma distributions specified in (5.18).

More concretely, by specifying $\nu \sim G\left(\tau_1, \tau_2\right)$, and introducing the auxiliary variable γ, we then take the following samples:

$$\left(\gamma \mid \nu, N^*\right) \sim Beta\left(\nu+1, N\right)$$

$$\left(\nu \mid \gamma, N^*\right) \sim \pi_\gamma G\left(\tau_1+N^*, \tau_2-\log\left(\gamma\right)\right)+\left(1-\pi_\gamma\right) G\left(\tau_1+N^*-1, \tau_2-\log\left(\gamma\right)\right)$$

where the weights π_γ is defined by $\pi_\gamma / \left(1-\pi_\gamma\right) = \left(\theta+N^*-1\right) / \left(N\left(\tau_2-\log\left(\gamma\right)\right)\right)$ after each of the steps outlined in sections 5.3.1 and 5.3.2. This concludes one complete iteration of our posterior inference procedure.

This three step procedure of 5.3.1, 5.3.2 and 5.3.3 is followed for both the real analytics and simulated studies presented in the subsequent sections. It should be noted, that the tuning of the MCMC parameters $(t_{init}, c, \sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda, T_{MH})$ - the iterations waited till the Metropolis Hastings procedure is initiated, the number auxiliary components used in during Neal's algorithm of step 5.3.1, the standard deviations of the proposal distributions and the number samples taken during the Metropolis Hasting algorithm of step 5.3.2 - should ultimately be tuned to produce good mixing amongst the final sampled atoms of $\left(\left(\alpha_k^*\right)^t, \left(\beta_k^*\right)^t, \left(\lambda_k^*\right)^t, \left(w_k^*\right)^t\right)$, across the index of the MCMC sampler $t = 1, \ldots, T$ (T being the total number of MCMC samples) as well as across the unique atoms $k = 1, \ldots, N_t^*$, where N_t^* is the number of unique clusters at iteration t. During the retail analytics and simulation studies described in this research, these MCMC tuning parameters are selected as $(t_{init}, c, \sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda, T) = (200, 150, 0.009, 0.05, 0.02, 0.05, 5)$ for the retail analytics dataset, and $(t_{init}, c, \sigma_w, \sigma_\alpha, \sigma_\beta, \sigma_\lambda, T) = (100, 100, 0.018, 0.010, 0.015, 0.015, 100)$ for the simulation studies.

5.4 Results

We now illustrate how our methodology works in practice by performing two simulation studies, before proceeding to a real retail analytics dataset. The first example (subsection 5.4.1) generates data from our model. The second example (subsection 5.4.2) generates data using a Gamma distribution as a kernel (rather than EW). We fit our model to both datasets using vague priors.

5.4.1 Simulated data 1

We generate data using parameters for the mixtures of (5.19) which demonstrate the various behaviours that variable length order statistics sequences from an EW kernel can exhibit, namely a mixture of light and heavy tails with varying rates of order statistics terms $x_{i,(20)}$ convergence to 0, lengths, different decay rates and varying modal behaviours. Specifically, we draw 1500 samples from the following DPMM of *variable length order statistics sequences* of (5.10):

$$\begin{aligned} G &= 0.4\delta_{\boldsymbol{\theta}_1} + 0.35\delta_{\boldsymbol{\theta}_2} + 0.25\delta_{\boldsymbol{\theta}_3} \\ (\alpha_i, \beta_i, \lambda_i, w_i) \mid G &\sim G,\ i = 1, \ldots, 1500, \\ l_i &\sim 1 + Binomial\left(19, w_i\right),\ i = 1, \ldots, 1500, \\ x_{i,j} &\sim EW\left(\alpha_i, \beta_i, \lambda_i\right),\ j = 1, \ldots, 20, \end{aligned} \tag{5.19}$$

where $\boldsymbol{\theta} = (\alpha, \beta, \lambda, w)$, with $\boldsymbol{\theta}_1^* = (0.15, 0.8, 0.91, 0.65)$, $\boldsymbol{\theta}_2^* = (2.5, 3.3, 0.35, 0.75)$, $\boldsymbol{\theta}_3^* = (0.64, 1.7, 0.4, 0.9)$. Thus, for each *variable length order statistics sequences* observation vector i, the first l_i entries (i.e. variability in truncation between observations) correspond to the top l_i order statistics of the random sample $x_{i,j}, j = 1, \ldots, 20$, with the remaining entries being censored.

5.4.2 Simulated data 2

This simulated example differs from the former simulation study in that the data is simulated from a mixture of gamma distributions rather than a mixture EW distributions. The purpose of fitting our model to a mixture of Gamma distributions instead of a mixture of EW distributions is to test the inference in a less optimistic setting and establish whether the EW kernel is sufficiently flexible to capture the decay of order statistics sequences from a set of mixtures that are not a mixture of EW distributions. The mixture components $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_3^*$ are selected to produce simulated mixtures that imitate the mixtures of (5.19).

We generate 1500 samples from the following DPMM of *variable length order statistics se-*

quences from the following mixture model

$$\begin{aligned} G &= 0.4\delta_{\boldsymbol{\theta}_1} + 0.35\delta_{\boldsymbol{\theta}_2} + 0.25\delta_{\boldsymbol{\theta}_3} \\ (\alpha_i, \beta_i, w_i) \mid G &\sim G, \; i = 1, \dots, 1500, \\ l_i &\sim 1 + \mathit{Binomial}\,(19, w_i)\,, \; i = 1, \dots, 1500, \\ x_{i,j} &\sim \mathit{Gamma}\,(\alpha_i, \beta_i)\,, \; j = 1, \dots, 20, \end{aligned} \tag{5.20}$$

where $\boldsymbol{\theta} = (\alpha, \beta, w)$, with $\boldsymbol{\theta}_1^* = (0.15, 0.5, 0.65)$, $\boldsymbol{\theta}_2^* = (1.7, 1.0, 0.75)$, $\boldsymbol{\theta}_3^* = (32, 10, 0.9)$, similarly as before.

5.4.3 Prior distributions and posterior sampling

We fit our EW mixture model using the following vague priors for $(\alpha, \beta, \lambda, w)$ and ν (DP scale parameter):

$$\begin{aligned} (\alpha, \beta, \lambda, w) &\sim \mathit{Gamma}\,(1, 0.1) \times \mathit{Gamma}\,(1, 0.1) \times \mathit{Gamma}\,(1, 0.1) \times \mathit{Beta}\,(1, 1) \\ \nu &\sim \mathit{Gamma}\,(1, 1) \end{aligned}$$

respectively.

The priors for α, β and λ imply a mean of 10 and variance 100, a rather vague choice centred away from the true values. The Beta prior for w corresponds to a uniform distribution, assuming no prior information about the number of non-censored entries. The motivation of these priors is to establish the effectiveness of inferential procedure, and we do this by specifying vague priors relative to the known true values of the mixture components. We use the steps outlined in Section 5.3 for parameter inference and perform 9000 MCMC iterations with 1000 burn-in, and thin every 9 samples. We present the MCMC output based on the inference methodology of Section 5.3 on the simulated data of (5.4.1) and (5.4.2).

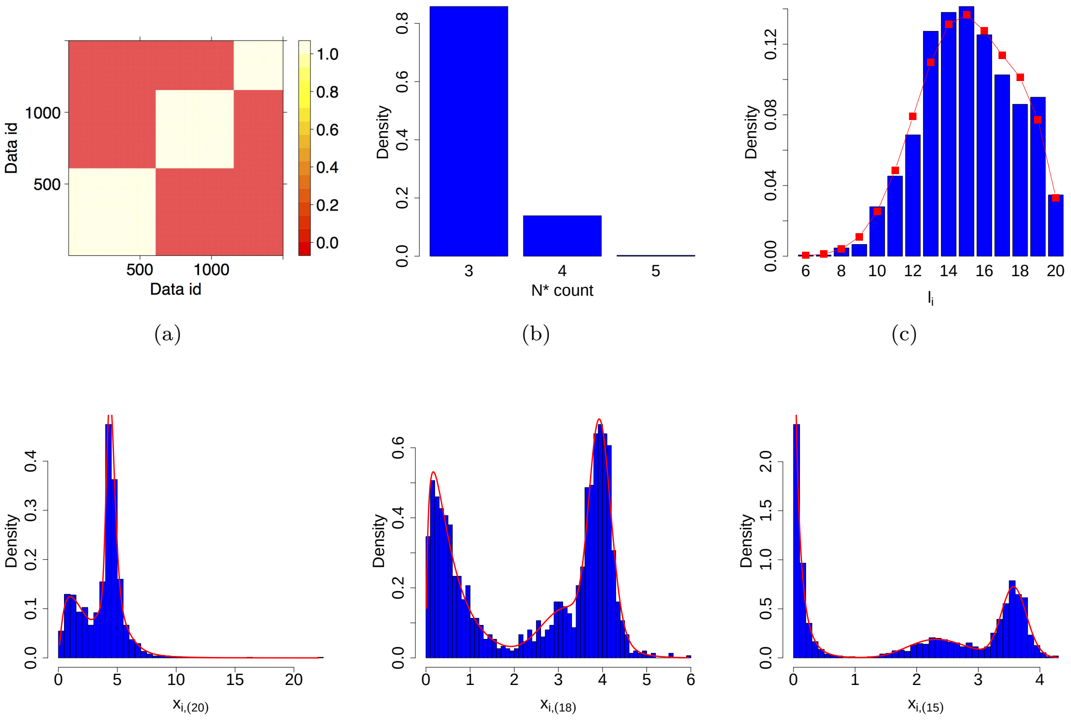
Since individual clusters are not identifiable (up to permutations), an additional identifiability criterion is required in order to perform cluster-wise inference. Binder [1978] proposed an approach of estimating the optimal co-memberships partitions based on a Bayesian clustering regime that involved the posterior coincidence probability matrix $\rho_{ij} = P(C_i = C_j)$ (computed as $\rho_{ij} = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}\left[\boldsymbol{\theta}_i^s = \boldsymbol{\theta}_j^s\right]$ where S is the number of MCMC samples and $\boldsymbol{\theta}_i^s$ is the location of i^{th} data point at the s^{th} sample). They in-particular proposed minimising the linear loss function of the posterior expected loss of the posterior marginal coincidence probabilities, which is equivalent to maximising:

$$l\left(\boldsymbol{C}^{*}, \mathrm{K}\right)=\sum_{(i,j)\in M} \mathbb{I}\left[C_{i}^{*}=C_{j}^{*}\right]\left(\rho_{ij}-K\right) \tag{5.21}$$

where $M = \{(i,j) : i < j; i,j \in \{1,\ldots,N\}\}$, $K = \frac{b}{a+b} \in [0,1]$ where b is the penalty of misclassifying two points into different clusters (when they should be) and a the penalty of misclassifying two points being the same cluster (when they shouldn't be), $\boldsymbol{C}$ is a given clustering of the observations (up to permutation). With such a partition $\boldsymbol{C}^{*}$, we are able to define cluster assignments C_i^{*} for each observation $\boldsymbol{x}_i$. Represented as an integer programming optimisation, solving (5.21) exactly is an NP hard problem, which makes it challenging to solve directly. A variety of methodologies have been devised to deal with this computational intensiveness, and during this work, we focus on using two approaches, namely, the Lau and Green [2007]'s integer programming approach and Medvedovic et al. [2004]'s agglomerative hierarchical clustering with average linkage approach, using $(1-\rho_{ij})$ as the metric representing the distance between observations i, j. We denote these respective methodologies as $\boldsymbol{C}_{LG}^{*}$ and $\boldsymbol{C}_{HC}^{*}$. The Lau and Green [2007] can be considered a more principled estimate of the optimal partitioning solution to (5.21) [Fritsch, Ickstadt, et al., 2009], as it devises a novel heuristic item-swapping algorithm guaranteed to approximate the true optimal partitioning, whereas the Medvedovic et al. [2004]'s approach only considers the subset of partitions induced from the hierarchical clustering cutting procedure, which are not guaranteed to be optimal. It should be noted however, that this more principled solution of $\boldsymbol{C}_{LG}^{*}$ comes with the computational caveat of being more challenging to compute compared to the $\boldsymbol{C}_{HC}^{*}$ partition. Where feasible, we opt for computing $\boldsymbol{C}_{LG}^{*}$ partition. Implementations calculating the partitions of $\boldsymbol{C}_{LG}^{*}$ and $\boldsymbol{C}_{HC}^{*}$ is performed using the minbinder() function from the mcclust R library [Fritsch, 2012].

Figures 5.3 and 5.4 present fitted MCMC output which includes the histogram of the number of occupied clusters N^{*}, heatmap of the posterior marginal coincidence probabilities (cluster co-membership) and density estimates of the order statistics sequences $\boldsymbol{x}_i$ for the each of the simulated studies (5.4.1) and (5.4.2) respectively. We observe that in both of simulation studies (5.4.1) and (5.4.2) that the marginal density estimates closely match the corresponding histograms of the data. This importantly demonstrates that even in the case when mixtures of variable order statistic sequences are generated from mixtures other than $EW(\alpha, \beta, \lambda)$, our model specified in (5.10) can successfully produce sound density estimates and correctly allocate the relative data-point partitioning induced by the mixture models in these examples.

During the simulated studies outlined in sections 5.4.1 and 5.4.2, we use the partition induced



(a) (b) (c)

(d) (e) (f)

Figure 5.3: Posterior probability cluster co-membership probability heatmap, histograms and density estimates for N^*, l and a few order statistics of simulated data (5.19).

by $\boldsymbol{C}^*_{HC}$ to compute cluster-wise point estimates of various quantities of interest, and table 5.2 summarises our MCMC output for the simulated study (5.4.1) using the clusters defined by $\boldsymbol{C}^*_{HC}$. The motivation for using $\boldsymbol{C}^*_{HC}$ over $\boldsymbol{C}^*_{LG}$ on the simulated datasets is based on the scaling issues related to computing $\boldsymbol{C}^*_{LG}$ for increasing N. The runtime of calculating $\boldsymbol{C}^*_{HC}$ over the simulated datasets of 5.4.1 and 5.4.2 was satisfactorily fast, taking less than 1 minute to execute. We opted for $\boldsymbol{C}^*_{HC}$ over $\boldsymbol{C}^*_{LG}$ in the simulation studies due to the runtime of calculating $\boldsymbol{C}^*_{LG}$ being computationally infeasible for $N = 1500$ (taking at least in the order of days rather minutes). Table 5.2 provides the estimates yielded from the inferential procedure outlined in Section 5.3 based on the simulated data of (5.19), grouped according to the partitions induced by $\boldsymbol{C}^*_{HC}$. This simulation study was the case where the data was generated from a mixture of $\text{EW}(\alpha, \beta, \lambda)$ distributions. We observe the estimates from table 5.2 are very close to the true parameter values, which are also contained within the 95% credibility intervals indicating the inference is working effectively.

Table 5.2: Posterior means and (2.5%,97.5%) credible intervals for the parameters $(\alpha, \beta, \lambda, w)$ of each cluster partitions induced by $\boldsymbol{C}^*_{HC}$ from simulated data (5.19). The top and bottom rows show the true parameters and number of observations assigned to each cluster respectively.

	Cluster 1	Cluster 2	Cluster 3
$(\alpha, \beta, \lambda, w) =$	(0.15,0.80,0.91,0.65)	(2.5,3.3,0.35,0.75)	(0.64,1.7,0.40,0.90)
α	0.15 (0.14, 0.17)	2.5 (2.2, 2.9)	0.67 (0.59, 0.75)
β	0.78 (0.72, 0.84)	3.3 (3.1, 3.5)	1.7 (1.5, 1.8)
λ	0.90 (0.79, 1.0)	0.35 (0.34, 0.36)	0.41 (0.38, 0.44)
w	0.64 (0.64, 0.65)	0.75 (0.75, 0.76)	0.90 (0.89, 0.91)
N	609	546	345

Table 5.3 provides the estimates yielded from the inferential procedure outlined in Section 5.3 based on the simulated data of (5.20), grouped according to the partitions induced by $\boldsymbol{C}^*_{HC}$. This simulation study was the study where the data was generated from a mixture of Gamma(α, β) distributions. The motivation of this simulation study was to challenge our inferential procedure with mixture distributions not generated under our initial EW(α, β, λ) assumptions. We observe the estimates from table 5.3 are not close to the true Gamma(α, β) parameter values, which is expected, as the parameters are unlikely to agree with the inferred EW parameters as assumed by our by inferential procedure. Though our inferential procedure does, as indicated by the partitions $\boldsymbol{C}^*_{HC}$, establish that the data was produced from a three mixture distributions and the density estimates indicated by figure 5.4 are effective at describing the decay of the order statistic sequences despite the data being generated from a mixture of Gamma distributions.

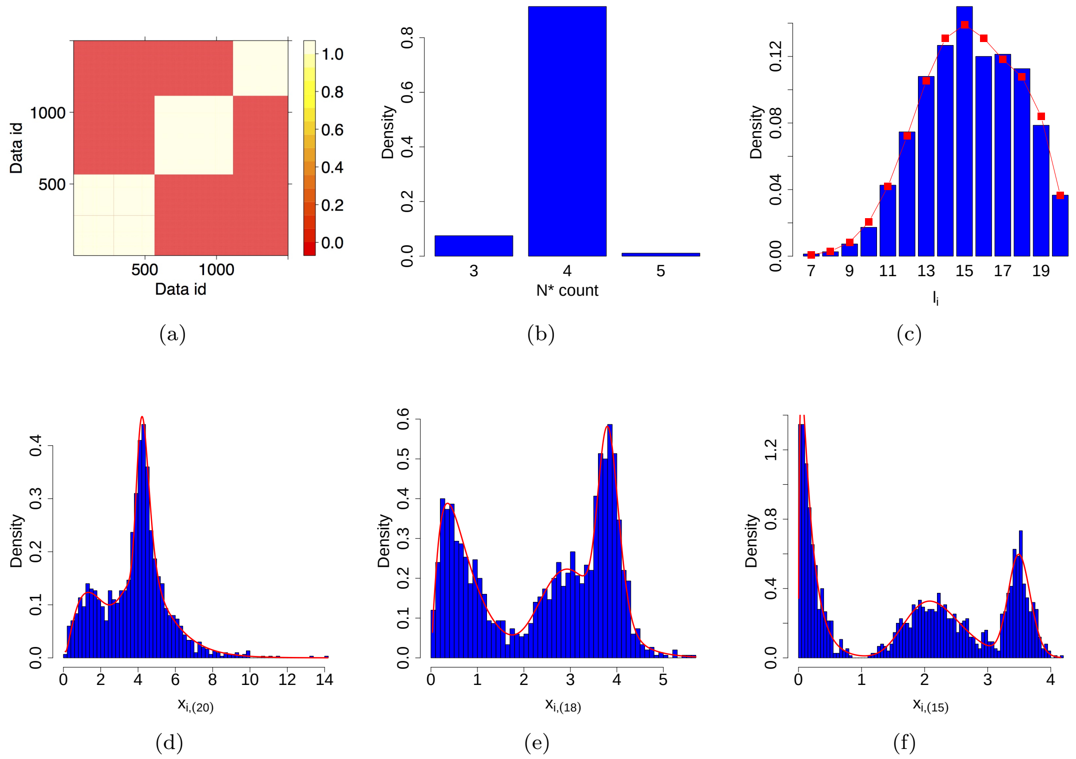
Table 5.3: Posterior means and (2.5%,97.5%) credible intervals for the parameters (α, β, w) of each cluster partitions induced by $\boldsymbol{C}^*_{HC}$ from simulated data (5.20). The top and bottom rows show the true parameters and the number of observations assigned to each cluster respectively

	Cluster 1	Cluster 2	Cluster 3
$(\alpha, \beta, w) =$	(0.15,0.5,0.65)	(1.7,1.0,0.75)	(32,10,0.9)
α	0.18 (0.16, 0.20)	1.4 (1.2, 1.7)	6.1 (5.1, 7.4)
β	0.87 (0.81, 0.94)	1.1 (1.07, 1.2)	2.7 (2.5, 2.9)
λ	0.89 (0.79, 1.0)	0.69 (0.62, 0.77)	0.42 (0.40, 0.45)
w	0.66 (0.65, 0.67)	0.75 (0.74, 0.76)	0.90 (0.89, 0.91)
N	535	591	374

5.4.4 Retail analytics dataset

We apply our method of order statistics clustering to a retail analytics dataset based on the aggregated demand from a large UK supermarket retailer. The dataset consists of the cross-elasticities for a category of supermarket products of the format described in Section 4.1.1.

$$\mathbf{X} = \{\boldsymbol{\eta}_{i,1:n} : \eta_{i,(n-l_i+1)} \leq \ldots \leq \eta_{i,(n)}, \text{with } \eta_{i,(j)} = \text{censored to 0 for } 1 \leq j \leq n - l_i\},$$



(a) (b) (c)

(d) (e) (f)

Figure 5.4: Posterior probability cluster co-membership probability heatmap, histograms and density estimates for N^*, l and a few order statistics of simulated data (5.20). These density plots demonstrate the EW kernel successfully describing the mixture of decay sequences. Although subfigure 5.4(b) indicates that the sampling procedure is often producing four occupied clusters, seemingly counter to subfigure 5.4(a)'s three clusters, we note that DP inference can produce partitions consisting of a few singleton clusters consisting of one datapoint. This is the case here, on the simulated mixture of Gamma distributions (the challenging study), where the inferential procedure largely achieves its task of good density estimates and overall partitions, but spends some time producing singleton clusters.

where we have observed only the top l_i order statistics of each cross-elasticity vector $\boldsymbol{\eta}_i$. To allow for straightforward interpretation we focus on the snacks category which consists of $N = 275$ products, consequently our data consists of $N = 275$ vectors of cross-elasticity coefficients. For this study, a maximum of $n = 10$ competitors is considered a priori to reflect a product's most significant competitors. The snack category consists of the following product line breakdown: 22.5% traditional flavoured crisps (salted, cheese and salt and vinegar), 33.1% exotic flavoured crisps (crisps excluding traditional flavours), 8.73% tortillas, 8.00% popcorn, 7.64% nuts, 4.73% dips, 2.18% pretzels and 13.1% other peripheral quick snack products. Figure 5.5 shows summary plots for the snacks category in this study. The plots provide histograms of the lengths l_i of $\boldsymbol{\eta}_i$ as well as the top two terms of the sequences and along with smooth density estimates produced from the DP model fit. The histogram of the top order statistics demonstrates spikes centred around 0.0 and 1.0, suggesting possible multi-modality.

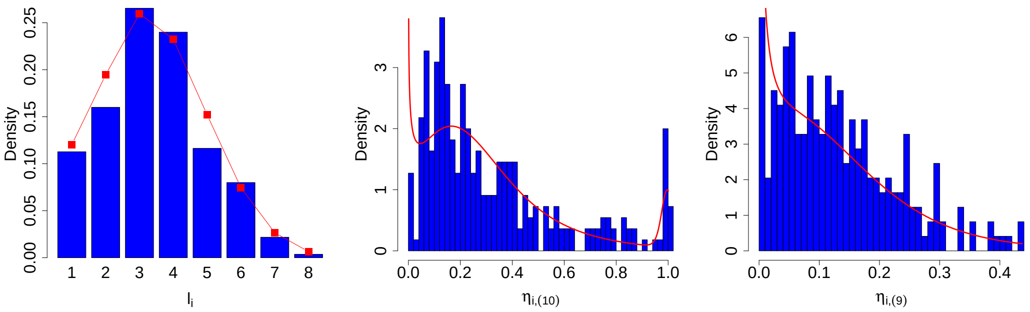


Figure 5.5: Histograms of the number of observed entries in each cross-elasticity vector, as well as the top two entries of the cross-elasticity vectors, with corresponding density estimates from our model. The censored entries (corresponding to 0 elasticities) have been omitted from the histograms.

5.4.4.1 Omitted competitors & aggregate competition

We introduce two statistics relevant to the retail analytics setting; *omitted competitors* and *mean aggregate competition*. These notions have key interpretations in the retails analytics context and will allow us to assess model fit.

Definition 1: *Omitted competitors*

Since cross-elasticity vectors arise as the outcome of penalised regression, it is natural to assume that coefficients are shrunk to zero as the result of a penalisation threshold. However under this regime, it is possible for potentially important competitor products to have been inadvertently omitted from the regression equation, meaning that the cross-elasticity vector should have included additional uncensored entries. The objective of the *omitted competitors* (OC) statistic is to assess whether the truncation has occurred prematurely by predicting the subsequent term of the observed order statistics sequence (i.e. $\eta_{i,(n-l_i)}$ of $\boldsymbol{\eta}_i$), and assessing whether this predicted value is sufficiently large. Concretely, we say an elasticity vector contains omitted competitors if its *variable length order statistics sequence* satisfies:

$$\text{OC} = \mathbb{E}_{\tilde{l},\tilde{\eta}_{(n-\tilde{l})}}\left[\tilde{\eta}_{(n-\tilde{l})} \mid \alpha, \beta, \lambda, w\right] \geq \epsilon,$$

for some truncation constant $\epsilon > 0$ and where $\tilde{\eta}_{(n-\tilde{l})}$ represents the random quantity of the $(n-\tilde{l})th$ order statistic of n i.i.d. $\mathit{EW}\left(\alpha, \beta, \lambda\right)$ samples with $\tilde{l} \sim 1 + \mathit{Binomial}\left(n-1, w\right)$. In other words, $\tilde{\eta}$ has the same distribution as η, but without any censoring. Thus the OC statistic represents the expected value of the 1^{st} censored term of a cross-elasticity vector $\tilde{\boldsymbol{\eta}}$, were we to have observed it. The value of ϵ should be chosen to represent a 'small value' within the modelling context. We set $\epsilon = 0.05$ as a sensible value to deem truncation (and will be fixed for our subsequent analysis) as it implies that if log price deviations of the next competitor is expected to account for more than 5% of equivalent log prices changes of the product's own cross-elasticity coefficient φ_i, we conclude this as a significant omission in the sales model. One of the benefits of interpreting the cross-elasticities as *variable length order statistic sequences* is the utility it provides with respect to defining OC statistic by casting censored observations into a missing data framework. The OC statistic crucially relies on being able to make a prediction of the subsequent value of a cross-elasticity vector were it to be observed. The *variable length order statistic sequence* model, by capturing the sequential decay of these decreasing sequences, allows inferences on subsequent entries of these cross elasticity vectors that flexibly incorporates the rates of decay across the previous entries.

Definition 2: *Aggregate Competition*

One of the primary interests of the analysis is characterising products in terms of their sales sensitivities with respect to their competitors' prices. We introduce the notion of *aggregate competition* (AC) to summarise the total effect of competition on a product's sales through its competitors' prices changes. We achieve this by defining the aggregate competition of product i as the sum of the top l cross-elasticity coefficients. Concretely, the AC of a cross elasticity vector distribution is given by:

$$\mathrm{AC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=n-l_i+1}^{n}\eta_{i,(j)}.$$

The AC can be thought of as the total percentage effect that log price deviations of the top l elasticity terms (where l is the expected number of competitors terms) has with respect to the equivalent prices changes of the product's own log price. For example, if a product's AC is 0.25, it means that if the log price decrease across each of its competitors was 1 unit, then the product's log price would need to decrease by 0.25 to offset the loss of sales its competitors prices changes would have had on the product's sales. Thus a large AC indicates a product's sales are significantly impacted by its competitors' prices.

5.4.4.2 MCMC output

We present the MCMC output of the real retail cross-elasticity dataset and using the following priors for $(\alpha, \beta, \lambda, w)$ and ν (DP scale parameter):

$$\begin{aligned}(\alpha, \beta, \lambda, w) &\sim \mathit{Gamma}\,(7, 7/10) \times \mathit{Gamma}\,(0.5, 1) \times \mathit{Gamma}\,(1, 1) \times \mathit{Beta}\,(2, 3)\\ \nu &\sim \mathit{Gamma}\,(5, 1)\end{aligned}$$

respectively. These priors are selected to reflect a prior expectation of the decay and typical length of the cross-elasticity vectors in the retail analytics context. The prior $w \sim Beta\left(2, 3\right)$ is selected to prefer cross elasticities of length 5, and λ, ν are uninformatively chosen. The priors over (α, β) are selected to reflect prior knowledge of the modal nature of the coefficients and the expected heavy tail nature of the cross elasticity coefficients. Figure 5.5 shows density estimates of the number of observed entries of the cross-elasticity vector, as well as the top two observed values in each vector, showing that our model is capturing these observed quantities very well. In particular, in Figure 5.5 we observe a spike of very small values for the top order statistics $\eta_{(10)}$ and $\eta_{(9)}$, which the model is able to accommodate through a small value of $\alpha\beta$. Figure 5.6 provides heatmaps of pairwise posterior distributions of the parameters which demonstrate a neat separation between pairwise atoms. Interestingly, we observe that larger values of λ (corresponding to a smaller mean) are associated with larger values of w; instead, β values are inversely associated to values of w, suggesting that the censoring in this case is largely driven by β.

Figure 5.7 presents a histogram of the number of unique clusters N^* and a heatmap of the posterior marginal coincidence probabilities (cluster co-membership). We see that the Maximum A Posteriori number of clusters is 3, with two large and one small cluster. Table 5.4 provides the category breakdown of each cluster, together with the number of observations in each as well as OC and AC values. It also includes the posterior mean and 2.5% and 97.5% posterior credible intervals of $(\alpha, \beta, \lambda, w)$ for each of optimal clusters.

For retail analytics dataset outlined in 5.4.4, we use the partition induced by $\boldsymbol{C}^*_{LG}$ to compute cluster-wise point estimates of various quantities of interest. The runtime of calculating the $\boldsymbol{C}^*_{LG}$ partitions over the retail analytics dataset was computationally feasible for $N = 275$ (executing in within one to two hours). To assess model fit, we calculate the posterior predictive p-values [Meng, 1994] of AC for each of the clusters defined by $\boldsymbol{C}^*_{LG}$. Posterior predictive p-values involves generating repetitions $\mathbf{X}^{rep}$ from the predictive distribution $p\left(\mathbf{X}^{rep} \mid \alpha, \beta, \lambda, w\right)$ for each MCMC sample and calculating p-value $= 2\left(1 - p\left(T\left(\mathbf{X}^{rep}\right) > T\left(\mathbf{X}\right) \mid \mathbf{X}\right)\right)$ for some test statistic $T\left(\mathbf{X}\right)$, in this case the aggregate competition. Figure 5.8 provides predictive posterior p-values plots on the observed aggregate competition AC over each cluster, compared against histograms of generated AC statistics over predictive replicates of $\mathbf{X}$. These all comfortably fall within the 95% prediction intervals.

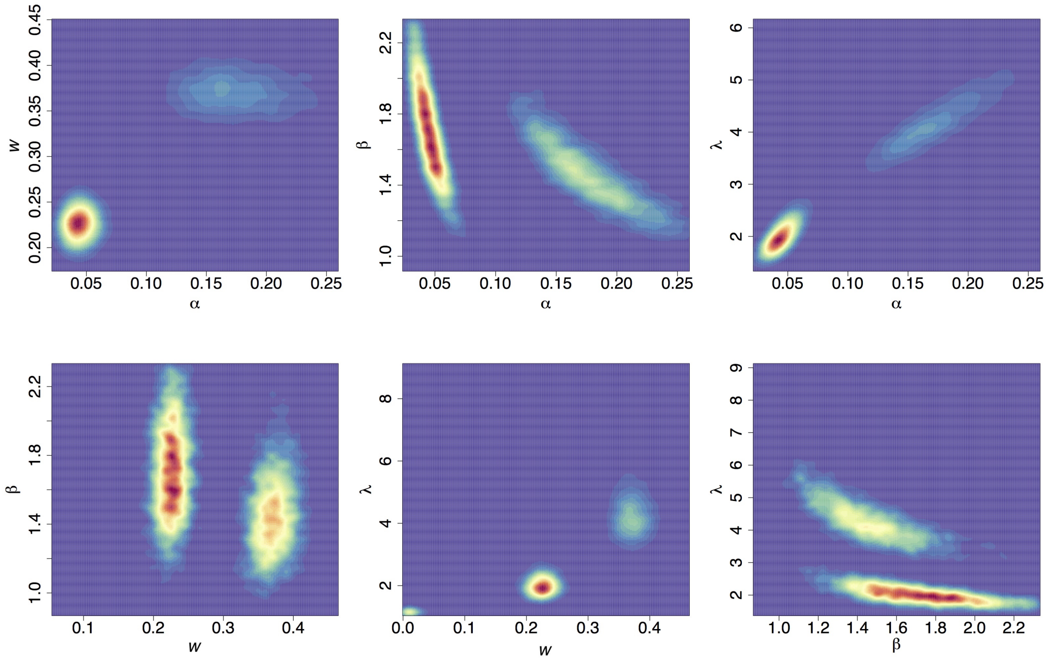


Figure 5.6: Heatmaps of pairwise posterior distributions of the parameters. The scales have been specifically selected here to produce interpretable heatmaps. It should be noted, that although three distinct clusters have been produced as a result of our inferential procedure, the smallest cluster ($\approx 6\%$ of snacks category) is only vaguely noticeable at this current scale. Even at scales where the smallest cluster could in theory be visible, it is difficult by-eye to pickup on a $\approx 6\%$ cluster through a heatmap due to its low density nature.

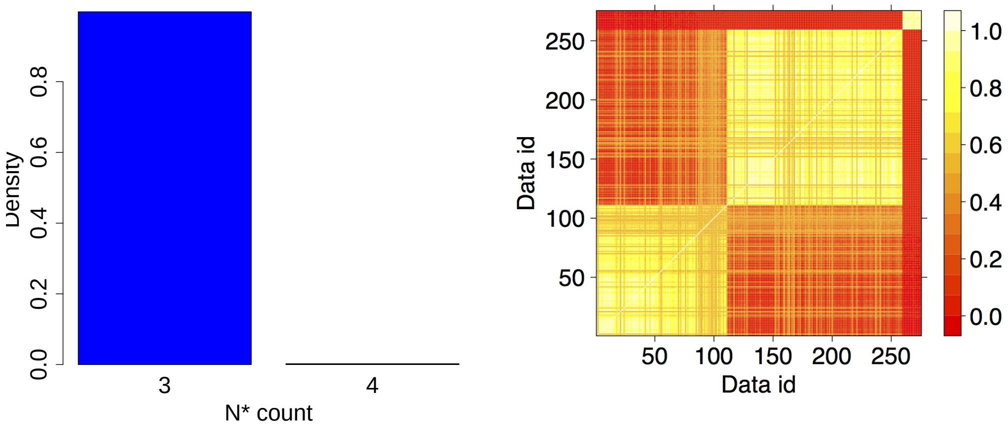
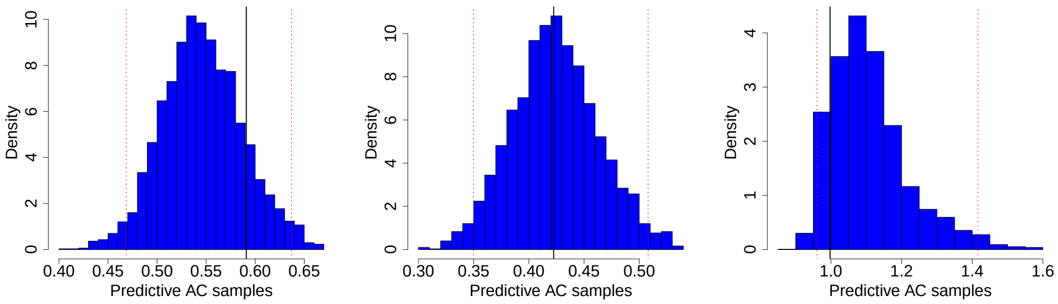


Figure 5.7: Left panel: histogram of N^*. Right panel: heatmap of cluster co-membership probabilities (re-grouped with respect to $\boldsymbol{C}^*_{LG}$).



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Figure 5.8: Histograms of AC samples with the 2.5%, 97.5% quantiles (red-dashed lines) and AC (solid black line) for each cluster induced by $\boldsymbol{C}^*_{LG}$.

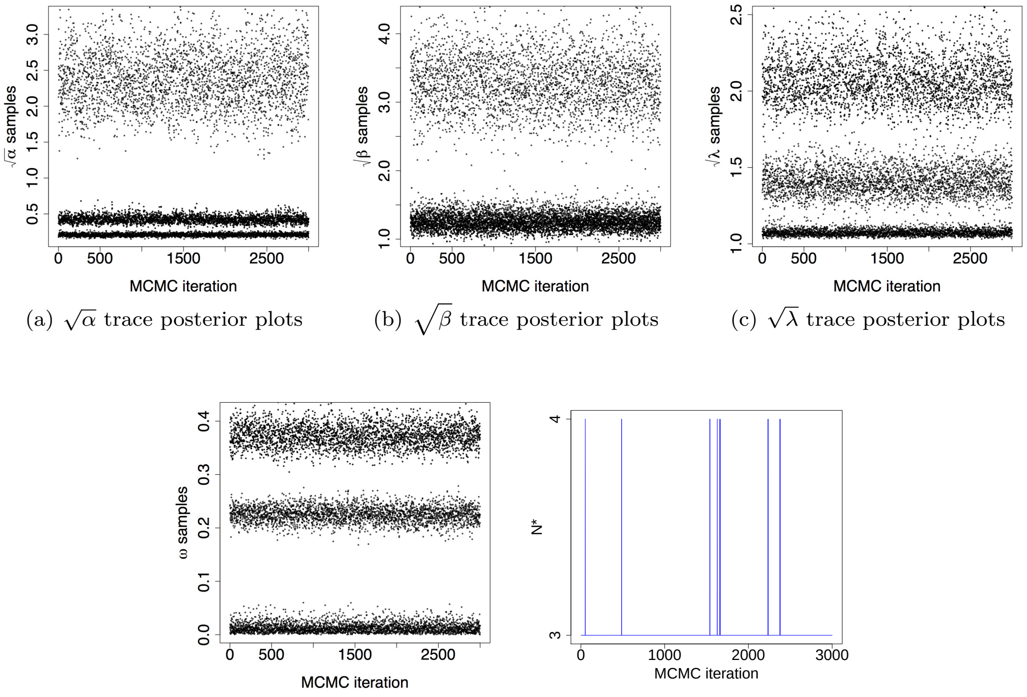
5.4.4.3 MCMC trace plots

Here we discuss convergence of our inferential procedure on the retail analytics dataset. One of issue in diagnosing MCMC convergence of a DPMM is the 'label switching' problem. The label switching problem relates to the non-identifiability of mixture components under symmetric priors, this makes it challenging to understand MCMC convergence of a DPMM since mixture components can merge, appear or disappear through MCMC sweeps which creates difficulties in diagnosing the convergence of clusterwise locations hard. We deal with this issue by providing trace plots of the atoms across all the unique clusters (demonstrating the convergence of the atom's locations) and the trace plot of the unique clusters N^* (demonstrating the convergence of $DP\left(\nu G_0\right)$ measure). Figure 5.9 provides traces of the atoms across all unique clusters $\left(\left(\alpha_k^*\right)^t, \left(\beta_k^*\right)^t, \left(\lambda_k^*\right)^t, \left(w_k^*\right)^t\right)$ of $DP\left(\nu G_0\right)$ samples for the iterations $t = 1, \ldots, T$ across the unique atoms $k = 1, \ldots, N_t^*$, where N_t^* is the number of unique clusters at iteration t and the trace of N_t^*. We plot the $\sqrt{\cdot}$ traces of (α, β, λ) to induce similar scales for graphical convenience. All plots indicate satisfactory convergence.

5.4.4.4 Retail analytics discussion

Considering the clusters given by $\boldsymbol{C}^*_{LG}$ and linking them to the corresponding categories, we see interesting breakdowns. Firstly, the first cluster has a high concentration of traditional flavoured crisps and no nut products, whereas the second cluster has a lower representation of traditional crisps. Finally, the third cluster comprises nuts, pretzels and the other product categories.

The first and second clusters appear not to have competitor products omitted from their regression models since $\mathrm{OC}_1 = 0.038, \mathrm{OC}_2 = 0.031 < \epsilon$ and thus indicate that we do not expect any of the unobserved cross-elasticities to be of any significance. However, the third cluster



(a) $\sqrt{\alpha}$ trace posterior plots (b) $\sqrt{\beta}$ trace posterior plots (c) $\sqrt{\lambda}$ trace posterior plots

(d) w trace posterior plots (e) N^* trace posterior plots

Figure 5.9: Trace plots of MCMC samples for unique atoms of $\left(\sqrt{\alpha}, \sqrt{\beta}, \sqrt{\lambda}, w\right)$ parameters and N^* on dunnhumby's cross elasticity data of the snack category. As discussed in section 5.3.2, for good overall mixing across all unique clusters induced by the DP(νG_0), it is important to have good mixing during the Metropolis Hastings phase for each of unique location updates of $(\alpha_k^*, \beta_k^*, \lambda_k^*, w_k^*)$ for $k = 1, \ldots, N^*$ (the second step of the inferential procedure outlined in 5.3). As a consequence, special attention should be paid to the Metropolis Hastings acceptance rate from sampling $p\left(\alpha^*, \beta^*, \lambda^*, w^* \mid \nu, x_{\{i:\, C_i = k\}}\right)$ across each of the unique clusters $k = 1, \ldots, N^*$. Depending on the modelling context, more sophisticated MCMC methodologies, such as adaptive Metropolis Hastings may be appropriate to produce optimal mixing rates across each of the unique clusters $k = 1, \ldots, N^*$.

exhibits competitor omission since $\mathrm{OC}_3 = 0.96 > \epsilon$. This implies that, according to the model, we expect to find at least one more competitor with a non-negligible cross-elasticity.

The posterior mean values of parameters of the first cluster are $\alpha_1 = 0.16$, $\beta_1 = 1.51$ with $w_1 = 0.34$ and an aggregate competition of $\mathrm{AC}_1 = 0.55$, which indicates that the marginal distributions of the cross-elasticities are of a light-tailed nature. This is in line with the fact that this cluster largely consists of traditional crisps, which are a fiercely competitive product line, where products have multiple substitutes and thus a high degree of sales sensitivity is expected. The second cluster exhibits similar behaviour, with posterior mean parameters $\alpha_2 = 0.06, \beta_2 = 1.71, w_2 = 0.24$ and an aggregate competition of $\mathrm{AC}_2 = 0.42$, also implying a light-tailed distribution. The third cluster is rather different; its posterior mean parameters $\alpha_3 = 5.73$, $\beta_3 = 10.88$ suggest a very light-tailed distribution. This cluster largely consists of vectors with only a single cross-elasticity entry (through $w_3 = 0.016$), and the model suggests that an additional competitor may have been missed (or does not exist) as indicated by $\mathrm{OC}_3 = 0.96 > \epsilon$. It is important to note however, that this cluster's attribute of having only a single cross-elasticity entry is not its defining trait. This cluster is also characterised by the parameters of $\alpha_3 = 5.73$, $\beta_3 = 10.88$, since this combination of parameters produces a very slow decay rate between respective entries of the order statistic sequences (markedly slower than the decay of the first and second clusters). This slow decay rate leads to the aggregate competition of the third cluster being $\mathrm{AC}_3 = 1.11$, i.e. price changes of the leading competitor products account for 1.11 of equivalent prices changes of the product's own price changes. These parameters suggest that these products are pure substitutes, i.e. products only bought as an alternative due to other equivalent products being unavailable or too expensive. This third cluster neatly demonstrates the value of clustering on both the parameters of (α, β, λ) and w, since the (α, β, λ) parameters control the decay between successive order statistic sequence entries, and w controls the number of relevant competitors a products receives throughout the market. If w were a global parameter, it is likely that products comprising of the third cluster would still need there own unique cluster, separate to that of the second and third clusters, due to the distinctive decay rate between the entries of the order statistic sequences. This is supported by figure 5.5, where we see a small spike centred around 1.0 in the histogram of $\eta_{i_i,(10)}$. This spike corresponds to the order statistic sequences of the third cluster, that are distinctly differentof the other entries of $\eta_{i_i,(10)}$ that decay markedly quicker.

With respect to the expected values of the order statistic entries themselves, we observe similar order statistic patterns between the first and second clusters; each of the first order statistics entries accounts for a roughly similar amount of its leading direct elasticity (28% and 30% respectively), however the decay rate between the subsequent order statistics of the first cluster is significantly slower than that of the second cluster (roughly 55% − 60% of their previous value compared with 40% − 45%). This decay rate observation between subsequent order statistics entries supports the discrepancy between each of the first and second cluster's AC statistics as well as the first cluster comprising of food items which traditionally have a high number of competitors than in the second cluster. Similarly as before, the third cluster differs significantly from the first and the second. Its first order statistic entry accounts for 98% of its leading direct elasticity and has a slower decay rate between successive order statistic sequences, each of these artefacts being significantly different from that of the previous clusters.

Retailers also wish to understand the behaviour of their product range at a less granular level, e.g. at a category level. Clustering of cross-elasticity profiles provides a means to extract a new summary profile for a subset of products through a principled data-driven approach. Crucially, these can aid store planners and business specialists in the retail analytics domain to better understand the optimal pricing and display combinations. For example, products in the third cluster are highly sensitive to their leading cross-effect products, but otherwise are unaffected by the bulk of products around them. On the other hand, products in the first and second clusters are cannibalized by their competitor products, meaning that increasing the sale of one product decreases the sale of another, but with the second cluster being more robust to these prices changes than the first.

Table 5.4: Retail analytics cluster-wise inference. Posterior means and (2.5%,97.5%) credible intervals for each of the four parameters $(\alpha, \beta, \lambda, w)$ along with other breakdown statistics for each the clusters induced by $\boldsymbol{C}^{*}_{LG}$.

Parameter	Cluster 1	Cluster 2	Cluster 3
α	0.16 (0.036, 0.28)	0.06 (0.030, 0.21)	5.73 (2.49, 10.19)
β	1.51 (1.10, 2.13)	1.71 (1.20, 2.39)	10.88 (5.88 ,17.06)
λ	3.89 (1.75, 5.81)	2.29 (1.59, 4.47)	1.17 (1.09, 1.29)
w	0.34 (0.21, 0.41)	0.24 (0.20, 0.39)	0.016 (0.0017, 0.042)
N	110	149	16
OC	0.038	0.031	0.96
AC	0.55	0.42	1.11
trad crisps (22.5 %)	30.9 %	18.1 %	6.25 %
exotic crisps (33.1 %)	33.6 %	35.6 %	6.25 %
tortillas (8.73 %)	11.81 %	7.38 %	0%
popcorn (8.00 %)	8.18 %	8.05%	6.25 %
nuts (7.64 %)	0%	8.72 %	50.0%
dip (4.73 %)	4.55%	5.37%	0 %
pretzels (2.18 %)	0.909 %	2.01 %	12.5%
other (13.1 %)	10.00%	14.8%	18.8%
$\mathbb{E}(\tilde{\eta}_{(10)})$	0.287	0.304	0.984
$\mathbb{E}(\tilde{\eta}_{(9)})$	0.156	0.124	0.963
$\mathbb{E}(\tilde{\eta}_{(8)})$	0.092	0.053	0.937
$\mathbb{E}(\tilde{\eta}_{(7)})$	0.055	0.022	0.926

5.5 Summary & future work

We have presented a Bayesian nonparametric mixture model for censored ordered data, using the Exponentiated Weibull distribution as a kernel. Our approach allows for flexible modelling of cross-elasticity coefficients and lends itself to meaningful interpretation. We implemented our methods on a dataset of cross-elasticities, focusing on quantities of interest in the retail analytics context, such as the aggregate competition and potential omitted competitors. Our model was able to capture several interesting features in the data through the corresponding clustering.

These methods can potentially be extended in several directions. Firstly, one could introduce structure between the distribution of the length of the order statistics sequences and the kernel distribution. This may allow borrowing of information between these two sources of information, although it will become more computationally cumbersome. Secondly, one could relax the assumption of ordered observations to account for observations only ordered in expectation. Although in the cross-elasticity context this was not appropriate, in applications such as sports analytics it may be more reflective of the data. For example, the best athlete will not always have the best performance at a competition; instead, the ranking corresponds to average performance. Finally, we would like to explore combinations of different product categories to investigate similarities in market behaviour between otherwise disparate products.

Chapter 6

Slow-moving-inventory & related methodologies

This Chapter introduces and defines the issues of demand forecasting for a class of retail products known as *slow-moving inventory* (SMI). Demand forecasting of products is of particular interest to retailers as it impacts their businesses on various operational levels; this Chapter introduces SMI forecasting and the issues related to these forecasts.

The subsequent sections are structured as follows; Section 6.1 defines the notion of SMI and describes the intermittent demand of such products along with related difficulties with forecasting SMI. The section continues on to outline the motivation behind forecasting SMI and articulates the opportunities that forecasting for this type of products offers. Section 6.2 outlines the existing work into intermittent demand forecasting of SMI. Section 6.3 describes a class of regression models known as zero-inflated regression models and discusses their relevance to the issues related to forecasting SMI. Section 6.4 describes a class of point process know as Hawkes processes and describes its relevance to the temporal aspects of forecasting SMI.

6.1 Slow-moving inventory background

One of the main objectives of retail analytics is to build predictive models for the demand of products that companies offer to their markets. Generally speaking, demand forecasting for products with high volumes of sales have been extensively studied in the literature [Seeger, Salinas, and Flunkert, 2016, Sahu, Baffour, Harper, Minty, and Sarran, 2014], and as a consequence, retailers have been successful at developing these forecasting models and have well understood the effect that traditional covariates such as price, cross-prices, seasonal indicators have on the demand of their products. However, there are a class of products known as *slow-moving inventory*, and for these products the sales volumes are significantly smaller than with most products, i.e. where a sale occurs only 5% of days. Crucially, at this level of sales volume, the data arising from the sales process of SMI is fundamentally different to products with a high sales volume, and this difference often makes demand forecasting harder. We will expand on why it is harder in subsequent sections. In spite of these difficulties, retailers remain interested in forecasting SMI for the same reasons that they are interested in forecasting demand for products with higher sales volumes, such as:

1. Understanding how variables such as a product's price, the promotional activity a product has undergone and how the seasonal trends affects a product's demand, and therefore revenue from the product, is crucially important to retailers as it allows them to understand the factors that drive demand. This allows retailers to make sensible decisions over how to set the variables they have control over. For example, a forecasting model allows experimenters to measure the effects variables have on revenue critically allows retailers to allocate marketing resources [Luan and Sudhir, 2005, Eagle and Ambler, 2002, Bass, Bruce, Majumdar, and Murthi, 2007], optimise prices [Ferreira, Lee, and Simchi-Levi, 2015, Caro and Gallien, 2012] and understand the affect of promotional activity [Deng, 2005, Zhang, Zhou, Ma, Chen, Zhang, and Agarwal, 2016]. Models capable of accurately assessing the link between demand and these variables create a competitive advantage by improving the retailers' decision making.

2. Demand forecasting for products allows retailers to manage their inventory. Thus, being able to forecast demand can allow retailers to manage their supply chain optimally both in terms of distribution between regional stores and to know when to reorder further inventory from suppliers, both of which provide a measurable improvement to their business operations [Yan and Wang, 2014, Yang, Xiao, and Kuo, 2017, Petruzzi and Dada, 1999, Oroojlooyjadid, Snyder, and Takáč, 2016, Ali and Yaman, 2013, Syntetos and Boylan, 2007]. For example, a demand forecaster at a store level allows retailers to reduce the opportunity costs associated with under/over stocking and avoid the potential loss of stock [Ferguson and Ketzenberg, 2006, Kärkkäinen, 2003, Vasconcellos and Sampaio, 2009]. Ghobbar and Friend [2003] showed that many companies have inventories that over stock products due to inaccurate demand forecasts.

Figure 6.1 provide plots for four SMI sales processes along with log(price) in £ over 364 trading days. For each product, the daily count corresponds to the aggregated sales of a touchscreen tablet across five large supermarkets within south London. These plots illustrate that the sales volumes are 'inflated' with an excess of zero sales and demonstrates an unclear correlation between demand and changes in prices and seasonal affects. We further observe a clustering effect in the succession of sales in a product's own demand series.

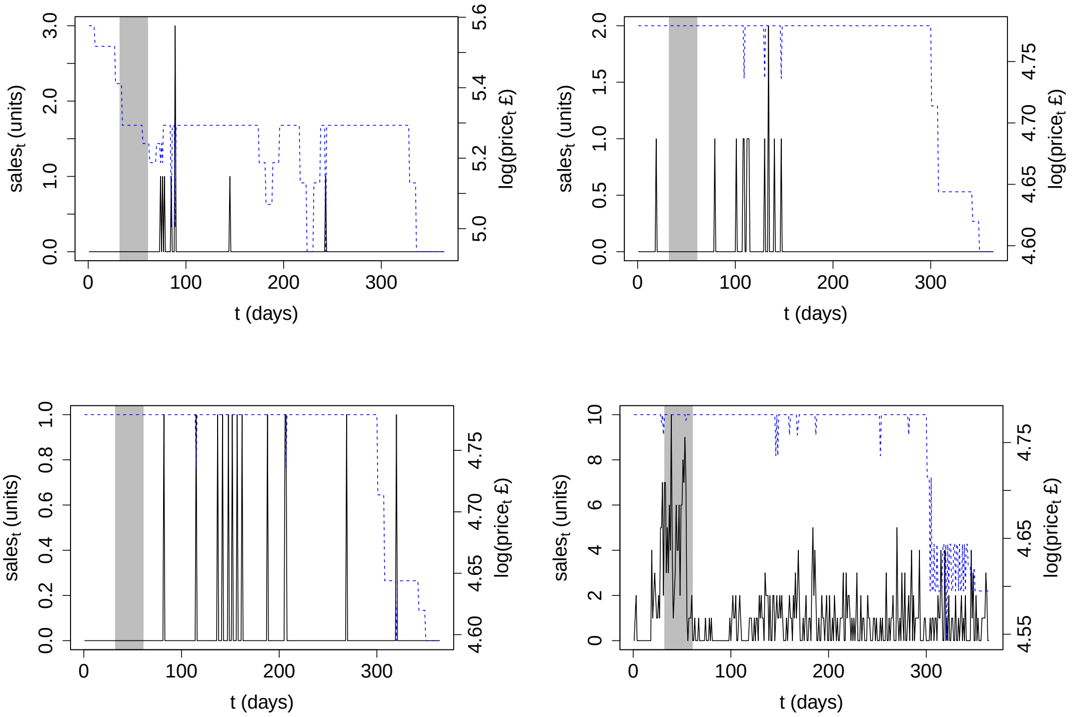


Figure 6.1: Sales plots (solid black line) for four tablets with their respective log prices in £ (dashed blue) over a subset of 364 days of demand data. The shaded region is the 30 days prior to the 25^{th} of December - a seasonal period typically associated with higher demand. The first three panels are of low volume tablets (i.e. lower number of sales) and the final forth panel is a high volume tablet (i.e. higher number of sales).

6.1.1 Challenges of slow-moving inventory forecasting

There are various aspects that make forecasting SMI difficult, but for the purposes of this work, we focus on the following three issues:

1. **Zero-inflation:** The sparsity of the sales signal that occurs for SMI products leads to an inflation of zeros (days with no sales), and these excess zero sales limit the degree to which forecasting methodologies can be deployed. In particular, this inflation of zeros has two impacts, firstly it means the implementation of traditional demand forecasting models such as (4.3) is untenable, due to the errors having non-standard distributions. Secondly, this zero-inflation often induces a low correlation between the covariates that retailers' traditionally utilise in forecasting and the sale response. This makes establishing a compelling explanatory narrative for what drives demand difficult due to the high level of uncertainty and low correlation between covariates and response.

2. **Temporal dynamics:** Another not fully understood aspect of SMI intermittent demand data is the dependency between future demand and historical demand. This temporal

component to SMI demand takes the form of 'bursty' sales across different products, i.e. that sales of product A cause further sales of product A in the future, and contemporaneous structure, i.e. that sales of product B cause sales of product C in the future. Such features have been shown to be prevalent in previous forecasting work [Seeger, Salinas, and Flunkert, 2016, Leeflang and Parreño-Selva, 2012]. During this work, the 'bursty' SMI demand is often referred to as self- and cross-excitation respectively. These bursts could be the result from some common external factor that cannot be accounted for by available covariates. An example of a common external factor could be an unexpected twitter campaign promoting a product whereas a strict dependency on the previous sales history could be the positive word-of-mouth between consumers having brought the product within a social network. This dependency of the sales processes on its recent history as well as on the sales process history of other products possibly offers a route to improve the performance of predictive models for the SMI demand process by extending the models to consider and capture autocorrelation/contemporaneous nature of sales.

3. **Short Sale cycles:** One of the practical concerns related to forecasting SMI is that SMI products are frequently stocked and sold for a relatively limited amount of time (short sale cycles). This has implications on training predictive SMI demand models, as over fitting issues can arise by little covariate and demand history and the added issue that the collective time series may not exist over the same entire time period. This lack of sales signal obfuscates how the traditional variables used in forecasting models (prices, promotions, seasonality) are linked with the volume of demand. This is particular important to retailers, as they want to understand the effect that controls have on the underlying sales process.

6.2 Analysis of slowing-moving-inventory forecasting

Much work has been done in the field of SMI forecasting, with a wide range of different methodological approaches having been used to address the challenges that forecasting SMI demand presents. Broadly speaking however, the bulk of the methodological contributions have been from the fields of machine learning, exponential smoothing (and related methods) to more traditional statistical approaches.

Exponential smoothing and related methods have been a popular class of methodology for intermittent demand forecasting of SMI products. Exponential smoothing is a sequential forecasting methodology with attempts to forecast future observations as a weighted moving average of past observations over time. More concretely, [Hunter et al., 1986] expressed exponential smoothing as:

$$\begin{aligned} l_0 &= y_0 \\ l_t &= \alpha y_{t-1} + (1-\alpha) l_{i-1} \end{aligned} \tag{6.1}$$

where y_t is sales observation at time t, l_t is the latent 'smoothed forecast' used to predict y_t and $\alpha \in (0, 1)$ is the smoothing factor. Exponential smoothing and related methods have been extensively applied to a wide range of signal processing applications [Kalekar, 2004, Ngo, Tager, and Hadley, 1996], but a relevant variation that has been heavily applied to forecast intermittent demand of SMI is known as Croston's method [Croston, 1972]. Croston's method decomposes the demand data into a count process y_t of instances of non-zero demand and l_t of inter demand intervals. Croston's approach then makes forecasts of future observations as the ratio of these two non-zero demand and inter demand intervals, i.e. $\frac{y_t}{l_t}$, assuming independence between the demand size and the inter-demand intervals. Extensions to Croston's method [Prestwich et al., 2014, Teunter et al., 2011] have included accommodating the possibility when there is no longer a demand for the marketed product and Syntetos and Boylan [2005] developed an unbiased estimator of Croston's method, which was shown to outperform Croston's original estimator on theoretically generated data. For a detailed review of Croston's method, its extensions and related exponential smoothing methods refer to [Xu, Wang, and Shi, 2012, Gardner, 2006].

In spite of exponential smoothing approaches being widely applied to forecasting intermittent demand of SMI, there have been significant methodological innovations in the literature that offer improved forecasts accuracy and explanatory power, which make these original exponential smoothing approaches less compelling [Willemain, Smart, and Schwarz, 2004, Seeger, Salinas, and Flunkert, 2016]. This relative underperformance in terms of forecast accuracy and explanatory power is arguably for a range of reasons, but it may be a result of the lack of statistical underpinning many of these methods have. Shenstone and Hyndman [2005] showed the stochastic process of Croston's method to be inconsistent with intermittent demand in that Croston's method is non-stationary and defined on a continuous space. This lack of statistical underpinning is a significant drawback to these methods as it means hypothesis testing, forecasting distributions and a framework for regression analysis are not readily available. Consequently, these approaches are often thought of as 'ad hoc' testing methods, and ones where it is often not straight-forward to optimise or select parameters [Kourentzes, 2014, Syntetos, Boylan, and Croston, 2005].

Machine learning approaches to intermittent demand forecasting of SMI have largely used neural networks and perceptrons methodologies [Kourentzes, 2013, Pour, Tabar, and Rahimzadeh, 2008, Mishra, Yuan, Huang, and Duc, 2014], and less extensively support-vector-machines [Bao, Zou, and Liu, 2006, Hua and Zhang, 2006]. Machine learning algorithms have a tradition as being methodologies capable of finding the complex non-linearities that underlie data generating processes with minimal model specification [Hornik, 1991, Hornik et al., 1989], and are appealing approaches to apply to intermittent demand forecasting. Consequently, neural networks and related methodologies have been widely applied to intermittent demand forecasting [Flunkert, Salinas, and Gasthaus, 2017, Gutierrez, Solis, and Mukhopadhyay, 2008], and have demonstrated the flexibility to accommodate for dependencies between non-zero demand and the inter-demand intervals, temporal phenomena of bursty or lumpy sales patterns as well as for the nuanced interactions between various intermittent demand between different time series.

One of the appealing aspects of neural networks and other related methodologies is the ability to fit complex and non-linear datasets. However, neural networks and alike often require a significant amount of data to train on [Kourentzes, 2013]. Markham and Rakes [1998] further demonstrated neural nets are outperformed by conventional statistical methods on smaller samples. This can inhibit machine learning algorithms' application to intermittent demand forecasting as the training signal is typically sparse and thus there can be a risk of overfitting. Regularised versions of machine learning methodologies have been successfully applied in the context on intermittent demand [Kourentzes, 2013], but a significant amount of data is necessary for these approaches to be applicable. Finally, machine learning methods' often do not have an interpretable stochastic process underpinning them. Consequently, answering hypothesis like questions such as the benefits of information sharing, measuring the effect variables have and quantifying certainty are not easily assessable. Machine learning approaches struggle with this interpretability, which is of particular interest in our SMI setting.

Finally, many statistical methodologies have been devised to handle intermittent demand forecasting of SMI. These approaches predominately make use of count models [Kocer, 2013], but also include state-space models [Seeger, Rangapuram, Wang, Salinas, Gasthaus, Januschowski, and Flunkert, 2017], modified Markov models [Kocer, 2013], or more traditional time series models [Rahman and Sarker, 2010, Mohammadipour, 2013]. Count models are often employed as a route to handling the zero-inflation observed in intermittent demand data, and consequently models such as generalised hurdle negative binomial model, beta-binomial model and hurdle shifted Poisson models, including others, have been successfully applied to forecasting intermittent demand [Hahn and Leucht, 2015, Dolgui and Pashkevich, 2008]. In addition to handling the excess of zeros in intermittent demand, extensions to static count models have included incorporating temporal dynamics to capture the lumpy and bursty nature that is often observed in such demand processes. Snyder et al. [2012] implemented various count models with damped and undamped recurrence relations on the mean of count distributions where they demonstrate a marked improvement over static traditional models that exclude temporal dynamics. Their approach however, does not make use of any explanatory variables or information borrowing between the intermittent demand series. Further extensions have included hierarchical expositions of count models that aim to pool information across related demand series. Chapados [2014] brought together a hierarchical Bayesian approach to information pool across the intermittent demand of different products with an $AR(1)$ process on the mean of the count process and further incorporated explanatory variables within the regression framework, but do not accommodate any regression or temporal framework on the zero-process.

State-space, Markov chain and time series approaches for intermittent demand forecasting of SMI have generally focused on capturing the temporal and bursty aspect of intermittent demand. Seeger et al. [2017] use an approximate Bayesian method with a latent state process to describe the burstiness of demand and Takahashi et al. [2016] similarly used a mixture of zero and Poisson distributions to demonstrate a significant improvement to Croston and related methods. Kocer [2013] used a modified Markov chain model to estimate intermittent demand and show that Markov chain based methods can capture the irregular nature of intermittent demand. Finally, with respect to more traditional times series modelling, Rahman and Sarker [2010] and Mohammadipour [2013] used conventional Bayesian times series and integer autoregressive moving average models for predicting intermittent demand.

However, there are open questions that the current statistical approaches do not sufficiently answer. Firstly, many existing approaches do not fully explore the effect covariates have on intermittent demand forecasting of SMI. Some have included a regression framework that could accommodate covariates, but the majority of applications have not. This leads us to ask whether the uncertainty exhibited in intermittent demand processes could be explained away if conditioned on the appropriates variables such as a prices or seasonality. Secondly, few approaches attempt to address the issues of hierarchical borrowing, temporal dynamics and zero-inflation in a unified way. For the approaches that do, they have not clearly separated out the benefits that covariates, hierarchical borrowing and temporal dynamics have in the context of intermittent demand forecasting. One of our key objectives is to understand the benefits that each of these modelling contributions bring to SMI forecasting. Thirdly, in the approaches incorporating temporal dynamics, though on the whole demonstrate significant benefits in each of the contexts they are applied to, are typically of linear forms that could arguably not be flexible enough in other modelling scenarios. These linear forms are typically moving averages or $AR(1)$ processes that only account for the most recent history. Temporal processes taking into account more of the history and allow for more complex linearities may be beneficial. Finally, none of approaches allow for any contemporaneous correlation across the intermittent demand series, i.e. the lumpiness in intermittent demand series occurs independently across products. It would be interesting to investigate whether there is an contemporaneous dependency across the bursty demand across products and whether such a structure could provide a utility to forecasting intermittent demand.

For the purposes of this work, we favour statistical modelling approaches. This is because on balance, statistical approaches' capability of neatly quantifying uncertainty, the ability to assess the effect of covariates and information borrowing in a hierarchical fashion are directly relevant to our objectives as outlined earlier in this Chapter.

6.3 Zero-modified distributions

The over dispersion of zero sales exhibited in intermittent demand is one of key difficulties when attempting to develop accurate demand forecasts. Much work into over dispersed count data has been done, and there are many contexts where common count distributions such as the Poisson and negative binomial distributions are not sufficiently flexible to capture the over dispersion of count data often exhibited in real-life settings [Lee, Han, Fulp, and Giuliano, 2012, Mihaylova, Briggs, O'hagan, and Thompson, 2011].

Many statistical approaches have been developed to model zero-inflated data count data, such as hurdle models, zero-inflated models, Neyman type A distribution, threshold models and Birth process models [Ridout, Demétrio, and Hinde, 1998]. For the purposes of this work, we focus on hurdle models as a route to modelling the inflation of zero sales, the reasons for which will be made clear in the subsequent section.

6.3.1 Hurdle models

Mullahy [1986] introduced the hurdle regression model to handle the inflation of zeros in count data that traditional count models could not adequately account for. The hurdle model defines a distribution over $\{0, 1, \ldots\}$, and assumes these counts can be split into two separate processes; a process accounting exclusively for the 0's (the hurdle), and a process accounting for the non-zero counts. More concretely, given an observation y assumed to be distributed according to a hurdle model, the probability mass function is given by:

$$p(y) = \begin{cases} \pi, \text{ for } y = 0 \\ (1-\pi)g(y), \text{ for } y \geq 1 \end{cases} \tag{6.2}$$

where $\pi \in [0, 1]$ is the probability of zero count and $g(\cdot)$ is some probability mass function over the positive integers. The mean and variance of the hurdle distribution is given by $\mathbb{E}[Y] = (1-\pi)\,\mathbb{E}_g[Y]$ and $\mathrm{Var}[Y] = (1-\pi)\,\mathrm{Var}_g[Y] + \pi\mathbb{E}_g[Y]$ respectively, where $\mathbb{E}_g[Y]$ and $\mathrm{Var}_g[Y]$ are the expectation and variance of the positive count distribution induced from $g(\cdot)$. Mullahy [1986] originally assumed a truncated Poisson parametrised by λ over the positive integers, and included explanatory variables $\boldsymbol{w}_i, \boldsymbol{x}_i$ with respective regression coefficients $\boldsymbol{\theta}, \boldsymbol{\beta}$ to allow π, λ to vary. Then, given a collection of y_i's with associated explanatory variables $\boldsymbol{w}_i, \boldsymbol{x}_i$, the hurdle model was given by:

$$p(y_i) = \begin{cases} \pi_i, \text{ for } y_i = 0 \\ \frac{(1-\pi_i)\exp(-\lambda_i)\lambda_i^{y_i}}{(1-\exp(-\lambda_i))y_i!} y_i \geq 0 \end{cases} \tag{6.3}$$

with link functions:

$$\mathrm{logit}(\pi_i) = \boldsymbol{w}_i\boldsymbol{\theta} \text{ and } \log(\lambda_i) = \boldsymbol{x}_i\boldsymbol{\beta}.$$

Here $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are $p \times 1$ and $q \times 1$ vectors of covariate data and $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are $p \times 1$ and $q \times 1$ vectors of regression coefficients.

The hurdle model has been applied to a range of contexts which include ecology modelling [Potts and Elith, 2006], resource planning within medicine [Molas and Lesaffre, 2010] to modelling insurance claims [Boucher, Denuit, and Guillén, 2008]. The hurdle model is closely related to the zero-inflated Poisson model [Lambert, 1992], which is essentially a two state mixture model that either mixes to a degenerate zero-distribution with probability $\pi \in [0, 1]$, or alternatively with probability $(1-\pi)$ takes a sample from a untruncated Poisson distribution. Though similar, these models differ in the following ways: firstly, the hurdle model assume the zero and non-zero processes are separable, as 0 observations arise exclusively from the degenerate 0 distribution and not from the count distribution over $\{1, \ldots\}$. Consequently, the likelihood of hurdle can be separable. Hurdle models further differ from their zero-inflated counterparts in their ability at accommodating deflated models that zero-inflated models cannot.

Extensions over the original hurdle model have been vast, but generally have included alternative specifications of link functions in both the count and zero-inflation components [Greene, 1994], as well as to allow for mixed effects [Hu, Pavlicova, and Nunes, 2011] to mention a few. One of the relevant developments has been the application of hierarchical Bayesian equivalents of hurdle and zero-inflated models. As discussed in Chapter 2, a Bayesian hierarchical model where samples are $i.i.d$ can be expressed as in (2.3), and has appealing benefits in terms of information borrowing. Many of these hierarchical extensions have been implemented in the context of multivariate longitudinal data analysis. Multivariate longitudinal data comprise multiple time series data whose observations are recorded at the same time across each of the multiple time series. Within the context of the hurdle model of (6.3), by denoting y_{it} as the observation at time t for i^{th} time series, $i = 1, \ldots, d$, and $\text{Hurdle}(\pi_i, \lambda_i)$ as the hurdle distribution with parameters π_i and λ_i, we can then express a d-dimensional longitudinal hurdle model across multiple longitudinal series as:

$$\begin{aligned} y_{it} &\overset{iid}{\sim} \text{Hurdle}(\pi_i, \lambda_i), \text{ for } t = 1, \ldots, n \\ \pi_i, \lambda_i &\sim \Upsilon(\omega) \\ \omega &\sim \Pi \end{aligned} \tag{6.4}$$

for $i = 1, \ldots, d$ where Π is the prior of the hyper-parameters ω that parametrise Υ. Such models are effective at information pooling across different time series whilst still capturing between-subject heterogeneity as well as demonstrating optimal small-sample properties [Buu, Li, Tan, and Zucker, 2012, Min and Agresti, 2005, Neelon, O'Malley, and Normand, 2010, Scheel, Ferkingstad, Frigessi, Haug, Hinnerichsen, and Meze-Hausken, 2013, Neelon, Ghosh, and Loebs, 2013]. This is particularly relevant to our intermittent demand setup, as the ability to pool parameter information between the intermittent demand series could allow a route to handling the issues related to the sparsity of the demand signal. Though it should be noted, to model the bustiness of demand, we disregard the i.i.d assumption of (6.4) to accommodate the more interest temporal dynamics typically exhibited in intermittent demand.

During this work, we opt for hurdle models over zero-inflated equivalents for the following reasons. Firstly, one of the appealing traits of hurdle models is the scope for the likelihood being separable. This decoupling of the count and zero processes can allow inference to be simplified. Furthermore, there are known issues associated with zero-inflated models in terms of estimation and identifiability as well as having poorer fit when compared to their hurdle model equivalents [Hall and Shen, 2010, Rose, Martin, Wannemuehler, and Plikaytis, 2006].

6.4 Temporal point processes

One of key aspects of intermittent demand is the bursty and lumpy nature of sales across time. As mentioned in section 6.2, the inclusion of temporal dynamics in SMI forecasting indicates significant benefits when compared to static models without such dynamics. There have been a range of approaches that incorporate temporal dynamics into demand forecasting. In the intermittent demand context, these have included state-space modelling, ARIMA process, modified Markov chain models as well as other approaches. For a detailed review of various forecasting approaches, refer to [De Gooijer and Hyndman, 2006].

For the purposes of this work however, we consider a relevant class of stochastic processes known as *temporal point processes*, that could have utility at describing the bursty and intermittent nature of sales of SMI products. Informally speaking, a temporal point process consists of an ordered sequence of arrival times of some particular events. Temporal point processes have been widely applied to a range of contexts from forecasting earthquake activity [Ogata, 1998], modelling market events data [Bowsher, 2007] to predicting Twitter tweet popularity [Zhao, Erdogdu, He, Rajaraman, and Leskovec, 2015]. Temporal point processes have proved to be a good class of model for predicting the arrival times of future events.

6.4.1 Point processes background

We now introduce some prerequisite terminology and mathematical definitions before covering the relevant temporal point processes that are relevant to this work. Namely, we define the concepts of a *counting process*, a *point process* and a *conditional intensity function.*

Definition 10. *Counting process*

A counting process is a stochastic process $\{N(t),\ \text{for } t \geq 0\}$ that satisfies the following:

1. $N(t)$ is defined over the positive integers $\mathbb{N}^+$.

2. $N(j) \leq N(t)$ for $j \leq t$.

3. $N(0) = 0$.

Definition 11. *Point process*

A point process is a collection of random variables $\{t_1, t_2, \ldots\}$ that satisfies the following:

1. $t_1 \in [0, \infty]$ for $t \geq 0$.

2. $\mathbb{P}\left(t_1 \leq t_2 \leq \ldots\right) = 1$ almost surely.

3. The number of points in some bounded region of $\mathbb{R}$ is almost surely finite.

We introduce $H(t)$ to denote the history of a point process instances prior to time t, i.e. $H(t) = (t_j : \text{for all } t_j < \text{t})$. Having now defined a point process, it is now necessary to specify a distributional form over a finite collection of $\{t_1, t_2, \ldots, t_J\}$, i.e.:

$$f(t_1, t_2, \ldots, t_J) = \prod_{j=1}^{J} f(t_j \mid H(t_j)). \tag{6.5}$$

Instead of specifying the conditional arrival distribution $f(t_j \mid H(t_j))$ directly, we can instead characterise a temporal process by a *conditional intensity function.*

Definition 12. *Conditional intensity function*

A conditional intensity function $\lambda(t)$ of an associated counting process $N(t)$ (or equivalently point process $\{t_1, t_2, \ldots, t_J\}$) is given as:

$$\lambda(t) = \lim_{h^+ \to 0} \frac{\mathbb{E}(N(t+h) - N(t) \mid H(t))}{h}.$$

given the righthand limit exists. Crucially, the conditional intensity function, if it exists, uniquely defines distributions over a point process given in (6.5) (in the finite dimensional case).

The conditional intensity function $\lambda(t)$ is a useful route to intuiting a temporal point process. More concretely, given a conditional intensity function $\lambda(t)$ and a sufficiently small interval of time $[t, t + \delta t)$, the probability of a new occurrence happening within this interval given $H(t)$ is:

$$\mathbb{P}\left(event \in [t, t + \delta t) \mid H(t)\right) = \lambda(t)dt$$

This allows an understanding of the count and point processes induced from the functional form of $\lambda(t)$. The greater $\lambda(t)$ is during the time interval $[t, t+\delta t)$, the more probable of observing an event occurring during this interval is. Consequently, $\lambda(t)$ should reflect a point processes dynamics observed in the data [Aalen, Borgan, and Gjessing, 2008]. Two popular point processes are the *homogeneous the non-homogenous Poisson process* (HPP) and *non-homogenous Poisson process* (NHPP). A counting process $N(t)$ defines a HPP if its conditional intensity function is given by $\lambda(t) = \lambda_0$. Similarly, a counting process $N(t)$ defines a NHPP if its conditional intensity function $\lambda(t) = \varphi(t)$, where $\varphi(t)$ is a function independent of its past history and depends only variables defined at current time t.

Homogeneous and non-homogeneous Poisson processes have been widely applied to a range of temporal settings [DasGupta, 2011, Saldanha, De Simone, and e Melo, 2001, Al Ajarmeh, Yu, and Amezziane, 2010]. However, in spite of these simpler models, in situations where contagion effects are observed, the assumption that the conditional intensity function is independent of its history can be violated. To accommodate this dependency on history, *exciting processes* have been introduced. Exciting processes are processes that are inclined to 'cluster' over the domain they occur over. One such process known to accommodate event arrival clustering is the Hawkes process.

6.4.2 Hawkes process

The Hawkes process was introduced as a self-exciting point process to capture the temporal clustering of events within a counting process [Hawkes, 1971]. More concretely, given a count process $N(t)$ and associated point process $\{t_1, t_2, \ldots\}$, a Hawkes process can be defined by its conditional intensity function $\lambda(t)$ given by:

$$\begin{aligned}\lambda(t) &= \varphi(t) + \int_0^t \kappa g(t-u)dN(u) \\ &= \varphi(t) + \sum_{t_j<t} \kappa g(t-t_j)\end{aligned} \tag{6.6}$$

where $\varphi(t)$ is the background intensity rate, $g(\cdot) \geq 0$ some continuous excitation kernel that controls the extent counts/events cluster together and $\kappa > 0$ is a trigger constant. A Hawkes process induces a clustering among the count process $N(t)$, where increases in κ in turn increases the probability of an event occurring in the future.

The Hawkes process can be thought of as a generalisation of a non-homogeneous Poisson process. The non-homogeneous Poisson processes assumes $\lambda(t) = \varphi(t)$, i.e. its conditional intensity is purely a function of its current time and independent of the history of previous events. This corresponds to a Hawkes process with $\kappa = 0$ and $\varphi(t) \neq constant$.

This demonstrates the differences between the Hawkes process and the closely related non-homogeneous Poisson process, in that a Hawkes process accounts for the history of event occurrences that allows for excitation dynamics. Figures 6.2 and 6.3 give an example of a counting process $N(t)$ induced from a homogeneous Poisson process and a Hawkes process respectively for 200 units of time. Interestingly, both have the same background rate, but the addition of $\sum_{t_j<t} e^{-(t-t_j)}$ in the conditional intensity function of the Hawkes process induces a strong clustering amongst the $N(t)$. This is indicated by the quick succession of blue dotted lines compared to the more evenly spaced counts generated under the HPP case.

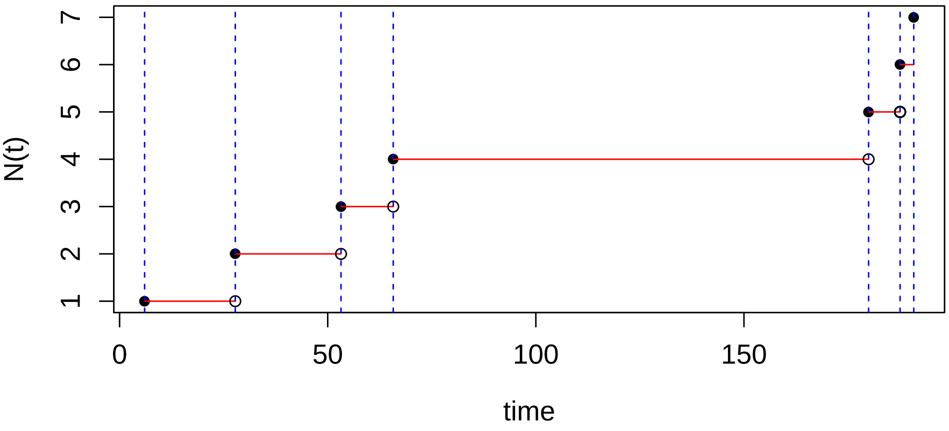


Figure 6.2: Plots of the count process $N(t)$ (as indicated by the solid blacks dots) and temporal point process $\{t_1, t_2, \ldots\}$ (as indicated by the sequence of vertical blue dotted lines) induced from the homogeneous Poisson process with conditional intensity function $\lambda(t) = \lambda_0 = 0.05$.

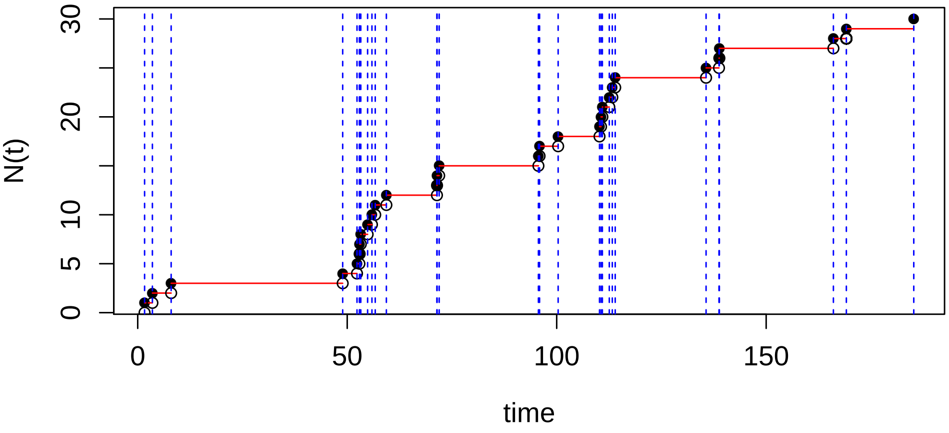


Figure 6.3: Plots of the count process $N(t)$ (as indicated by the solid blacks dots) and temporal point process $\{t_1, t_2, \ldots\}$ (as indicated by the sequence of vertical blue dotted lines) induced from the Hawkes process with conditional intensity function $\lambda(t) = 0.05 + 0.3 \sum_{t_j < t} e^{-(t-t_j)}$.

The incorporation of a point process's history in the conditional intensity function of the Hawkes process allows for excitation and contagion dynamics to be captured that a HPP and NHPP can not. Consequently, in situations where self-excitation phenomena exist, the Hawkes process has demonstrated to be a good temporal point process that accommodates for such dynamics [Da Fonseca and Zaatour, 2014, Yang and Zha, 2013].

6.4.3 Multivariate Hawkes Process

A Hawkes process can be generalised to a multivariate setting. The corresponding multivariate Hawkes Process can be thought of as a multivariate point process whose individual point process entries are defined by a conditional intensity function given by a Hawkes process that incorporates excitation from the other point process entries. Before defining the multivariate Hawkes process more formally, we introduce the concept of a *multivariate counting process.*

Definition 13. *Multivariate counting process*

A stochastic process $\{\boldsymbol{N}(\mathrm{t}),\ \text{for t} \geq 0\}$ is a multivariate counting process of d dimensions if it satisfies:

1. $N^i(t)$ defines a counting process for every $i \in \{1, \ldots, d\}$, where $N^i(t)$ is the i^{th} entry of $\boldsymbol{N}(\mathrm{t})$.

2. The sum of the coordinates of $\boldsymbol{N}(\mathrm{t})$, $\sum_{i=1}^{d} N^i(t)$, also defines a counting process.

We then denote $\{t^i_j\}_{\{i=1,\ldots,d\}}$ as the associated d-dimensional multivariate point process induced from the multivariate count process $\boldsymbol{N}(\mathrm{t})$ [Zocher, 2005]. The multivariate Hawkes process is then defined through its conditional intensities $\lambda^i(t)$, $i = 1, \ldots, d$, given by:

$$\begin{aligned}
\lambda^i(t) &= \varphi^i(t) + \sum_{k=1}^{d} \int_0^t \kappa_{ki} g_{ki}(t-u) dN^k(u) \\
&= \varphi^i(t) + \sum_{k=1}^{d} \sum_{t^k_j < t} \kappa_{ki} g_{ki}(t - t^k_j)
\end{aligned} \tag{6.7}$$

where $\varphi^i(t)$ is the background intensity of the counting process $N^i(t)$, $g_{ki}(\cdot)$ and $\kappa_{ki} > 0$ are the excitation kernel and constant corresponding to the counting process $N^k(t)$'s effect on counting process $N^i(t)$. By the inclusion of $g_{ki}(\cdot)$ with $k \neq i$ in the intensity function of $\lambda^i(t)$ allows for mutual excitation across the multivariate counting processes. Multivariate Hawkes processes have been successfully applied to a wide range of disciplines. Lai et al. [2014] proposed a scheme allowing for inter-excitation and inhibition across different social media events and themes and use a triggering kernel exponential in time and Gaussian in space to capture cross excitation and inhibition in tweets in different topics and geographies. Zhou et al. [2013] use a multivariate Hawkes process to model information spread across sparse low-rank social networks. A multivariate Hawkes process is a possible methodology that could capture the suspected dependency between the 'lumpy' sales of intermittent demand series.

6.4.4 Discretised Hawkes process

As outlined earlier, the Hawkes process has useful applications in modelling the clustering phenomena often exhibited in point process data. However, it is important to note that the Hawkes process is defined over the continuous space. Consequently, it is necessary to create a discretised equivalent to the Hawkes process that is applicable in the intermittent demand forecasting setting where forecasts are made on a daily level.

One relevant discretisation of excitation processes in modelling zero-inflated count data was that of Porter and White [2012], who interpreted a Hawkes process within the discrete setting. In particular, they let E_t be the indicator for an event day where $E_t = 1$ if on day t, there was at least one non-zero count observed, and $E_t = 0$ if on day t, there was only a 0 count observed. They then assumed $E_t \sim \text{Bernoulli}(\pi_t)$ where π_t is the probability of observing non-zero count on day t with link function:

$$\eta(\pi_t) = -\log(1 - \pi_t) = \varphi(t) + \kappa \sum_{j<t} E_j g(t - j) \tag{6.8}$$

where $\varphi(t)$ is mean intensity function on day t (which takes no account of the history), $g(\cdot)$ is a discrete excitation function and κ is the real valued excitation or inhibition constant. This essentially achieves the self-excitation dynamics that a traditional Hawkes process captures in the temporal point process setting. We observe from (6.8), that $\eta(\pi_t)$ elicits a behaviour such that for $\kappa > 0$, increases in $\varphi(t) + \kappa \sum_{j<t} E_j g(t - j)$ lead to increases the probability (π_t) of further such events occurring in the near future.

Porter and White [2012] demonstrated that this adequately captures the self-excitation exhibited in their count dataset when compared to other benchmark models. Such an approach is closely related to intermittent demand forecasting. Figure 6.4 plots two series of samples from a Bernoulli distribution with a Hawkes process term. It illustrates the variation in Bernoulli samples according to the differing parameters of the excitation kernel and trigger constant. We observe from this plot, that the maroon curve, by having a higher excitation constant κ, experienced much more excitation as exhibited by the densely packed maroon events dots, as opposed to the blue which are mostly isolated events.

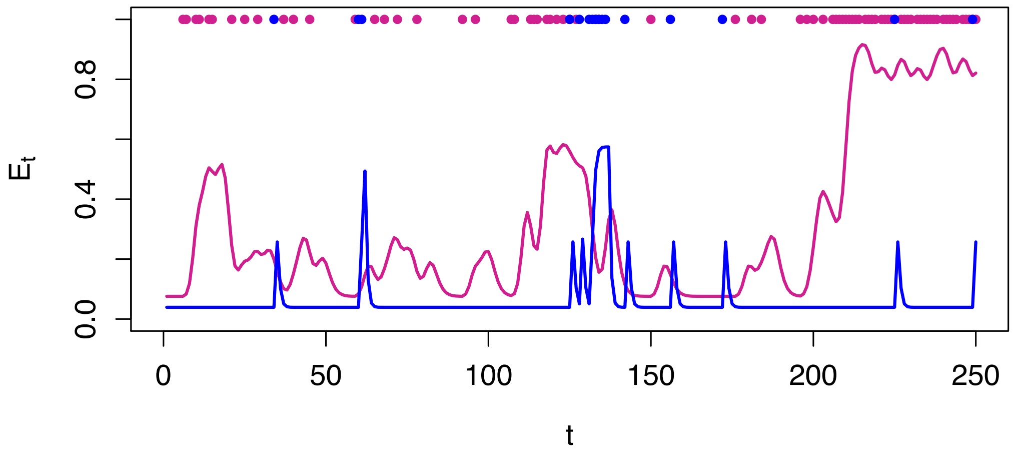


Figure 6.4: Simulated example. Two series of samples are generated from $E_t \sim \text{Bernoulli}(p_t)$, with $\text{logit}(p_t) = \theta + \kappa \sum_{j<t} E_j g(t - j \mid \mu, \tau)$ for $t = 1, \ldots, 364$ where $g(\cdot \mid \mu, \tau)$ is the negative binomial density on the positive integers with mean and scale μ, τ. The blue dots are E_t samples generated from $(\theta, \kappa, \mu, \tau) = (-3.2, 3.1, 1.0, 5.0)$ and the solid blue line is the corresponding p_t. The maroon dots are E_t samples generated from $(\theta, \kappa, \mu, \tau) = (-2.5, 5, 5, 60)$ and the solid maroon line is the corresponding p_t. We observe how the differing $(\theta, \kappa, \mu, \tau)$ lead to different clustering patterns and the underlying shape of the probability of seeing events.

Chapter 7

Slow-moving inventory prediction model

This Chapter is largely from a paper due to be submitted to The Annals of Applied Statistics titled "Bayesian hierarchical modelling of sparse count processes in retail analytics". arXiv:1805.05657

This Chapter presents a novel forecasting methodology for the intermittent demand of slow-moving inventory. Our approach accommodates the structural features exhibited in slow-moving inventory sales data, namely; zero-inflation of sales, the temporal clustering within and across intermittent demand series and the inherent information sparsity within each series. We achieve this by developing a modelling, inferential and predictive method able to learn the dynamics of sparse count processes for SMI products with few to no sales. In particular, we flexibly introduce covariates into the self-exciting model for sparse processes through the link function of the hurdle model of (6.3) similarly to that of Porter and White [2012], introduce pricing covariates into the discretised background intensity of (6.6), and further extend the model to include a cross-excitation contribution allowing for differing intermittent demand series to excite one another. Similarly to the work of Chapados [2014], we integrate individual products into a Bayesian hierarchical model that accommodates shrinkage and information passing across differing sparse count process, but further allows for excitation, seasonality and information pooling across intermittent demand series to exist in the zero-process component of the hurdle model.

The rest of this Chapter is organised as follows; section 7.1 describes the SMI demand data used in this work. Section 7.2 outlines our hierarchical Bayesian hurdle model with self- and cross-excitation components to model the multiple sparse count processes simultaneously. Section 7.3 presents the results of our sparse count process on the demand data of touchscreen tablets across five South London supermarkets. We conduct a detailed investigation to compare our model to its non-hierarchical equivalent and models without the self- and cross-excitation terms to highlight the benefits the information borrowing and excitation components and discuss the implications of these results within the context of retail analytics. Section 7.4 concludes with a summary of our contributions and a discussion of possible future developments.

7.1 Data

Our data consist of 17 longitudinal SMI sales processes over 464 days of trading between the dates 1^{st} October 2013 to 7^{th} January 2015. For each product, the daily count corresponds to the aggregated sales of a touchscreen tablet across five large south London supermarkets of a leading UK supermarket retailer. Daily prices as well as seasonality characteristics are available as covariates during the 464 trading days, during which all of the 17 tablets were stocked and in circulation. We split the data into training and test sets, the first 364 trading days between 1^{st} October 2013 to 29^{th} September 2014 (a full trading year excluding Christmas), and the remaining 100 trading days between 30^{th} September 2014 to 7^{th} January 2015 kept as hold out test set. These training and test split gives a balance between providing sufficient training periods where we observe one full year to allow the learning of seasonal trends, whilst having test sets of a reasonable size to allow meaningful forecasts to be made on. This dataset is challenging since we only have one year to learn seasonality from and thus makes a hierarchical model formulation particularly applicable. It should be noted that this data was fully anonymised for general research purposes such that no individual shoppers, or any other sensitive data could be identified.

Table 7.1 provides summary statistics over the training set of the sale counts across the 17 tablet products. The demand across the category is primarily driven by one product, as it accounts for 75% of sales. However, the remaining products are extremely slow moving as indicated by the majority of them only having 0.5-5% non-zero sales days.

Table 7.1: Summary statistics of SMI demand within tablet category on the training set. The brands have been anonymised with fictitious names for privacy purposes.

Product	Brand	total sales	% non-zero sale days
1	SPARK	1	0.27
2	TECHY	409	53.57
3	TECHY	36	4.12
4	GADGET	9	1.92
5	TECHY	5	1.37
6	TECHY	13	3.57
7	TECHY	13	3.57
8	GADGET	13	3.30
9	GADGET	2	0.27
10	GADGET	5	1.37
11	TECHY	1	0.27
12	TECHY	12	1.92
13	TECHY	2	0.55
14	TECHY	3	0.82
15	TECHY	9	0.82
16	TECHY	6	1.10
17	TECHY	3	0.82

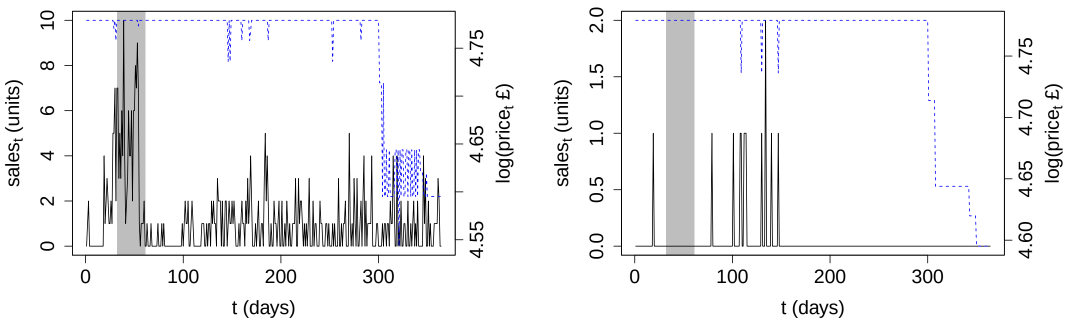


Figure 7.1: Plots of demand series (solid black) for two tablets with their respective log(prices) in £ (dashed blue) over 364 days of training data. The left and right panel demand is a high volume and low volume tablet respectively. The shaded region is the month prior to Christmas.

These data demonstrate many of the pertinent features of SMI sales processes. Figure 7.1 contrasts the sales and respective prices of a faster-selling tablet against a slower one. The plots illustrate the zero-inflation and that the sales do not show a straightforward dependence on either the prices or the seasonal effects, as indicated by the little movement in demand with respect to changes in prices and season. A clustering effect in the succession of sales within their own demand series is also evident. For example, sales of the right-hand plot in Figure 7.1 fall during the month prior to the festive period, typically thought of as driving demand, but a quick succession of sales follows shortly after this month. This suggests that an excitation process not accounted for by covariate information, as sales bursts occur outside the effects explained by covariate data. Figure 7.2 provides plots suggesting the existence of possible cross-excitation of tablet sales within a particular brand. We see that successive sales of a tablet in a given brand is often followed by a subsequent sale of another tablet of the same brand.

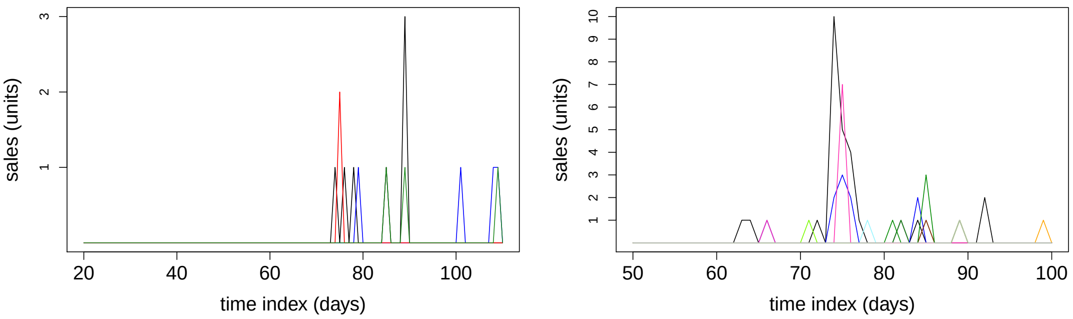


Figure 7.2: Plots of tablet sales across two brands over proportions of the training set. The left plot corresponds to the GADGET brand and the right plot to the TECHY brand. For each plot, the differing colours correspond to the sales of a particular product within the given brand. The model proposed in subsequent sections (in-particular section 7.2.2), will incorporate the cross-excitation dynamic that the sales of products within the GADGET brand can trigger further sales of differing products within the same GADGET brand (and likewise for the TECHY brand).

The model proposed in the latter section will have various boolean variables that indicate whether a sale has been triggered within a product's own recent sales history, but also whether a recent sale has been triggered of a different product within the same brand category.

7.2 Model

We model the daily sales of SMI by explicitly modelling the absence of a sale, termed the 'zero-process', and the number of sales by the 'count-process'. Our model uses a Bayesian hierarchical version of the hurdle model of (6.3), with self- and cross-excitation terms in the zero components and self-excitation terms in the count components. More concretely, given y_{it} sales of some product i on day t (where $y_{it} \in \{0, 1, \ldots\}$), the probability density function of y_{it} given covariates $\boldsymbol{x}_{it}, \boldsymbol{w}_{it}$ is specified as:

$$p(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i, \boldsymbol{\beta}_i) = \begin{cases} p(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i), \text{ for } y_{it} = 0 \\ \left(1 - p(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i)\right) f(y_{it} \mid \lambda\left(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i\right)), \, y_{it} \in \mathbb{N}^+ \end{cases}$$

with link functions:

$$\operatorname{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi_i^z(t, \boldsymbol{w}_{it}, \boldsymbol{\theta}_i) + S_{it}^z(H_{it}, \boldsymbol{\theta}_i) + \tilde{S}_{it}^z(\tilde{H}_{it}, \boldsymbol{\theta}_i), \tag{7.1}$$

$$\log(\lambda(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i)) = \varphi_i^c(t, \boldsymbol{x}_{it}, \boldsymbol{\beta}_i) + S_{it}^c(H_{it}, \boldsymbol{\beta}_i). \tag{7.2}$$

Here

$\boldsymbol{x}_{it}$ and $\boldsymbol{w}_{it}$ are the $p \times 1$ and $q \times 1$ vectors of covariate data at time t,

$\boldsymbol{\theta}_i, \boldsymbol{\beta}_i$ are the collection of coefficients of the zero and count processes respectively,

H_{it} & $\tilde{H}_{it}$ are the history self and cross events until t,

$\varphi_i^z(t, \boldsymbol{w}_{it}, \boldsymbol{\theta}_i)$ is the background intensity function of $t, \boldsymbol{w}_{it}, \boldsymbol{\theta}_i$ for the zero process,

$\varphi_i^c(t, \boldsymbol{x}_{it}, \boldsymbol{\beta}_i)$ is the background intensity function of $t, \boldsymbol{x}_{it}, \boldsymbol{\beta}_i$ for the count process,

$S_{it}^z(H_{it}, \boldsymbol{\theta}_i)$ is the self-excitation function of the zero process (as a function of $H_{it}, \boldsymbol{\theta}_i$),

all indexed by product i, and with $\tilde{S}_{it}^z(\tilde{H}_{it}, \boldsymbol{\theta}_i)$ and $S_{it}^c(H_{it}, \boldsymbol{\beta}_i)$ defined similarly. Here $p(\boldsymbol{w}_{it}, \boldsymbol{\theta}_i)$ is the probability of observing a zero sale at time t, and $f(\cdot \mid \lambda\left(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i\right))$ is a probability mass function defined on the positive integers parametrised by $\lambda\left(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i\right)$ (as function of $\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i$). It is important to note that $\boldsymbol{\theta}_i$ is the a vector coefficients that represent the collection of coefficients parametrising the $\varphi_i^z(t, \boldsymbol{w}_{it}, \boldsymbol{\theta}_i)$, $S_{it}^z(H_{it}, \boldsymbol{\theta}_i)$ and $\tilde{S}_{it}^z(\tilde{H}_{it}, \boldsymbol{\theta}_i)$ processes. $\boldsymbol{\beta}_i$ is defined similarly. For notational purposes, we express $\varphi_i^z(t, \boldsymbol{w}_{it}, \boldsymbol{\theta}_i) = \varphi_i^z(t)$, $\varphi_i^c(t, \boldsymbol{x}_{it}, \boldsymbol{\beta}_i) = \varphi_i^c(t)$, $S_{it}^z(H_{it}, \boldsymbol{\theta}_i) = S_{it}^z$, $\tilde{S}_{it}^z(\tilde{H}_{it}, \boldsymbol{\theta}_i) = \tilde{S}_{it}^z$ and $S_{it}^c(H_{it}, \boldsymbol{\beta}_i) = S_{it}^c$ in subsequent sections. We let E_{it} be the indicator for an event day such that $E_{it} = 1$ if $y_{it} \geq 1$ (a day t where at least one sales instance is observed) and $E_{it} = 0$ if $y_{it} = 0$ (a day t with no sales for product i). The functional forms of the distributions, covariates, parameters, intensity, link and excitation functions for each of the zero and count processes will be specified in more detail in subsequent sections.

Our proposed model makes the following three extensions to existing approaches. Firstly, we use covariates beyond seasonal information, in particular, we use price to assist in forecasting the demand of products along boolean seasonal variables. Secondly, we extend the zero process of hurdle models to include covariates in the background intensity, along with self- and cross-excitation terms that aims to capture the auto-correlative and contemporaneous nature of demand bursts across the SMI category. Thirdly, we build a Bayesian hierarchical model across the sales y_{it} (the sales of product i at time t) of a SMI category to allow information borrowing.

7.2.1 Covariate data

We introduce covariate data into the model through the background intensity functions $\varphi_i^z(t)$ and $\varphi_i^c(t)$ of (7.1) and (7.2). In the supermarket sales context, this corresponds to a product's own price along with seasonal effects (which are common for all products). In particular, these covariates for a product i at time t are logarithm of its price, along with the indicator functions of week day, month and Christmas period (where we define the Christmas period being the 30 trading days prior to the 25^{th} of December). We summarise these covariates as:

$$\log(\mathrm{p}_{it}) = \log(\mathrm{price}_{it}) = \text{logarithm price of SMI product } i \text{ at time } t,$$

$$\mathrm{s}_t = \left(\mathbb{1}_{(t\in\mathrm{Christmas})}, \mathbb{1}_{(t\in\mathrm{Mon})}, \ldots, \mathbb{1}_{(t\in\mathrm{Sat})}, \mathbb{1}_{(t\in\mathrm{Jan})}, \ldots, \mathbb{1}_{(t\in\mathrm{Nov})}\right).$$

We use December and Sunday as reference values when all indicators are equal to zero. Using boolean indicators allows for a natural interpretation in an information borrowing scheme, and further avoids any explicit aggregation across the SMI product data, allowing us to easily handle any issues relating to products coming in and out of circulation. As mentioned in section 7.1, all of the 17 tablets during this analysis were stocked and in circulation, but it is important to note, that the use of boolean seasonal indicator variables is the primary mechanism by which our model handles the issue of non-overlapping sale periods.

We specify the background intensities $\varphi_i^z(t)$, $\varphi_i^c(t)$ of the zero and count processes of (6.3) for product i, as:

$$\varphi_i^z(t) = \theta_{i1} + \theta_{i2}\log(\mathrm{p}_{it}) + \sum_{k=1}^{18}\theta_{i(k+2)}\mathrm{s}_{kt} \tag{7.3}$$

$$\varphi_i^c(t) = \beta_{i1} + \beta_{i2}\log(\mathrm{p}_{it}) \tag{7.4}$$

where $\{\theta_{i1}, \ldots, \theta_{i20}\}$ and $\{\beta_{i1}, \beta_{i2}\}$ are the parameters associated with the zero and count processes respectively for product i. The j index of θ_{ij} ranges from $1-20$ to include the 1 additive constant, 1 log price variable, 6 week day, 11 month and 1 Christmas indicators. These drift functions (7.3) and (7.4) describe the background intensities of the processes absent of excitation. Thus, in the zero process, we expect the background intensity to depend on a linear combination of log(price), seasonal effects and some additive constant through a given link function, whereas in the count process, we expect the background intensity to depend on a linear combination of log(price) and some additive constant through a given link function. We restrict the background intensity of the count process to exclude seasonal effects to reduce model complexity and the possibility of over-fitting. It is important to note, that for a given product i the count process only exists for t with $E_{it} = 1$. This reduces the count process data has to train on compared to the zero process. We now denote these covariates as $\boldsymbol{w}_{it} = (\mathrm{p}_{it}, \mathrm{s}_t)$ and $\boldsymbol{x}_{it} = (\mathrm{p}_{it})$ for the zero and count processes respectively in line with notation of (7.7).

7.2.2 Self- and cross-excitation

SMI demand of different but comparable products may occur in auto-correlative and contemporaneous ‘bursts’, in that, sales of a particular product may be followed by sales of a comparable product in the immediate future. These bursts can be a result of external advertising campaigns or viral dynamics, but importantly the apparent excitation not only happens auto-correlatively, but also contemporaneously across products. In the SMI context, cross-excitation is suspected to occur within brand, i.e. an instance of demand for a product leads to a higher probability of demand of a product from the same brand over the subsequent days. Concretely, we define $\tilde{E}_{it}$ as the indicator for a *cross event day* of product i of some brand such that $\tilde{E}_{it} = 1$ if $\sum_{k \in B \setminus \{i\}} y_{kt} \geq 1$, where B is the set of indices corresponding to products of the brand, and $\tilde{E}_{it} = 0$ if $\sum_{k \in B \setminus \{i\}} y_{kt} = 0$. Thus the indicator $\tilde{E}_{it}$ is 1 if there is at least one sale within the brand at time t and 0 otherwise. We denote the history of cross-events up to but not including t as $\tilde{H}_{i(t-1)} = \big(\tilde{E}_{i1}, \ldots, \tilde{E}_{i(t-1)}\big)$.

The corresponding shot noise process with the self and cross-excitation of product i then becomes:

$$S_{it} = \sum_{j<t} \kappa_i E_{it} g(t - j \mid \zeta_i) \tag{7.5}$$

$$\tilde{S}_{it} = \sum_{j<t} \tilde{\kappa}_i \tilde{E}_{it} g(t - j \mid \tilde{\zeta}_i) \tag{7.6}$$

where $\kappa_i, \tilde{\kappa}_i$ are the trigger constants for the self- and cross-excitation respectively and g is some probability mass function parametrised by ζ_i and $\tilde{\zeta}_i$ controlling the shape of future self and cross-excitation respectively. Our cross-excitation formulation of (7.6) is closely related to the multivariate Hawkes process [Hawkes, 1971], where we fix all cross-excitation kernels of a given product to 0 that correspond to a different brand, and have shared cross-excitation kernels with shared parameters for products corresponding to the same brand. We denote these collections of self- and cross-excitation parameters as $\gamma_i = (\kappa_i, \zeta_i)$ and $\tilde{\gamma}_i = \big(\tilde{\kappa}_i, \tilde{\zeta}_i\big)$ respectively.

7.2.3 Self and cross exciting hurdle model

Our SMI model uses the hurdle model specification of (7.7). In particular, for the zero process, we use the background intensity $\varphi_i^z(t)$ of (7.3) along with self- and cross-excitation components specified in (7.5) and (7.6). For the count process, we use background intensity $\varphi_i^c(t)$ of (7.4) with self-excitation term of (7.5). Our model is indexed by 17 longitudinal demand series from the tablets category over 464 (training+test) days of trading between the dates 1^{st} October 2013 to 7^{th} January 2015. The probability mass function of the hurdle model is specified as:

$$p(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i, \boldsymbol{\beta}_i) = \begin{cases} p(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i), \text{ for } y_{it} = 0 \\ \left(1 - p(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i)\right) f(y_{it} \mid \lambda\left(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i\right), \phi), \ y_{it} \in \mathbb{N}^+ \end{cases} \tag{7.7}$$

with $f(y_{it}|\lambda, \phi) = \binom{y_{ik}-2+\phi}{y_{ik}-1} \left(\frac{\lambda-1}{\lambda-1+\phi}\right)^{y_{ik}-1} \left(\frac{\phi}{\lambda-1+\phi}\right)^{\phi}$ and $\phi = 1$ which is the probability mass function of the shifted negative binomial distribution (NB) and H_{it}, $\tilde{H}_{it}$, $\boldsymbol{w}_{it}$ and $\boldsymbol{x}_{it}$ are as defined in sections 7.2.2 and 7.2.1 respectively. We opt for a shifted NB distribution over a shifted Poisson distribution on the positive counts due to the known shortcomings of the Poisson distribution at not accommodating over-dispersion adequately [Weaver, Ravani, Oliver, Austin, and Quinn, 2015]. We specify the link functions as:

$$\operatorname{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi_i^z(t) + S_{it}^z + \tilde{S}_{it}^z \tag{7.8}$$

$$\log(\lambda(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i)) = \varphi_i^c(t) + S_{it}^c \tag{7.9}$$

$\varphi_i^z(t)$ and $\varphi_i^c(t)$ are as defined from (7.3) and (7.4) respectively. We define $S_{it}^z = \sum_{s<t} \kappa_i^z E_{it} g(t-s \mid \mu_i^z, \tau_i^z)$ and $\tilde{S}_{it}^z = \sum_{s<t} \tilde{\kappa}_i^z \tilde{E}_{it} g(t-s \mid \tilde{\mu}_i^z, \tilde{\tau}_i^z)$ similarly to (7.5) and (7.6) respectively with $g(t \mid \mu, \tau) = \binom{t-2+\tau}{t-1} \left(\frac{\mu-1}{\mu-1+\tau}\right)^{t-1} \left(\frac{\tau}{\mu-1+\tau}\right)^{\tau}$ as the shifted NB distribution. We similarly define $S_{it}^c = \sum_{s<t} \kappa_i^c E_{it} g(t-s \mid \mu_i^c, \tau_i^c)$. We denote the collection of shot parameters as $\tilde{\boldsymbol{\gamma}}_i^z = (\tilde{\kappa}_i^z, \tilde{\mu}_i^z, \tilde{\tau}_i^z)$, $\boldsymbol{\gamma}_i^z = (\kappa_i^z, \mu_i^z, \tau_i^z)$ and $\boldsymbol{\gamma}_{\boldsymbol{i}}^c = (\kappa_i^c, \mu_i^c, \tau_i^c)$ and collectively denote $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{i20}, \boldsymbol{\gamma}_{\boldsymbol{i}}^z, \tilde{\boldsymbol{\gamma}}_i^z)$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \boldsymbol{\gamma}_{\boldsymbol{i}}^c)$.

Special attention is paid to the specification of hierarchical priors over the collection $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$, as they are the mechanism through which we penalise complexity and pool information. In particular, we specify $\theta_{ij} \sim \mathrm{N}(\rho_j^z, (\sigma_j^z)^2)$ and $\rho_j^z \sim \mathrm{N}(\vartheta_j^z, (\zeta_j^z)^2)$ and fix $(\sigma_j^z)^2$ for $j = 1, \ldots, 20$ and similarly specify $\beta_{ij} \sim \mathrm{N}(\rho_j^c, (\sigma_j^c)^2)$ and $\rho_j^c \sim \mathrm{N}(\vartheta_j^c, (\zeta_j^c)^2)$ and fix $(\sigma_j^c)^2$ for each $j = 1, 2$. For parameters of the shot function S_{it}^z, we specify $\gamma_{ij}^z \sim \mathrm{Gamma}(\eta_j^z, \nu_j^z)$ with $\eta_j^z \sim \mathrm{Gamma}(\alpha_j^z, \delta_j^z)$ and fix ν_j^z for each $j = 1, 2, 3$. We specify priors on $\tilde{\gamma}_{ij}^z$ and γ_{ij}^c similarly. The full details of hierarchical prior specification are contained in appendix B.1.3 and B.1.5. Thus, by denoting:

$$\boldsymbol{Y} = \left(y_{(1)(1)}, \dots, y_{(1)(364)}, y_{(2)(1)}, \dots, y_{(2)(364)}, \dots, y_{(17)(1)}, \dots y_{(17)(364)}\right)$$

$$\boldsymbol{X} = \left(\boldsymbol{x}_{(1)(1)}, \dots, \boldsymbol{x}_{(1)(364)}, \boldsymbol{x}_{(2)(1)}, \dots, \boldsymbol{x}_{(2)(364)}, \dots, \boldsymbol{x}_{(17)(1)}, \dots \boldsymbol{x}_{(17)(364)}\right)$$

$$\boldsymbol{W} = \left(\boldsymbol{w}_{(1)(1)}, \dots, \boldsymbol{w}_{(1)(364)}, \boldsymbol{w}_{(2)(1)}, \dots, \boldsymbol{w}_{(2)(364)}, \dots, \boldsymbol{w}_{(17)(1)}, \dots \boldsymbol{w}_{(17)(364)}\right)$$

$$\boldsymbol{H} = \left(H_{(1)(1)}, \dots, H_{(1)(364)}, H_{(2)(1)}, \dots, H_{(2)(364)}, \dots, H_{(17)(1)}, \dots H_{(17)(364)}\right)$$

$$\tilde{\boldsymbol{H}} = \left(\tilde{H}_{(1)(1)}, \dots, \tilde{H}_{(1)(364)}, \tilde{H}_{(2)(1)}, \dots, \tilde{H}_{(2)(364)}, \dots, \tilde{H}_{(17)(1)}, \dots \tilde{H}_{(17)(364)}\right)$$

we can write the full posterior $p_{post}\left(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{17}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{17} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H}, \tilde{\boldsymbol{H}}\right)$ as:

$$\begin{aligned} &p^z_{post}\left(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{17}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{17} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H}, \tilde{\boldsymbol{H}}\right) = \\ &p^z_{post}\left(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{17} \mid \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{H}, \tilde{\boldsymbol{H}}\right) \times p^c_{post}\left(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{17} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}\right) \end{aligned}$$

since the posteriors over the zero and count processes are separable, each denoted as $p^z_{post}\left(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{17} \mid \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{H}, \tilde{\boldsymbol{H}}\right)$ and $p^c_{post}\left(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{17} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}\right)$ respectively, where:

$$\begin{aligned} p^z_{post}\left(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{17} \mid \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{H}, \tilde{\boldsymbol{H}}\right) &= \prod_{i=1}^{17} \prod_{t=1}^{364} p(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i)^{E_t} \left(1 - p(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i)\right)^{(1-E_t)} \\ &\times \prod_{j=1}^{20} f_{norm}(\theta_i \mid \rho^z_j, (\sigma^z_j)^2) f_{norm}(\rho^z_j \mid \vartheta^z_j, (\zeta^z_j)^2) \\ &\times \prod_{j=1}^{3} f_{gamma}(\gamma^z_{ij} \mid \eta^z_j, \nu^z_j) f_{gamma}(\eta^z_j \mid \alpha^z_j, \delta^z_j) \\ &\times \prod_{j=1}^{3} f_{gamma}(\tilde{\gamma}^z_{ij} \mid \tilde{\eta}^z_j, \tilde{\nu}^z_j) f_{gamma}(\tilde{\eta}^z_j \mid \tilde{\alpha}^z_j, \tilde{\delta}^z_j) \end{aligned}$$

$$\begin{aligned} p^c_{post}\left(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{17} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}\right) &= \prod_{i=1}^{17} \prod_{t \in T_i} f(y_{it} \mid \lambda\left(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i\right), \phi) \\ &\times \prod_{j=1}^{20} f_{norm}(\beta_i \mid \rho^c_j, (\sigma^c_j)^2) f_{norm}(\rho^c_j \mid \vartheta^c_j, (\zeta^c_j)^2) \\ &\times \prod_{j=1}^{3} f_{gamma}(\gamma^c_{ij} \mid \eta^c_j, \nu^c_j) f_{gamma}(\eta^c_j \mid \alpha^c_j, \delta^c_j) \end{aligned}$$

where $T_i = \{t | y_{it} > 0\}$, i.e. T_i are the set time indices corresponding to sale days for product i over some interval of time, E_{it} defined as in section 7.2.2 and $f_{gamma}(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}$ and $f_{norm}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

7.3 Results

We fit variations of the model (7.7) to the 17 longitudinal SMI sales processes over 364 days of trading between the dates 1^{st} October 2013 to 29^{th} September 2014. We denote time interval over which we train our models as T^{train}. A hold out test set over 100 trading days between 30^{th} September 2014 to 7^{th} January 2015 is used to evaluate the predictive performance of the model variations for both the zero and count processes. We denote this test interval as T^{test}. As the zero and count processes are completely separable, we perform model inference and analysis separately.

7.3.1 Zero process variations

To assess the predictive benefits of the additions of self-excitation, cross-excitation and hierarchical components to the zero process of the hurdle model of (7.7), we implement cumulative variations of both the link functions as well as the hierarchical layering used in the modelling. These model variations are the following:

Z.1 Baseline model ($\mathbf{Base}^z_1$): We learn the zero process of the hurdle model (7.7) with link function:

$$\text{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi^z_i$$

for each $i = 1, \ldots 17$, i.e. a constant probability per product. This is the Bayesian baseline model as it estimates the zero-process independent of covariate information. The φ^z_i is estimated using non-informative priors. The performance of this model is used to verify the relative benefits that covariate information brings to SMI zero-process modelling.

Z.2 Hierarchical Bayesian ($\mathbf{HB}^z$): We learn the zero process of the hurdle model (7.7) with link function:

$$\text{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi^z_i(t)$$

for each $i = 1, \ldots 17$ with the hierarchical prior formulation discussed in section 7.2.3. This model is implemented to establish a benchmark of the simplest regression model, i.e. a model that excludes information of previous events and is used to verify the relative benefits of self excitation and cross-excitation.

Z.3 Bayesian with self-excitation ($\mathbf{BE}^z$): We learn the zero process of the hurdle model

(7.7) with link function:

$$\operatorname{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi_i^z(t) + S_{it}^z$$

for each $i = 1, \dots 17$ but exclude the hierarchical layer of the priors articulated in section 7.2.3. More concretely, we fix the parameters $\left(\rho_j^z, (\sigma_j^z)^2\right)$ and $\left(\eta_j^z, \nu_j^z\right)$ as constants rather for each j. This model is implemented to establish a benchmark of a model with excitation but without information borrowing between products and is used to verify the relative benefits of information borrowing between products.

Z.4 Hierarchical Bayesian with self-excitation (HBE^z): We learn the zero process of the hurdle model (7.7) with link function:

$$\operatorname{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi_i^z(t) + S_{it}^z$$

for each $i = 1, \dots 17$ with the hierarchical prior formulation discussed in section 7.2.3. This model is implemented to demonstrate the possible benefits of self-excitation in the standard zero inflated regression model. We use HB^z as reference as to what self-excitation provides over the model that exclusively uses the regression covariates.

Z.5 Fixed Bayesian with self-excitation (FBE^z): We learn the zero process of the hurdle model (7.7) with link function:

$$\operatorname{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi^z(t) + S_t^z$$

for each $i = 1, \dots 17$. In this implementation the parameters across the products are shared, i.e. the parameters of $\varphi_i^z(t)$ and S_{it}^z are identical across all 17 products, and hence $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{i20}, \boldsymbol{\gamma}_i^z) = (\theta_1, \dots, \theta_{20}, \boldsymbol{\gamma}^z)$. This is the maximal information sharing regime where the borrowing is to the extent that the parameters are identical across all products. We fix the parameters $\left(\rho_j^z, (\sigma_j^z)^2\right)$ and $\left(\eta_j^z, \nu_j^z\right)$ as constants. This is compared with models BE^z and HBE^z to assess the benefits of information borrowing.

Z.6 Bayesian with self and cross-excitation (BEC^z): We learn the zero process of the hurdle model (7.7) with link function:

$$\text{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi_i^z(t) + S_{it}^z + \tilde{S}_{it}^z$$

for each $i = 1, \ldots 17$ but exclude the hierarchical layer of the priors articulated in section 7.2.3. Prior specification is similar to that of BE^z but extended to include $\tilde{\boldsymbol{\gamma}}_i^z$. This is a benchmark of a model with self and cross-excitation but without an information borrowing scheme.

Z.7 Hierarchical Bayesian with self and cross-excitation (HBEC^z): This model is the full model discussed in the section 7.7. We learn the zero process of the hurdle model (7.7) with link function:

$$\text{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi_i^z(t) + S_{it}^z + \tilde{S}_{it}^z$$

for each $i = 1, \ldots 17$ with the hierarchical prior formulation discussed in section 7.2.3. The hyper-priors are selected to balance borrowing across products and penalising complexity. This hierarchical model will be cross referenced with model BEC^z.

Z.8 Fixed Bayesian with self and cross-excitation (FBEC^z): We learn the zero process of the hurdle model (7.7) with link function:

$$\text{logit}\left(p\left(\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i\right)\right) = \varphi^z(t) + S_t^z + \tilde{S}_t^z$$

for each $i = 1, \ldots 17$. In this implementation the parameters across the products are shared, i.e. the parameters of $\varphi_i^z(t), S_{it}^z, \tilde{S}_{it}^z$ are identical across all of the 17 products, and hence $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{i20}, \boldsymbol{\gamma}_i^z, \tilde{\boldsymbol{\gamma}}_i^z) = (\theta_1, \ldots, \theta_{20}, \boldsymbol{\gamma}^z, \tilde{\boldsymbol{\gamma}}^z)$. This is the maximal information sharing regime to the extent that the parameters are identical across all products. We fix the parameters $\left(\rho_j^z, (\sigma_j^z)^2\right)$, $\left(\tilde{\eta}_j^z, \tilde{\nu}_j^z\right)$ and $\left(\eta_j^z, \nu_j^z\right)$ as constants for each j. This model is compared with models BEC^z and HBEC^z to assess the benefits of information borrowing.

Parameter inference of models Base$_1^z$, HB^z, BE^z, HBE^z, FBE^z, BEC^z, HBEC^z and FBEC^z is performed by Hamiltonian Monte Carlo algorithm as outlined during section 2.2.3, and is implemented by the rstan library [Stan Development Team, 2016]. Inference is performed by the Hamiltonian Monte Carlo algorithm (via the rstan library) due to its ease at implementing Bayesian hierarchical models and because of its success at efficiently producing uncorrelated

MCMC samples. The RStan code for the model HBEC^z is included in appendix B.1.2. All other RStan models are simplifications of the HBEC^z implementation. Convergence was confirmed by Heidelberger Welch statistic across all models and parameters [Heidelberger and Welch, 1981]. The specification of hyper-priors is included in appendix B.1.3. Further MCMC implementation details are included in B.1.6.

7.3.2 Zero process fits

The predictive performance of models Base$_1^z$, HB^z, BE^z, HBE^z, FBE^z, BEC^z, HBEC^z and FBEC^z is assessed by calculating how capable each model is at predicting the probability of a sale occurring on a given day over the test interval T^{test} (30th September 2014 to 7th January 2015) for each $i = 1, \ldots, 17$ given the history of self and cross events $H_{it}, \tilde{H}_{it}$, covariate information $\boldsymbol{w}_{it}$ and posterior samples. We denote the s^{th} posterior sample of $\boldsymbol{\theta}_i$ of the i^{th} product as $\boldsymbol{\theta}_i^s$. The sales occurrence probabilities are based on the posterior samples $\boldsymbol{\theta}_i^s$ inferred from the training interval T^{train} (between 1st October 2013 to 29th September 2014). More precisely, we apply the following methodology over the test interval:

1. On given day t on the test interval and s^{th} posterior sample, we compute the full predictive posterior distribution of the probability of a sale occurring based conditioned on $\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i^s$ for each product $i = 1, \ldots, 17$.

2. We observe $y_{i(t+1)}$ (the number of sales of product i on day $t+1$) for each $i = 1, \ldots, 17$ and update the self and cross event histories $H_{i(t+1)}, \tilde{H}_{i(t+1)}$ for $i = 1, \ldots, 17$.

3. Repeat steps for each t, for each sample s and i over the test period of 30th September 2014 to 7th January 2015.

This builds up a set of daily predictive posterior probabilities p_{its} for each $s = 1, \ldots, S$ for the probability of a sale on a given day over T^{test} for each $i = 1, \ldots, 17$ based on posterior samples inferred from T^{train} conditioned on $\boldsymbol{w}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i$. Parameter inference is performed only once (over the training interval), with the predictive posterior probabilities being computed from the inferred posterior values from the fixed training interval (i.e. parameter inference is not rerun over the additional days observed during the test interval).

To evaluate the predictive performance of the models we use the log pointwise predictive density [Gelman, Hwang, and Vehtari, 2014] for each of the products $i = 1, \ldots, 17$. The log pointwise predictive density is a score that indicates the predictive accuracy of a model over a dataset - the larger the log pointwise predictive density score, the better predictive accuracy of a model. The log pointwise predictive density lppd^z for the zero process is given by:

$$\text{lppd}_i^z = \sum_{t \in T} \log \left(\frac{1}{S} \sum_{s=1}^{S} p_{its}^{E_{it}} (1 - p_{its})^{(1-E_{it})} \right)$$

where p_{its} is the prediction probability of a sale occurring for product i from posterior sample s for some model of interest. Table 7.2 provides the lppd^z scores across products and models Base_1^z, HB^z, BE^z, HBE^z, FBE^z, BEC^z, HBEC^z and FBEC^z.The subscript of lppd denotes the log pointwise predictive density for a given fitted model and product (e.g. $\text{lppd}_{HBEC,i}^z$ is the log pointwise predictive density for model HBEC^z and product i). We compute the lppd_i^z over both the test and training intervals T^{test} and T^{train} which we denote as $\text{lppd}_i^{z,\text{test}}$ and $\text{lppd}_i^{z,\text{train}}$ respectively.

Table 7.2: $\text{lppd}_i^{z,test}$ and $\text{lppd}_i^{z,train}$ scores of the zero process fits for the models Base_1^z, HB^z, BE^z, HBE^z, FBE^z,BEC^z, HBEC^z and FBEC^z and each product.

Product i	$\text{lppd}^{z,test}_{Base_1,i}$	$\text{lppd}^{z,test}_{HB,i}$	$\text{lppd}^{z,test}_{BE,i}$	$\text{lppd}^{z,test}_{HBE,i}$	$\text{lppd}^{z,test}_{FBE,i}$	$\text{lppd}^{z,test}_{BEC,i}$	$\text{lppd}^{z,test}_{HBEC,i}$	$\text{lppd}^{z,test}_{FBEC,i}$
1	-0.37	-3.16	-0.32	-2.04	-2.38	**-0.32**	-1.97	-2.30
2	-73.47	-65.66	-60.85	-55.87	-57.70	-60.42	**-55.18**	-57.20
3	-7.33	-6.81	-6.18	-5.56	**-5.24**	-6.23	-5.59	-5.27
4	-29.44	-28.27	-29.30	-28.54	-26.47	-29.00	-28.35	**-26.36**
5	-14.16	-13.09	-10.46	-12.12	-15.14	**-10.27**	-11.81	-14.89
6	-3.67	-5.80	-2.55	-3.63	-2.77	**-2.54**	-3.63	-2.79
7	-6.92	-7.42	-5.91	-5.98	**-5.70**	-6.00	-6.07	-5.76
8	-6.74	-8.95	-6.47	-6.91	-6.38	-6.42	-6.77	**-6.29**
9	-5.97	-7.27	**-5.68**	-5.98	-6.03	-5.69	-5.93	-5.99
10	-9.91	-11.30	-10.76	-10.45	-9.74	-10.60	-10.22	**-9.57**
11	-17.16	**-11.48**	-14.01	-11.79	-13.51	-13.97	-11.80	-13.47
12	-9.80	-11.86	-10.48	-10.53	-9.66	-10.30	-10.27	**-9.49**
13	-15.84	-15.25	**-9.75**	-9.99	-12.23	-9.81	-9.91	-12.10
14	-10.34	**-8.66**	-11.15	-9.93	-10.09	-11.11	-9.95	-10.13
15	**-10.36**	-11.15	-10.78	-10.49	-10.51	-10.83	-10.52	-10.55
16	**-5.61**	-7.47	-6.12	-6.60	-6.27	-6.19	-6.61	-6.27
17	-15.01	-15.23	-13.60	-13.09	-14.14	-13.66	**-13.07**	-14.14
$\sum_{i=1}^{17} \text{lppd}^{z,\text{test}}_{model,i}$	-242.10	-238.82	-214.37	-209.50	-213.98	-213.35	**-207.65**	-212.58
$\sum_{i=1}^{17} \text{lppd}^{z,\text{train}}_{model,i}$	-708.26	-699.89	-609.45	-662.65	-675.64	**-608.48**	-662.84	-675.89

Interpreting Table 7.2's lppd^{z,test} and lppd^{z,train} scores reveal some interesting findings. Firstly, we observe the model HB^z, the zero process model with covariate information, provides a significant improvement in predictive performance compared to baseline models Base$_1^z$ without covariate information. We further see that inclusion of a self-excitation component in 7.3.1 provides a marked improvement over the model HB^z without self-excitation. Figure 7.3 demonstrates an example of the benefit of self-excitation inclusion by comparing the event day prediction performance between models HBE^z and HB^z over a portion of the test set. We observe inclusion of self-excitation produces a 95% credibility interval of model HBE^z that captures a subsequent sale that model HB^z does not immediately after the first sale at $t = 382$.

Table 7.2 further indicates the predictive benefits that hierarchical extensions provide over its non-hierarchical equivalents. Figure 7.4 illustrates an example of the benefit of these hierarchical extensions by comparing event day prediction performance between models HBE^z and BE^z over a portion of the test set. We observe that by information pooling across the intermittent demand series produces a 95% credibility interval of model HBE^z that captures a sale at $t = 446$ (during the Christmas period). This is in spite of there being no sales over the Christmas period of the previous year for this product. In this way, the hierarchical model benefits from inferring parameter values from other intermittent demand series which have observed sales over the previous the Christmas period.

Table 7.2 indicates that the cross-excitation expositions of models BEC^z, HBEC^z and FBEC^z offer an improvement in event day prediction over the test set compared to their non cross-excitation counterparts (i.e. HBE^z, HBE^z and BE^z). Interesting, cross-excitation does not offer benefits in terms of the training set; but shows significant predictive gains in the test set.

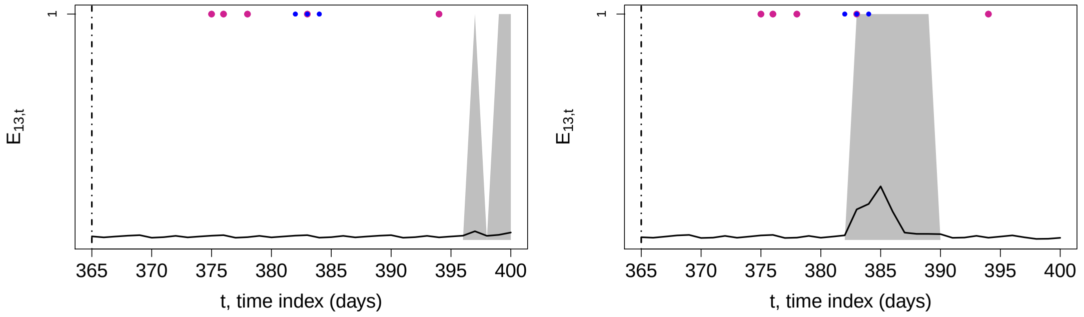


Figure 7.3: Plots of the predictive models HB^z (left plot) and HBE^z (right plot) for product $i = 13$ over a portion of the test set. The blue and magenta dots represent self and cross event days respectively (i.e. E_{it} and $\tilde{E}_{it}$). The black line is the estimated posterior mean of an event day observation (i.e. p_{it}) and the shaded region is the 95% credible interval of these estimates.

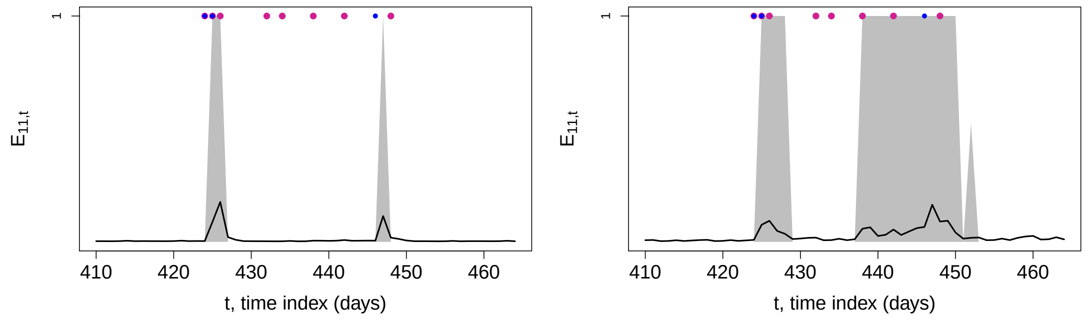


Figure 7.4: Plots of the predictive models BE^z (left plot) and HBE^z (right plot) for product $i = 11$ over a portion of the test set. The blue and magenta dots represent self and cross event days respectively (i.e. E_{it} and $\tilde{E}_{it}$). The black line is the estimated posterior mean of an event day observation (i.e. p_{it}) and the shaded region is the 95% credible interval of these estimates.

7.3.3 Count process variations

Similarly to section 7.3.2, the benefits of the excitation and hierarchical component to the count process are verified by implementing cumulative variations in the link functions and hierarchical layerings of the model. These model variations follow the same rationale as with the zero process. In particular:

C.1 Baseline model (Base$_1^c$): We learn the count process of the hurdle model (7.7) with link function:

$$\log(\lambda(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i)) = \varphi_i^c$$

for each $i = 1, \ldots 17$, i.e. a constant rate per product. This is the Bayesian baseline model as it estimates the zero-process independent of covariate information. The φ_i^c is estimated using non-informative priors.

C.2 Hierarchical Bayesian (HB^c): We learn the count process of the hurdle model (7.7) with link function:

$$\log(\lambda(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i)) = \varphi_i^c(t)$$

for each $i = 1, \ldots 17$ with the hierarchical prior formulation discussed in section 7.2.3.

C.3 Bayesian with self-excitation (BE^c): We learn the count process of the hurdle model (7.7) with link function:

$$\log(\lambda(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i)) = \varphi_i^c(t) + S_{it}^c$$

for each $i = 1, \ldots 17$ but exclude the hierarchical layer of the priors articulated in section 7.2.3.

C.4 Hierarchical Bayesian with self-excitation (HBE^c): This is the full model discussed in the section 7.2.3. We learn the count process of the hurdle model (7.7) with link function:

$$\log(\lambda(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i)) = \varphi_i^c(t) + S_{it}^c$$

for each $i = 1, \ldots 17$ with the hierarchical prior formulation discussed in section 7.2.3.

C.5 Fixed Bayesian with self-excitation (FBE^c): We learn the count process of the hurdle model (7.7) with link function:

$$\log(\lambda(\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i)) = \varphi^c(t) + S_t^c$$

for each $i = 1, \ldots 17$.

Parameter inference of models HB^c, BE^c, HBE^c and FBE^c is performed by Hamiltonian Monte Carlo sampling algorithm and is implemented by the `rstan` library [Stan Development Team, 2016]. The RStan code for the model HBE^c is included in appendix B.1.4. All other RStan models are simplifications of the HBE^c implementation. Convergence was confirmed by Heidelberger Welch statistic across all models and parameters [Heidelberger and Welch, 1981]. The specification of these hyper-priors and constant of models HB^c, BE^c, HBE^c, and FBE^c is included in appendix B.1.5. For further MCMC implementation details refer to appendix B.1.7.

7.3.4 Count process fits

Similarly as with the zero processes outlined in section 7.3.2, we test the performance of the count variation models Base$_1^c$, HB^c, BE^c, HBE^c and FBE^c by calculating how capable each model is of predicting the volume of sales on event days (i.e. days when sale has been observed) over the test interval T^{test} (between 30^{th} September 2014 to 7^{th} January 2015) for each $i = 1, \ldots, 17$ given the history of self events H_{it}, covariate information $\boldsymbol{x}_{it}$ and posterior samples. We denote the s^{th} posterior sample of $\boldsymbol{\beta}_i$ of the i^{th} product as $\boldsymbol{\beta}_i^s$. The predictive distribution is based on the posterior samples fits inferred from the training interval T^{train} (between 1^{st} October 2013 to 29^{th} September 2014). We apply the following methodology over the test interval:

1. On event day t (i.e. $E_t = 1$) on the test interval and s^{th} posterior sample, we compute the full predictive posterior distribution of the volume of sales occurring conditioned on

$H_{it}, \boldsymbol{x}_{it}, \boldsymbol{\beta}_i^s$ for each $i = 1, \ldots, 17$.

2. We observe $y_{i(t+1)}$ (the volume of sales of product i on day $t+1$) for each $i = 1, \ldots, 17$ and update the self event histories $H_{i(t+1)}$ for $i = 1, \ldots, 17$.

3. Repeat steps for each t, for each sample s and i over the test period of 30^{th} September 2014 to 7^{th} January 2015.

This builds up a set of posterior rates λ_{its} for samples $s = 1, \ldots, S$ for the probability of the number of sales on a given event day over T^{test} for each $i = 1, \ldots, 17$ based on our posterior sample fits inferred from T^{train} conditioned on $\boldsymbol{x}_{it}, H_{it}, \boldsymbol{\beta}_i$. As with the zero process, parameter inference is performed only once over the training interval, and not rerun over the additional days observed during the test interval.

Similarly as with the zero process, we evaluate the predictive performance by calculating the log pointwise predictive density for each of the products $i = 1, \ldots, 17$. The log pointwise predictive density lppd^c for the count process is given by:

$$\text{lppd}_i^c = \sum_{t \in T_i} \log \left(\frac{1}{S} \sum_{s=1}^{S} \binom{y_{ik} - 2 + \phi}{y_{ik} - 1} \left(\frac{\lambda_{its} - 1}{\lambda_{its} - 1 + \phi} \right)^{y_{ik}-1} \left(\frac{\phi}{\lambda_{its} - 1 + \phi} \right)^{\phi} \right)$$

where $\phi = 1$ and λ_{its} is the prediction mean of count sales occurring for product i from the s^{th} posterior sample for some model of interest and $T_i = \{t | y_{it} > 0\}$, i.e. T_i are the set time indices corresponding to sale days for product i over some interval of time. Table 7.3 provides the lppd^c scores for across products and models Base$_1^c$, HB^c, HBE^c, BE^c and FBE^c. Interpreting Table 7.3's lppd$^c_{model,i}$ scores reveals some interesting findings. Firstly, we observe that the model variations of HB^c, BE^c, HBE^c and FBE^c perform significantly better than the baseline model Base$_1^c$ with no covariates. This provides evidence in favour of the hypothesis that the model of HB^c, BE^c, HBE^c and FBE^c are capturing the SMI count process in a meaningful way. Similarly as with the zero process, Table 7.3 indicates the count process uniformly benefits from the inclusion of self-excitation. We further see that the count process benefits more from the hierarchical borrowing across the intermittent demand series.This is understandable given the level of sparsity in the count process. As Table 7.1 indicates, the order of sales that the each intermittent demand series has is very small (typically in the order 3-20 sales), and thus it may be expected that information borrowing would particularly benefit the individual models. An example of this additive strength of the hierarchical exposition of the count model variations is illustrated by Figure 7.5. This plot shows a histogram of y_{it} against the sum of $\sum_{t:y_{it}=k} y_{it}$ (for product 12) with corresponding 95% credibility intervals of posterior predictive distributions

Table 7.3: lppd_i^c scores of the count process fits for the models Base_1^c, HB^c, BE^c, HBE^c and FBE^c for each product and fitted model.

Product i	$\text{lppd}^{c,test}_{Base_0,i}$	$\text{lppd}^{c,test}_{HB,i}$	$\text{lppd}^{c,test}_{BE,i}$	$\text{lppd}^{c,test}_{HBE,i}$	$\text{lppd}^{c,test}_{FBE,i}$
1	0.00	0.00	0.00	0.00	0.00
2	-18.10	-18.78	**-13.59**	-14.18	-14.27
3	-0.91	-0.55	-0.62	-0.48	**-0.30**
4	**-1.60**	-1.78	-1.66	-1.77	-1.73
5	-0.08	**-0.07**	-0.08	-0.66	-0.82
6	0.00	0.00	0.00	0.00	0.00
7	-0.01	**-0.00**	-0.04	-0.22	-0.33
8	-4.99	-4.16	-7.92	-3.17	**-2.89**
9	-2.54	**-1.40**	-1.50	-1.60	-1.63
10	-3.98	-3.98	-3.80	**-2.04**	-2.05
11	**-7.05**	-7.07	-7.45	-10.95	-12.24
12	-1.02	-1.09	-1.03	-0.68	**-0.62**
13	-3.46	-3.47	-3.47	**-2.33**	-2.55
14	-6.19	6.46	-6.48	**-5.23**	-5.51
15	-2.04	-2.05	-1.95	-0.66	**-0.51**
16	**-1.57**	-2.64	-1.63	-1.80	-1.73
17	-0.10	**-0.08**	-0.09	-0.55	-0.54
$\sum_{i=1}^{17} \text{lppd}^{c,\text{test}}_{model,i}$	-53.64	-53.60	-51.32	**-46.33**	-47.70
$\sum_{i=1}^{17} \text{lppd}^{c,\text{train}}_{model,i}$	-336.81	-335.21	**-308.58**	-325.15	-329.90

for the models BE^c and HBE^c. We observe that the hierarchical model variation (even without the excitation) produces much tighter credibility intervals around the observed data than the model without information borrowing. However, the best performing models are ones with both information borrowing and self-excitation. Figure 7.6 illustrates the optimal performance of HBE^c over HB^c. In this plot, we see the 95% credibility intervals produced from model HBE^c for the higher count instances (7+) capture the observed aggregated count instances, whereas the HB^c credibility intervals fail to do so. We further see the aggregate log pointwise predictive density of $\sum_{i=1}^{17} \text{lppd}^{c,\text{train}}_{model,i}$ of table 7.3 provides more evidence that model HBE^c is the best fitting model, as this is maximised relative to the other hierarchical model variations.

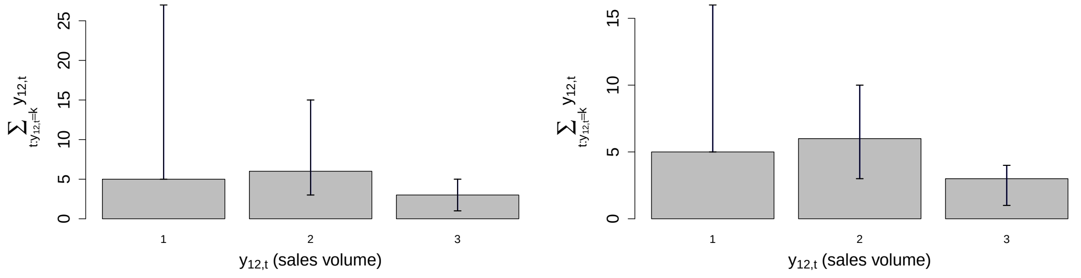


Figure 7.5: Histograms of $\sum_{t:y_{it}=k} y_{it}$ with corresponding 95% credible intervals of the posterior predictive distributions for models BE^c (left plot) and HBE^c (right plot) for product $i = 12$. The lower of 2.5% credible interval (the lower bound of the whisker bars) for $\sum_{t:y_{it}=1} \tilde{y}_{it}$ will at best be $\sum_{t:y_{it}=1} 1$. This is since the count distribution is lower bounded by 1.

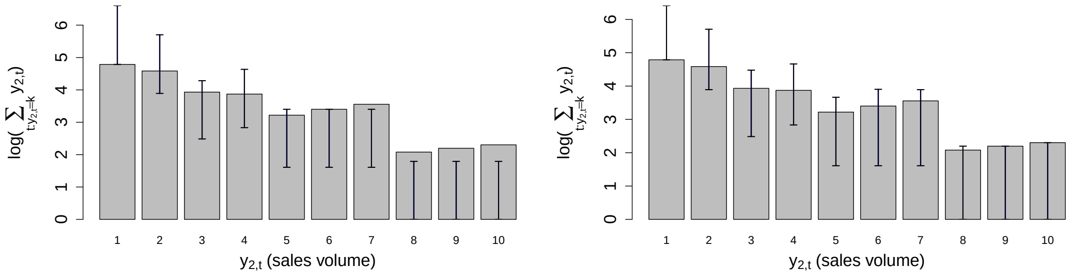


Figure 7.6: Histograms of $\sum_{t:y_{it}=k} y_{it}$ with corresponding 95% credible intervals of the log of the posterior predictive distributions of $\sum_{t:y_{it}=k} \tilde{y}_{it}$ (sale counts) for models HB^c (left plot) and HBE^c (right plot) for product $i = 2$.

7.3.5 Retail analytics discussion

The output of models outlined in sections 7.3.1 and 7.3.3 provides interesting interpretations from a retail analytics perspective. Firstly, we observe that covariate data $\boldsymbol{w}_{it}, \boldsymbol{x}_{it}$ as specified in 7.2.1 improves forecasting performance for the intermittent demand series of SMI products. This is indicated in both HB^c and HB^z - models with regression parameters and no form of excitation - outperforming their baseline counterparts on both the training and test sets. This importantly sheds light into the intermittent demand of SMI, in that it demonstrates covariate data such as prices and seasonality ought to be incorporated into training forecasting models as it seems predictions are improved from their inclusion.

Our findings further support the hypothesis that intermittent demand forecasting is improved when excitation dynamics are incorporated into models. This supports the findings of Snyder et al. [2012] and Chapados [2014] in which they establish that models incorporating the recent demand history outperform temporally static models. This is important because it ultimately allows retailers to circumvent over-stocking that typically results from inaccurate forecasting [Ghobbar and Friend, 2003]. However, our findings reveal some aspects of intermittent demand forecasting that go beyond the work of Snyder et al. [2012] and Snyder et al. [2012]. Namely, we establish that the temporal excitation exists even if you condition on the seasonal trends and pricing information of $\boldsymbol{w}_{it}, \boldsymbol{x}_{it}$. This suggests that temporal excitation is systematic and occurs beyond the variables traditionally utilised in forecasting models. We furthermore find that temporal excitation is manifested at lags greater than 1. Figure B.2 demonstrates that μ_i^z (the mean of excitation function of $g(\cdot \mid \mu, \tau)$) is approximately 2 across the majority of products, which implies that 2/3 of the probability mass of $g(\cdot \mid \mu, \tau)$ is placed on lags greater than or equal to 2. This is crucially important, as it indicates that a simple $AR(1)$, or equivalent model only taking the most recent observation into account is possibly not enough compared to the

Hawkes process that incorporates the entire history of events.

Thirdly, we also see strong support for the hypothesis that intermittent demand forecasting of SMI products benefits from hierarchical modelling. This is evident when we cross reference the parameter estimates with the posterior predictive distributions produced from each of the models. For example, Figure B.1 shows how a non-hierarchical model can suffer from not observing a range of sale counts on the training set which then translates to poor predictive performance on the test set. Figure B.3 further shows how information pooling to the extent that parameters are fixed across all the intermittent demand series can lead to misfit when heterogeneity appears to exist in the parameter estimates when compared to their hierarchical model counterparts. This has significant implications in the retail analytics context, as it suggests retailers should take into account the hierarchical structure exhibited in intermittent demand forecasting, as prediction is significantly improved when such structure is taken into account. Figure 7.7 are the forecasts of the intermittent demand of two slow-moving-inventory products using the combined zero and count models.

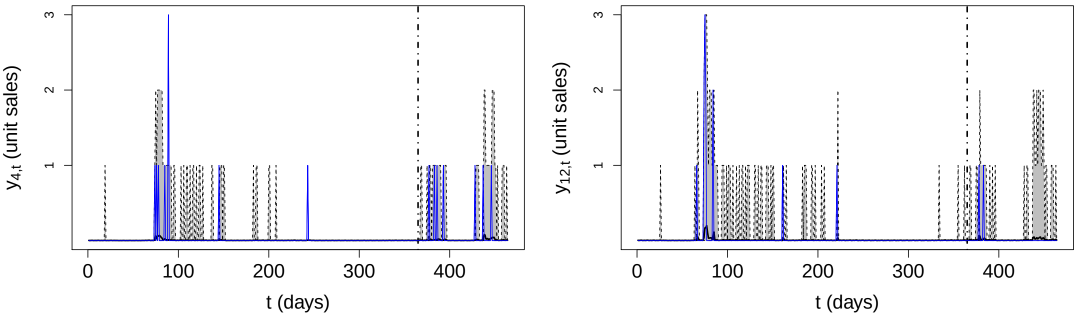


Figure 7.7: Plots of the combined models HBE^z and HBE^c for product $i = 4$ (left plot) and product $i = 12$ (right plot) over the entire training and test sets. The solid blue lines represent the sales of the respective touchscreen tablets and the black dashed lines are the 95% credible interval of the predictive posterior distribution of the sales counts. The dashed-dotted vertical black line at $t = 365$ represents the end and start of the training and test sets respectively.

7.4 Summary & future work

During this work we introduce a hierarchical model for the sales of the slow-moving-inventory category of touchscreen tablets across five large supermarkets in south London. We modelled the sales process as a Bayesian hierarchical zero-inflated hurdle regression model with self and cross-excitation components. The model specification is interpretable and allows a deeper understanding of the predictive role that covariates, self-excitation and cross-excitation play in the sales process of slow-moving-inventory and further provides a fully specified predictive distribution over this process. We demonstrated that the hierarchical structure as well as the self and cross-excitation additions offer a significant improvement in the predictive accuracy of this SMI sales process.

This model has important implications to the challenging issues that retail analytics face when developing SMI models. Firstly, it offers utility in terms of demand and profit forecasting that will allow retailers more accurate predictions of the sales distributions to aid with the issue of inventory management as well as price optimisation over short term horizons. It helps to explain the sources of variation and uncertainty that is exhibited in intermittent demand processes that previously was not well understood. The model also reveals a strong excitation component to these sales which could warrant further investigation as to what the potential underlying factors that could explain the observed excitation (e.g. marketing campaigns). We further note, that though there are many other approaches of specifying the cross-excitation relationship between pairwise products, our adopted approach of cross-excitation within brand provides an intuitive and computationally simple method of expressing the suspected temporal cross-correlation.

This work could be extended in many different directions. For example, a variable selection methodology could be introduced into the covariate predictors for each of the regression models. Our approach specified a priori the cross-excitation structure by defining an excitation event as a sale occurring within the same brand; it could be an interesting to assess whether the excitation structure could instead be inferred from the data.

Chapter 8

Conclusion

The field of retail analytics is a research area characterised by having a vast range of interesting challenges to choose from. These problems vary from the design of recommendation strategies awarding customer loyalty, to optimal inventory management, to understanding the full effect that marketing campaigns have on consumer demand, to name but a few. All of these problems, as a result of the growth of available data, offer a range of interesting routes that research could develop from.

During our work, we focused on two specific subclasses of problems within product clustering and demand forecasting. In particular, we tackled the issues of clustering products in terms of their cross-elasticity coefficients and forecasting the intermittent demand of SMI products. In both of these problem subclasses, we broadly achieved our objective. With respect to clustering products in terms of their sensitivities in sales, we developed a Bayesian nonparametric methodology that flexibly clusters products in terms of their cross-elasticities coefficients in a way that reflects the underlying structure and assumptions of the data. Similarly with forecasting the intermittent demand of SMI products, we developed a Bayesian hierarchical forecasting methodology that incorporates excitation dynamics and offers significant improvement over other forecasting benchmarks.

The contributions of this work subtly answer deeper questions about each of the problem areas of cross-elasticity coefficient analysis and intermittent demand forecasting. Namely, with interpreting products via their cross-elasticities, we were able address issues around designing a clustering strategy that accommodates the structure exhibited in cross-elasticity data and establishes a fundamental pattern in heterogeneity among these coefficients across different products. In terms of intermittent demand forecasting, we were able to develop a forecasting methodology able to incorporate excitation dynamics together with price and seasonal affects.

8.1 Future opportunities

This work helped to illustrate the possible applications that Bayesian nonparametric mixture modelling, excitation processes and hierarchical modelling has to the field of retail analytics. In particular, we further identify the following areas as interesting routes yet to be fully investigated:

1. **New product prediction:** As mentioned in section 6.1.1, one of the issues that retailers' face is the problem of forecasting the demand of a product that has limited demand history, or has no demand history as it has not yet been released for general consumption. Consequently, making accurate demand forecasts with such little information can be challenging, and this is particularly felt with products yet to be launched, which makes optimal inventory management hard. Such a scenario could be a fruitful application of data dependent mixture of regressions, in which the sales across various products is described as a mixture model whose mixtures are dependent on data related to various product features (for example a product's category, its initial selling price, brand information etc). Such a model would be capable at making forecasts of a product's demand prelaunch by allocating a newly launched product to an appropriate mixture of regression models depending on the product's relevant features.

2. **Social media data & temporal excitation:** One of the interesting aspects of intermittent demand our work established was the existence of a temporal excitation over and beyond what traditional covariates such as season and prices are able to explain. Such a finding opens up the question of whether this excitation can be described by the use of additional covariates that are typically not used in demand modelling. A possible avenue of investigation would be to better understand whether incorporating information of social media activity, for example the number @mentions a particular product or brand receives on twitter, or the amount of shares a photograph of a product around various social network sites has had, could help to explain movements of a product's sales that was previously thought of as excitation. Investigating a regression framework that studies the link between publicly available data and demand models to answer questions of whether temporal excitation still exists or whether social media data can be used to predict demand could be worthy of future research.

3. **Store-level hierarchical demand modelling:** A further avenue of research would be to investigate whether there are any temporal or spatial dependencies in demand across different stores and products. In the work developed during Chapter 7, the demand of a

particular touchscreen tablet was aggregated across five different south London stores. An alternative approach could be instead to model the demand of each product at a store level and incorporate a multivariate Hawkes process across the temporal domain that allows demand excitation in a product to occur as a result of the same products being purchased from a neighbouring store. Such an approach may have important implications in terms of a retailers' inventory management strategy, and one that lends itself to a hierarchical exposition in which information can be shared across the different regional stores as well as the multivariate Hawkes processes themselves.

We further believe our work has possible applications to fields other than retail analytics. Our Bayesian nonparametric clustering methodology of order statistics sequences could be applied to fields where data is inherently ordered and censored, such as is frequently the case in hazard rate modelling and software reliability analysis [Navarro and Shaked, 2006, Wilson and Samaniego, 2007]. Such fields could benefit from our clustering approach by identifying possible heterogeneity that could offer powerful interpretations within their relevant contexts. With respect to the proposed Bayesian hierarchical forecasting methodology, our approach could be applied to modelling scenarios where there exists hierarchical sparse count processes where future counts are dependent on historical counts. An example of such a setting is the case with healthcare resource planning, which is often characterised by an excess of zero counts, longitudinal data and temporal dynamics [Mihaylova, Briggs, O'hagan, and Thompson, 2011]. Such problem areas could greatly benefit from our forecasting approach by more accurately forecasting the demand for resources and services.

Appendix A

Appendix to Chapter 3

A.1 Appendix

A.1.1 Code for sampled DP measures and density mixtures of figure 3.1

The code below produces the generating process as described during section 3.5.2 along with the figure 3.1.

```
N = 50; alpha_vec = c(1, 5, 15, 40); base_mean = 0;
base_sd = 3; reponse_sd = 0.5;


for(j in 1:length(alpha_vec)){
  ### Samples from DP(alpha, N(base_mean, base_sd))
  theta_atoms = array(0,1)
  for(i in 1:N){
    if(i==1){
      theta_atoms[i] = rnorm(1,base_mean,base_sd)
    }
    else{
      probability_vector = array(0,i)
      probability_vector[1] = alpha_vec[j]/(i-1+alpha_vec[j])
      probability_vector[2:i] = 1/(i-1+alpha_vec[j])
      temp_allocation_vec = rmultinom(1, 1, probability_vector)
      if(temp_allocation_vec[1]){
        theta_atoms = c(theta_atoms, rnorm(1,base_mean,base_sd))
      }else{
        temp_allocation_vec = temp_allocation_vec[c(-1)]
        theta_atoms = c(theta_atoms, theta_atoms[temp_allocation_vec==1])
    }
  }
}
unique_atoms = unique(theta_atoms)
no_unique_atoms = length(unique(theta_atoms))
weights = array(0,no_unique_atoms)

for(i in 1:no_unique_atoms){
  weights[i] = sum(theta_atoms == unique_atoms[i])/N
}
x = seq(-10,10,0.01)
for(i in 1:no_unique_atoms){
  if(i==1){
    truth = weights[i]*dnorm(x,unique_atoms[i],reponse_sd)
  }else{
    truth = truth+weights[i]*dnorm(x,unique_atoms[i],reponse_sd)
  }
}
par(mar= c(5,5.0,4,5)+0.1)
plot(x, truth, type="l", xlab=""
, ylim=c(0,max(truth, table(theta_atoms)/N))
, main="", col="blue", cex.axis=1.5, cex=1.5, ylab="")
points(unique_atoms, weights)
axis(side = 4, cex.axis = 1.5)
mtext(side = 4, line = 3, expression(pi[i]), cex=1.5)
mtext(side = 1, line = 3, 'x', cex=1.5)
mtext(side = 2, line = 3, 'density', cex=1.5)
}
```

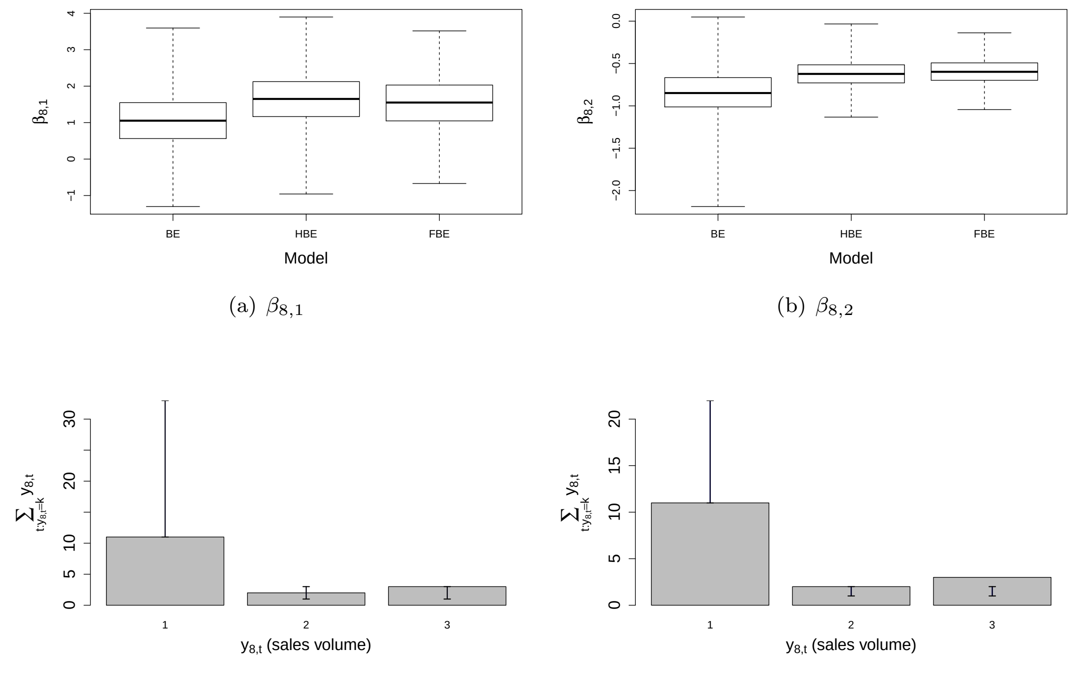
Appendix B

Appendix to Chapter 7

B.1 Appendix

B.1.1 Parameter analysis

We now provide an analysis of the parameters generated from model of 7.3.3 and 7.3.1.



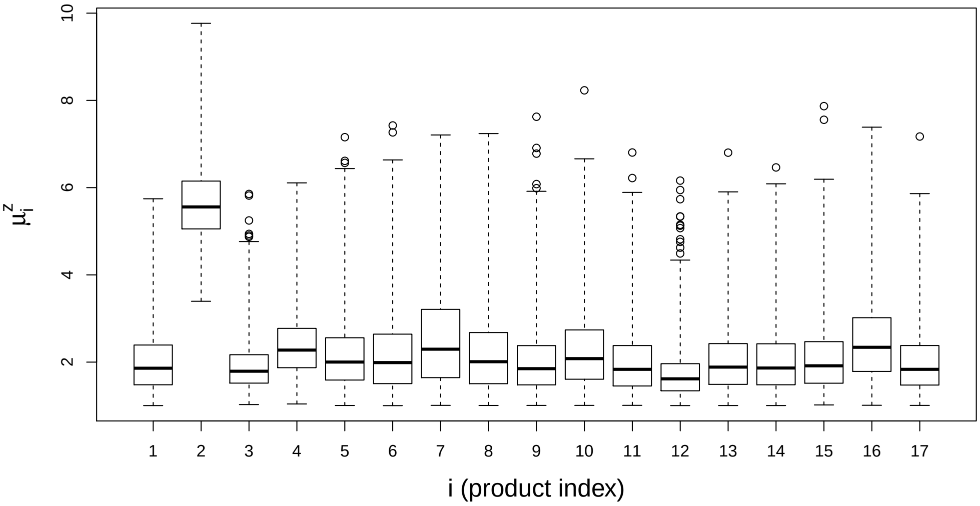
(a) $\beta_{8,1}$

(b) $\beta_{8,2}$

(c) Posterior predictive distributions of $\sum_{t:y_{8t}=k} y_{8t}$ from model BE^c.

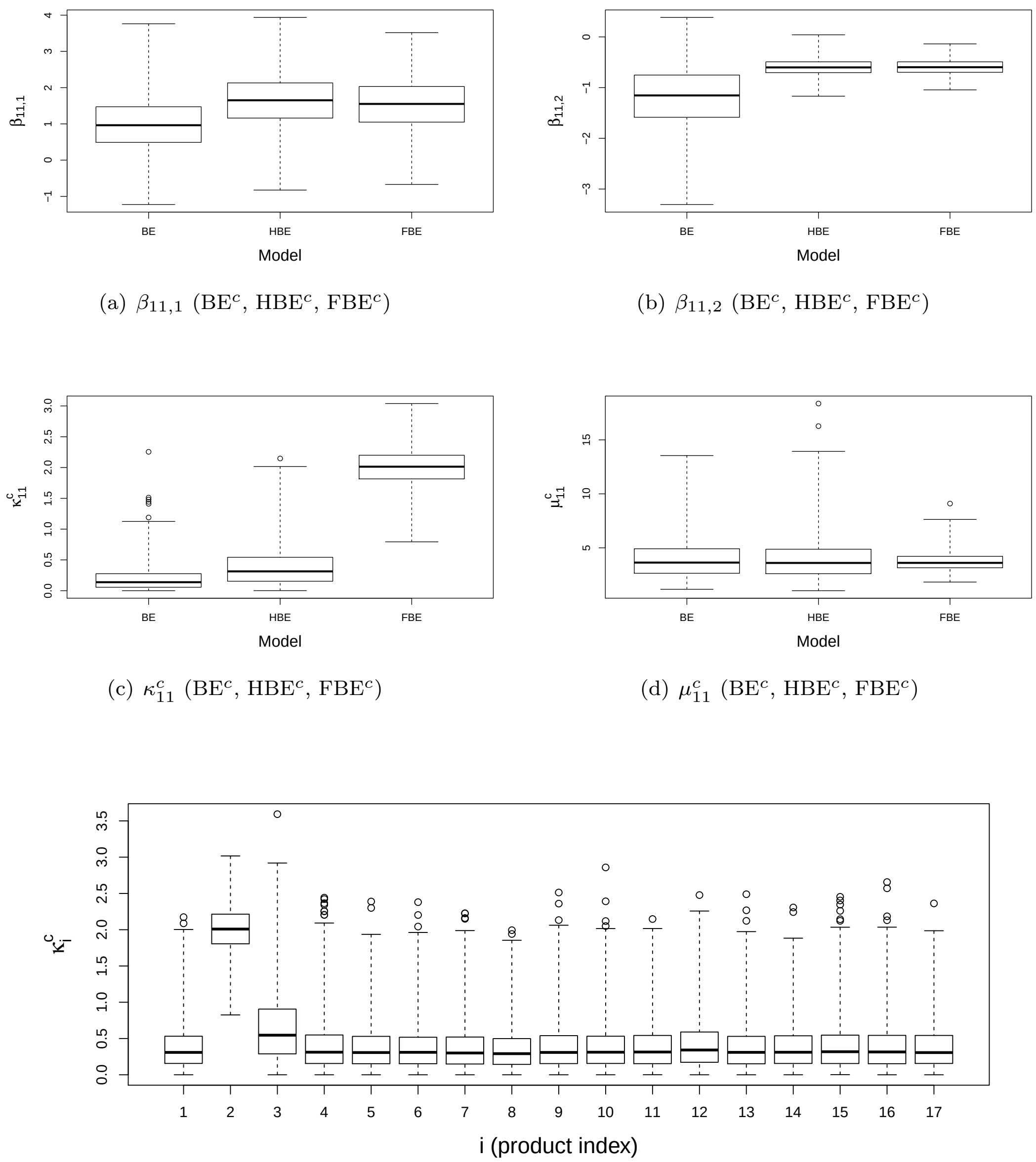
(d) Posterior predictive distributions of $\sum_{t:y_{8t}=k} y_{8t}$ from model HBE^c.

Figure B.1: Box plots of count process parameters β_{81}, β_{82} from models BE^c, HBE^c and FBE^c with histograms of $\sum_{t:y_{8t}=k} y_{8t}$ with 95% credible intervals of the posterior predictive distributions for models HBE^c and BE^c. These demonstrate model BE^c being penalised compared to models FBE^c and HBE^c that allow information pooling. The box plot of β_{81}, β_{82} estimates for model BE^c are different to those of models HBE^c and FBE^c, which leads to poor predictions for product $i = 8$ in model BE^c. These discrepancies arise from sale counts of product $i = 8$ being $y_{8t} < 3$ over the training set. However, over the test interval a sale count 3 is observed. This shows models HBE^c and FBE^c benefit from having 'seen' sale counts > 2 from other intermittent demand series that non-hierarchical models cannot.



(a) Box plots of μ_i^z for $i = 1, \ldots, 17$ for model HBE^z

Figure B.2: Box plots of μ_i^c across all products for model HBE^z. The μ_i^c estimates being greater than 2 indicates the temporal excitation exhibited in that data typically occurs at lags greater than 1.



(a) $\beta_{11,1}$ (BE^c, HBE^c, FBE^c)

(b) $\beta_{11,2}$ (BE^c, HBE^c, FBE^c)

(c) κ^c_{11} (BE^c, HBE^c, FBE^c)

(d) μ^c_{11} (BE^c, HBE^c, FBE^c)

(e) Box plots of $\beta_{i,11}$ for $i = 1, \ldots, 17$ for model HBE^c (HBE).

Figure B.3: Box plots of various count process parameters. This figure shows that parameters across the models HBE^c and FBE^c to be almost identical except for parameter κ^c_{11}. The root of this discrepancy derives from the significant excitation exhibited in product $i = 2$, as demonstrated from the HBE^c estimate of κ^c_2 in plot B.3(e). This creates a skew in the shared parameter of κ^c in model FBE^c, a skew that does not exist in the κ^c_{11} of HBE^c. This lack of heterogeneity in model FBE^c reduces the predictive accuracy on the test set compared to hierarchical equivalent model of HBE^c

B.1.2 Stan code for HBEC^z

Below is the hierarchical exposition of the HBEC^z as implemented by STAN.

```
data{

int <lower=0> no_models;
int <lower=0> N;
int <lower=0> zero_reg_dim;
int <lower=0> HP_zero_dim;

matrix[N, zero_reg_dim] X_zero[no_models];
int <lower=0> events_times[no_models, N];
int <lower=0> no_events_lesst[no_models, N];
int <lower=0> y[no_models, N];

int <lower=0> N_test;

matrix[N_test, zero_reg_dim] X_zero_test[no_models];
int <lower=0> events_times_test[no_models, N_test];
int <lower=0> no_events_lesst_test[no_models, N_test];
int <lower=0> y_test[no_models, N_test];

real <lower=0> excitation_hyperparameters[HP_zero_dim, 4];
real regression_hyperparameters[zero_reg_dim, 4];

// CROSS data
real <lower=0> excitation_hyperparameters_CROSS[HP_zero_dim, 4];

int <lower=0> events_times_CROSS[no_models, N];
int <lower=0> no_events_lesst_CROSS[no_models, N];

int <lower=0> events_times_test_CROSS[no_models, N_test];
int <lower=0> no_events_lesst_test_CROSS[no_models, N_test];
}
parameters {
vector[zero_reg_dim] beta_zero[no_models];
real <lower=0> kappa_zero[no_models];
real <lower=1> mu_zero[no_models];
real <lower=0> tau_zero[no_models];
// CROSS
```

```
real <lower=0> kappa_zero_CROSS[no_models];
real <lower=1> mu_zero_CROSS[no_models];
real <lower=0> tau_zero_CROSS[no_models];
real  beta_normal_mu_priors[zero_reg_dim];
real <lower=0> excitation_alphapriors[HP_zero_dim];

// CROSS
real <lower=0> excitation_alphapriors_CROSS[HP_zero_dim];
}
model{
real ps[no_models,N];
real HP_zero[no_models,N];
real HP_zero_CROSS[no_models,N];

for(k in 1:no_models){
beta_normal_mu_priors[1] ~ normal(regression_hyperparameters[k,1]
                                ,regression_hyperparameters[k,2]);
}
excitation_alphapriors[1] ~ gamma(excitation_hyperparameters[1,1]
                                ,excitation_hyperparameters[1,2]);
excitation_alphapriors[2] ~ gamma(excitation_hyperparameters[2,1]
                                ,excitation_hyperparameters[2,2]);
excitation_alphapriors[3] ~ gamma(excitation_hyperparameters[3,1]
                                ,excitation_hyperparameters[3,2]);
// CROSS
excitation_alphapriors_CROSS[1] ~ gamma(excitation_hyperparameters_CROSS[1,1]
                                ,excitation_hyperparameters_CROSS[1,2]);
excitation_alphapriors_CROSS[2] ~ gamma(excitation_hyperparameters_CROSS[2,1]
                                ,excitation_hyperparameters_CROSS[2,2]);
excitation_alphapriors_CROSS[3] ~ gamma(excitation_hyperparameters_CROSS[3,1]
                                ,excitation_hyperparameters_CROSS[3,2]);

for(m in 1:no_models){
for (k in 1:20) {
```

```
beta_zero[m,k] ~ normal(beta_normal_mu_priors[k],0.05);
}
kappa_zero[m] ~ gamma(excitation_alphapriors[1],1);
(mu_zero[m]-1) ~ gamma(excitation_alphapriors[2],2);
tau_zero[m] ~ gamma(excitation_alphapriors[3],2.5);
kappa_zero_CROSS[m] ~ gamma(excitation_alphapriors_CROSS[1],8);
(mu_zero_CROSS[m]-1) ~ gamma(excitation_alphapriors_CROSS[2],2);
tau_zero_CROSS[m] ~ gamma(excitation_alphapriors_CROSS[3],2.5);

for (i in 1:N) {
HP_zero[m,i] <- 0;
HP_zero_CROSS[m,i] <- 0;

if (no_events_lesst[m,i]>0) {
for (j in 1:no_events_lesst[m,i]) {
HP_zero[m,i] <- HP_zero[m,i]
+kappa_zero[m]*exp(neg_binomial_2_lpmf(i-events_times[m,j]-1 |
(mu_zero[m]-1), tau_zero[m]));}
}
if (no_events_lesst_CROSS[m,i]>0) {
for (j in 1:no_events_lesst_CROSS[m,i]) {
HP_zero_CROSS[m,i] <- HP_zero_CROSS[m,i]
+kappa_zero_CROSS[m]*exp(neg_binomial_2_lpmf(i-events_times_CROSS[m,j]-1 |
(mu_zero_CROSS[m]-1), tau_zero_CROSS[m]));}
}
ps[m,i] <- 1/(1+exp(-X_zero[m,i]*beta_zero[m]
                              -HP_zero[m,i]-HP_zero_CROSS[m,i]));
y[m,i] ~ bernoulli(ps[m,i]);
}}}
```

B.1.3 Prior formulation of zero processes

Table B.1 specifies the prior structure of models Base_1^z, HB^z, BE^z, HBE^z, FBE^z, BEC^z, HBEC^z and FBEC^z.

Table B.1: Prior formulation of models Base$_1^z$, HB^z, BE^z, HBE^z, FBE^z, BEC^z, HBEC^z and FBEC^z.

Parameter	Base$_1^z$	HB^z	BE^z	HBE^z	FBE^z	BEC^z	HBEC^z	FBEC^z
$\varphi_i \sim$	$\mathrm{N}(-3, 3)$							
$\theta_{i1} \sim$		$\mathrm{N}(\mu_1^z, 0.05)$	$\mathrm{N}(-3, 0.75)$	$\mathrm{N}(\mu_1^z, 0.05)$		$\mathrm{N}(-3, 0.75)$	$\mathrm{N}(\mu_1^z, 0.05)$	
$\theta_{i2} \sim$		$\mathrm{N}(\mu_2^z, 0.05)$	$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_2^z, 0.05)$		$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_2^z, 0.05)$	
$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	
$\theta_{i20} \sim$		$\mathrm{N}(\mu_{20}^z, 0.05)$	$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_{20}^z, 0.05)$		$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_{20}^z, 0.05)$	
$\theta_{i1} = \theta_1 \sim$					$\mathrm{N}(-3, 0.75)$			$\mathrm{N}(-3, 0.75)$
$\theta_{i2} = \theta_2 \sim$					$\mathrm{N}(0, 0.75)$			$\mathrm{N}(0, 0.75)$
					$\vdots$			$\vdots$
$\theta_{i20} = \theta_{20} \sim$					$\mathrm{N}(0, 0.75)$			$\mathrm{N}(0, 0.75)$
$\gamma_{i1}^z \sim$			$\mathrm{G}(5, 1)$	$\mathrm{G}(\eta_1^z, 1)$		$\mathrm{G}(5, 1)$	$\mathrm{G}(\eta_1^z, 1)$	
$\gamma_{i2}^z \sim$			$1 + \mathrm{G}(1, 2)$	$1 + \mathrm{G}(\eta_2^z, 2)$		$1 + \mathrm{G}(1, 2)$	$1 + \mathrm{G}(\eta_2^z, 2)$	
$\gamma_{i3}^z \sim$			$\mathrm{G}(10, 2.5)$	$\mathrm{G}(\eta_3^z, 2.5)$		$\mathrm{G}(10, 2.5)$	$\mathrm{G}(\eta_3^z, 2.5)$	
$\gamma_{i1}^z = \gamma_1^z \sim$					$\mathrm{G}(5, 1)$			$\mathrm{G}(5, 1)$
$\gamma_{i2}^z = \gamma_2^z \sim$					$1 + \mathrm{G}(1, 2)$			$1 + \mathrm{G}(1, 2)$
$\gamma_{i3}^z = \gamma_3^z \sim$					$\mathrm{G}(10, 2.5)$			$\mathrm{G}(10, 2.5)$
$\widetilde{\gamma}_{i1}^z \sim$						$\mathrm{G}(2, 8)$	$\mathrm{G}(\widetilde{\eta}_1^z, 8)$	
$\widetilde{\gamma}_{i2}^z \sim$						$1 + \mathrm{G}(1, 2)$	$1 + \mathrm{G}(\widetilde{\eta}_2^z, 2)$	
$\widetilde{\gamma}_{i3}^z \sim$						$\mathrm{G}(10, 2.5)$	$\mathrm{G}(\widetilde{\eta}_3^z, 2.5)$	
$\widetilde{\gamma}_{i1}^z = \widetilde{\gamma}_1^z \sim$								$\mathrm{G}(2, 8)$
$\widetilde{\gamma}_{i2}^z = \widetilde{\gamma}_2^z \sim$								$1 + \mathrm{G}(1, 2)$
$\widetilde{\gamma}_{i3}^z = \widetilde{\gamma}_3^z \sim$								$\mathrm{G}(10, 2.5)$
$\rho_1^z \sim$		$\mathrm{N}(-3, 0.75)$		$\mathrm{N}(-3, 0.75)$			$\mathrm{N}(-3, 0.75)$	
$\rho_2^z \sim$		$\mathrm{N}(0, 0.75)$		$\mathrm{N}(0, 0.75)$			$\mathrm{N}(0, 0.75)$	
$\vdots$		$\vdots$		$\vdots$			$\vdots$	
$\rho_{20}^z \sim$		$\mathrm{N}(0, 0.75)$		$\mathrm{N}(0, 0.75)$			$\mathrm{N}(0, 0.75)$	
$\eta_1^z \sim$				$\mathrm{G}(50, 10)$			$\mathrm{G}(50, 10)$	
$\eta_2^z \sim$				$\mathrm{G}(10, 10)$			$\mathrm{G}(10, 10)$	
$\eta_3^z \sim$				$\mathrm{G}(500, 50)$			$\mathrm{G}(500, 50)$	
$\widetilde{\eta}_1^z \sim$							$\mathrm{G}(30, 15)$	
$\widetilde{\eta}_2^z \sim$							$\mathrm{G}(10, 10)$	
$\widetilde{\eta}_3^z \sim$							$\mathrm{G}(500, 50)$	

B.1.4 Stan code for HBE^c

Below is the hierarchical exposition of the HBE^c as implemented by STAN.

```
functions {
  real neg_binomial_den(int t, real mu, real size) {
    real binomial_term;
    binomial_term <- exp(binomial_coefficient_log(t+size -1, t));

    return binomial_term*pow(mu/(mu+size), t)*pow(size/(mu+size), size);
  }

  real poisson_den(int t, real lambda_par) {
    return pow(lambda_par, t)*exp(-lambda_par)/tgamma(t+1);
  }
}
data{
  int <lower=0> no_models;
  int <lower=0> N;
  //int <lower=0> no_events[no_models];
  int <lower=0> count_reg_dim;
  int <lower=0> HP_count_dim;

  matrix[N, count_reg_dim] X_count[no_models];
  int <lower=0> events_times[no_models, N]; // event occurrence index
  int <lower=0> no_events_lesst[no_models, N];
  int <lower=0> y_count[no_models, N];

  int <lower=0> N_test;

  matrix[N_test, count_reg_dim] X_count_test[no_models];
  int <lower=0> events_times_test[no_models, N_test];
  int <lower=0> no_events_lesst_test[no_models, N_test];
  int <lower=0> y_count_test[no_models, N_test];

  real <lower=0> excitation_hyperparameters[HP_count_dim,4];
  real regression_hyperparameters[count_reg_dim,4];
}
parameters {
  vector[count_reg_dim] beta_count[no_models];
  //real <lower=0, upper=10> kappa_count[no_models];
  real <lower=0> kappa_count[no_models];
  real <lower=1> mu_count[no_models];
  real <lower=0> tau_count[no_models];

  real beta_mupriors;
  real <lower=0> beta_sigmapriors;

  real beta_price_alpha_priors;
  real <lower=0> beta_price_beta_priors;

  real <lower=0> excitation_alphapriors[HP_count_dim];
  real <lower=0> excitation_betapriors[HP_count_dim];
  }
model{
  real HP_count[no_models, N];
  real lambda[no_models, N];

  beta_mupriors ~ normal(regression_hyperparameters[1,1]
                        ,regression_hyperparameters[1,2]);
  beta_sigmapriors ~ gamma(regression_hyperparameters[1,3]
                          ,regression_hyperparameters[1,4]);

  beta_price_alpha_priors ~ normal(regression_hyperparameters[2,1]
                                  ,regression_hyperparameters[2,2]);
  beta_price_beta_priors ~ gamma(regression_hyperparameters[2,3]
                                ,regression_hyperparameters[2,4]);

  excitation_alphapriors[1] ~ gamma(excitation_hyperparameters[1,1]
                                   ,excitation_hyperparameters[1,2]);
  excitation_betapriors[1] ~ gamma(excitation_hyperparameters[1,3]
                                  ,excitation_hyperparameters[1,4]);

  excitation_alphapriors[2] ~ gamma(excitation_hyperparameters[2,1]
                                   ,excitation_hyperparameters[2,2]);
  excitation_betapriors[2] ~ gamma(excitation_hyperparameters[2,3]
                                  ,excitation_hyperparameters[2,4]);

  excitation_alphapriors[3] ~ gamma(excitation_hyperparameters[3,1]
                                   ,excitation_hyperparameters[3,2]);
  excitation_betapriors[3] ~ gamma(excitation_hyperparameters[3,3]
                                   ,excitation_hyperparameters[3,4]);

  for(m in 1:no_models){
    beta_count[m,1] ~ normal(beta_mupriors
                            ,beta_sigmapriors);
    beta_count[m,2] ~ normal(beta_price_alpha_priors
                            ,beta_price_beta_priors);

    kappa_count[m] ~ gamma(excitation_alphapriors[1]
                          ,excitation_betapriors[1]);
    (mu_count[m]-1) ~ gamma(excitation_alphapriors[2]
                           ,excitation_betapriors[2]);
    tau_count[m] ~ gamma(excitation_alphapriors[3]
                        ,excitation_betapriors[3]);

    for (i in 1:N) {
      HP_count[m,i] <- 0;

      if(y_count[m,i]>0){
        if (no_events_lesst[m,i]>0) {
          for (j in 1:no_events_lesst[m,i]) {
            HP_count[m,i] <- HP_count[m,i]
+kappa_count[m]*neg_binomial_den(i-1-events_times[m,j]
                                 ,(mu_count[m]-1), tau_count[m]);
          }
        }
        lambda[m,i] <- exp(X_count[m,i]*beta_count[m]+HP_count[m,i]);
        target += neg_binomial_2_log(y_count[m,i] , lambda[m,i], 1)
                -neg_binomial_2_ccdf_log(0 , lambda[m,i], 1);
      }
    }
  }
}
```

Table B.2: Prior formulation of models Base_1^c, HB^c, BE^c, HBE^c and FBE^c.

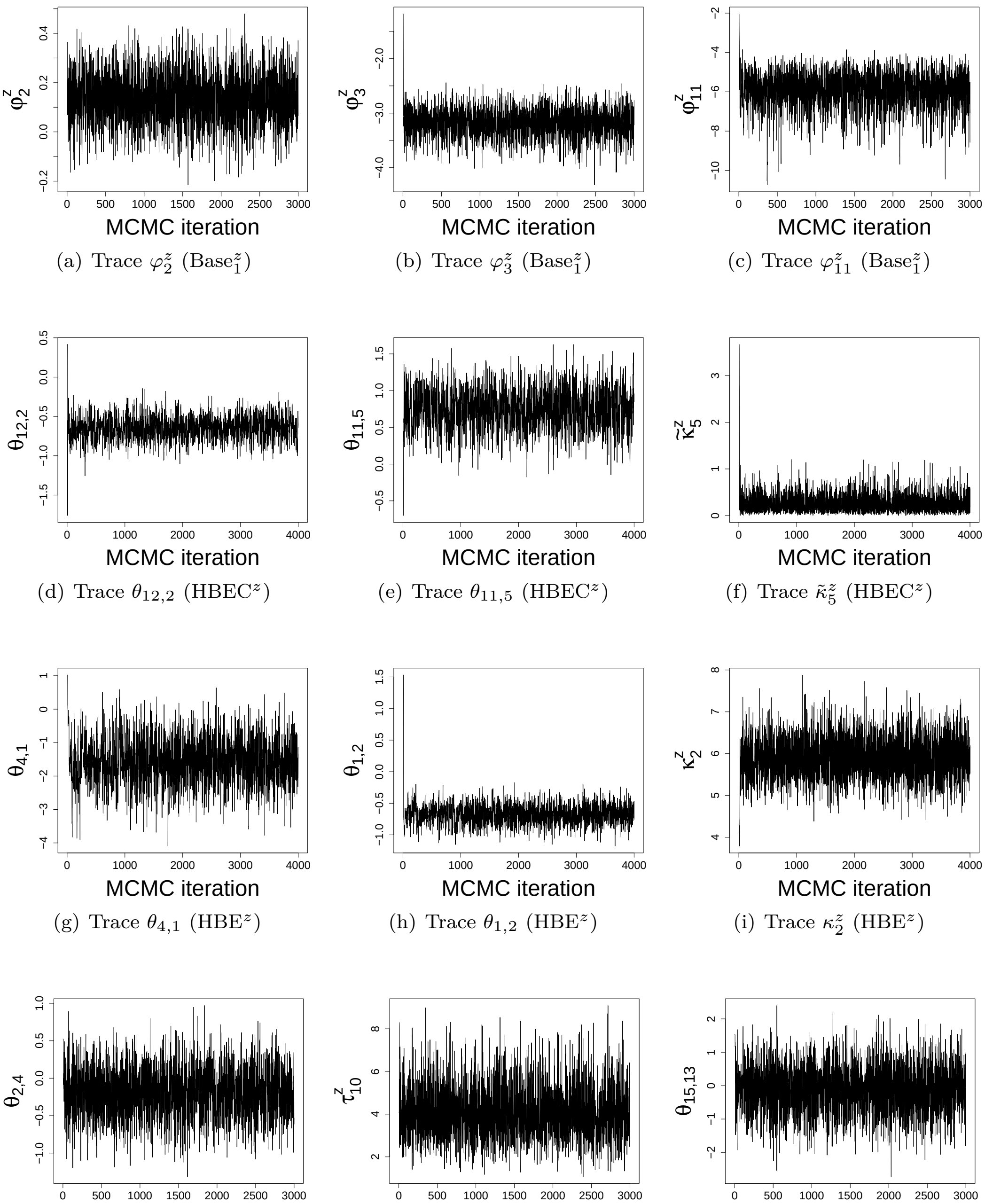
Parameter	Base_1^c	HB^c	BE^c	HBE^c	FBE^c
$\varphi_i^c \sim$	$\text{N}(-4,4)$				
$\beta_{i1} \sim$		$\text{N}(\mu_1^c,0.05)$	$\text{N}(1,0.75)$	$\text{N}(\mu_1^c,0.05)$	
$\beta_{i2} \sim$		$\text{N}(\mu_2^c,0.05)$	$\text{N}(-1,0.75)$	$\text{N}(\mu_2^c,0.05)$	
$\beta_{i1} = \beta_1 \sim$					$\text{N}(1,0.75)$
$\beta_{i2} = \beta_2 \sim$					$\text{N}(-1,0.75)$
$\gamma_{i1}^c \sim$			$\text{G}(1,5)$	$\text{G}(\eta_1^c,5)$	
$\gamma_{i2}^c \sim$			$1+\text{G}(3,1)$	$1+\text{G}(\eta_2^c,1)$	
$\gamma_{i3}^c \sim$			$\text{G}(4,1)$	$\text{G}(\eta_3^c,1)$	
$\gamma_{i1}^c = \gamma_1^c \sim$					$\text{G}(1,5)$
$\gamma_{i2}^c = \gamma_2^c \sim$					$1+\text{G}(3,1)$
$\gamma_{i3}^c = \gamma_3^c \sim$					$\text{G}(4,1)$
$\rho_1^c \sim$		$\text{N}(1,0.75)$		$\text{N}(1,0.75)$	
$\rho_2^c \sim$		$\text{N}(-1,0.75)$		$\text{N}(-1,0.75)$	
$\eta_1^c \sim$				$\text{G}(5,5)$	
$\eta_2^c \sim$				$\text{G}(15,5)$	
$\eta_3^c \sim$				$\text{G}(40,10)$	

B.1.5 Prior formulation of count processes

Table B.1.5 specifies the prior structure of models Base_1^c, HB^c, BE^c, HBE^c and FBE^c.

B.1.6 Zero process MCMC output

Parameter inference of zero process models are performed by HMC algorithm. For models HB^z and HBE^z we take 3000 samples with 1000 burn-in respectively and 2000 samples with 1000 burn-in for models FBEC^z, BE^z and FBE^z. For model HBEC^z we take 2500 samples and 1500 burnin and we take 4000 samples and 6000 burn-in for BEC^z. For model Base$^z_1$ we take 2000 samples and 1000 burnin. Figure B.4 provides some typical trace plots of fitted models.



(a) Trace φ^z_2 (Base$^z_1$)

(b) Trace φ^z_3 (Base$^z_1$)

(c) Trace φ^z_{11} (Base$^z_1$)

(d) Trace $\theta_{12,2}$ (HBEC^z)

(e) Trace $\theta_{11,5}$ (HBEC^z)

(f) Trace $\tilde{\kappa}^z_5$ (HBEC^z)

(g) Trace $\theta_{4,1}$ (HBE^z)

(h) Trace $\theta_{1,2}$ (HBE^z)

(i) Trace κ^z_2 (HBE^z)

(j) Trace $\theta_{2,4}$ (BE^z)

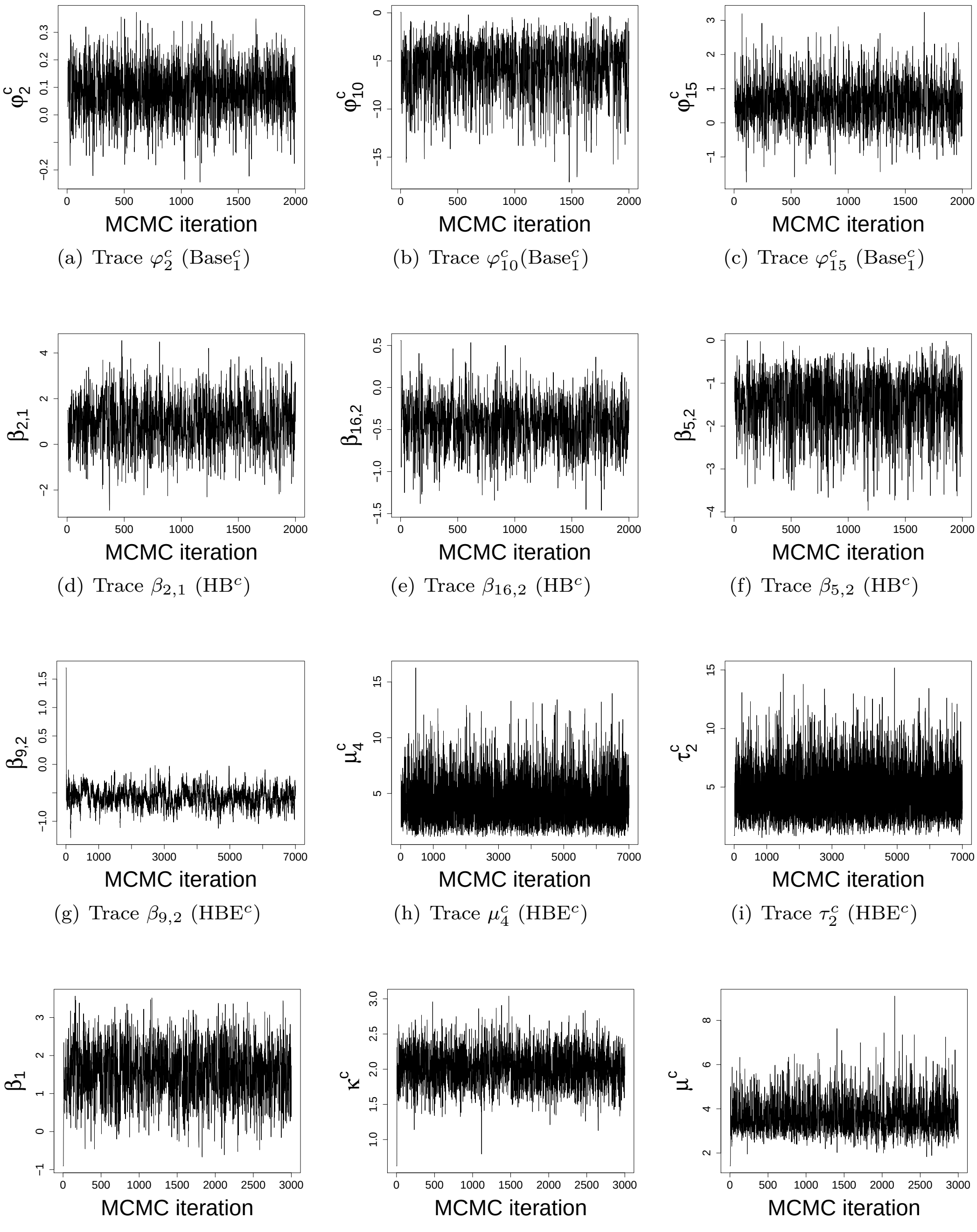
(k) Trace τ^z_{10} (BE^z)

(l) Trace $\theta_{5,13}$ (BE^z)

Figure B.4: Trace plots of selected zero process parameters.

B.1.7 Count process MCMC output

Parameter inference is performed by the HMC sampling algorithm. For models BE^c and FBE^c we take 2000 samples with 1000 burn-in respectively. For model HB^c we take 3000 samples and 2000 burnin and for model HBE^c we take 5000 samples and 2000 burn-in. For model Base$_1^c$ we take 1000 samples and 1000 burnin. Figure B.5 provides typical trace plots of fitted models.



(a) Trace φ_2^c (Base$_1^c$) (b) Trace φ_{10}^c(Base$_1^c$) (c) Trace φ_{15}^c (Base$_1^c$)

(d) Trace $\beta_{2,1}$ (HB^c) (e) Trace $\beta_{16,2}$ (HB^c) (f) Trace $\beta_{5,2}$ (HB^c)

(g) Trace $\beta_{9,2}$ (HBE^c) (h) Trace μ_4^c (HBE^c) (i) Trace τ_2^c (HBE^c)

(j) Trace β_1 (FBE^c) (k) Trace κ^c (FBE^c) (l) Trace μ^c (FBE^c)

Figure B.5: Trace plots of selected count process parameters.

‘

Bibliography

Office of national statistics. *Retail sales index*, Main reference tables, Data series AGG1 J448.

Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view.* Springer Science &amp; Business Media, 2008.

Alp Eren Akcay. *Statistical Estimation Problems in Inventory Management.* PhD thesis, Tepper School of Business, 2013.

Imad Al Ajarmeh, James Yu, and Mohamed Amezziane. Framework of applying a non-homogeneous poisson process to model voip traffic on tandem networks. In *Proc. 10th WSEAS Int. Conf. Applied Informatics and Communications AIC 2010*, pages 164–169, 2010.

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

Özden Gür Ali and Kübra Yaman. Selecting rows and columns for training support vector regression models with large retail datasets. *European Journal of Operational Research*, 226 (3):471–480, 2013.

Tatiana Andreyeva, Michael W Long, and Kelly D Brownell. The impact of food prices on consumption: a systematic review of research on the price elasticity of demand for food. *American journal of public health*, 100(2):216–222, 2010.

Charles E Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, pages 1152–1174, 1974.

Barry C Arnold, Narayanaswamy Balakrishnan, and Haikady Navada Nagaraja. *A first course in order statistics*, volume 54. Siam, 1992.

Pradip Kumar Bala. Improving inventory performance with clustering based demand forecasts. *Journal of Modelling in Management*, 7(1):23–37, 2012.

Yukun Bao, Hua Zou, and Zhitao Liu. Forecasting intermittent demand by fuzzy support vector machines. In *IEA/AIE*, pages 1080–1089. Springer, 2006.

May Barghout, Bev Littlewood, and Abdalla Abdel-Ghaly. A non-parametric order statistics software reliability model. *Software Testing Verification and Reliability*, 8(3):113–132, 1998.

Frank M Bass, Norris Bruce, Sumit Majumdar, and BPS Murthi. Wearout effects of different advertising themes: A dynamic Bayesian model of the advertising-sales relationship. *Marketing Science*, 26(2):179–195, 2007.

Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjen, and Michael Teucke. A survey on retail sales forecasting and prediction in fashion markets. *Systems Science & Control Engineering*, 3(1):154–161, 2015.

James O Berger and Luis R Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

E Beutner and U Kamps. Order restricted statistical inference for scale parameters based on sequential order statistics. *Journal of Statistical Planning and Inference*, 139(9):2963–2969, 2009.

Ram Bezawada, Subramanian Balachander, PK Kannan, and Venkatesh Shankar. Cross-category effects of aisle and display placements: a spatial modeling approach and insights. *Journal of Marketing*, 73(3):99–117, 2009.

David A Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 1978.

David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

David M Blei, Michael I Jordan, et al. Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

Jean-Philippe Boucher, Michel Denuit, and Montserrat Guillén. Modelling of insurance claim count with hurdle distribution for panel data. *Advances in Mathematical and Statistical Modeling: Statistics for Industry and Technology*, pages 45–60, 2008.

Clive G Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

Raymond R Burke and Alex Leykin. Identifying the drivers of shopper attention, engagement, and purchase. In *Shopper Marketing and the Role of In-Store Marketing*, pages 147–187. Emerald Group Publishing Limited, 2014.

Anne Buu, Runze Li, Xianming Tan, and Robert A Zucker. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in medicine*, 31(29):4074–4086, 2012.

Bo Cai, Renate Meyer, and François Perron. Metropolis–Hastings algorithms with adaptive proposals. *Statistics and Computing*, 18(4):421–433, 2008.

Juan Juan Cai. Lecture 1-introduction and the empirical cdf. 2014.

Felipe Caro and Jérémie Gallien. Clearance pricing optimization for a fast-fashion retailer. *Operations Research*, 60(6):1404–1422, 2012.

François Caron and Yee W Teh. Bayesian nonparametric models for ranked data. In *Advances in Neural Information Processing Systems*, pages 1520–1528, 2012.

Nicolas Chapados. Effective Bayesian modeling of groups of related count time series. *arXiv preprint arXiv:1405.3738*, 2014.

Benaissa Chidmi and Rigoberto A Lopez. Brand-supermarket demand for breakfast cereals and retail competition. *American Journal of Agricultural Economics*, 89(2):324–337, 2007.

Lee G Cooper and Masako Nakanishi. *Market-share analysis: Evaluating competitive marketing effectiveness*, volume 1. Springer Science &amp; Business Media, 1989.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

J Do Croston. Forecasting and stock control for intermittent demands. *Operational research quarterly*, pages 289–303, 1972.

José Da Fonseca and Riadh Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

Anirban DasGupta. Poisson processes and applications. In *Probability for Statistics and Machine Learning*, pages 437–462. Springer, 2011.

Jan G De Gooijer and Rob J Hyndman. 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473, 2006.

Maria De Iorio, Peter Müller, Gary L Rosner, and Steven N MacEachern. An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99(465): 205–215, 2004.

Maria De Iorio, Wesley O Johnson, Peter Müller, and Gary L Rosner. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771, 2009.

Rene A de Wijk, Anna J Maaskant, Ilse A Polet, Nancy TE Holthuysen, Ellen van Kleef, and Monique H Vingerhoeds. An in-store experiment on the effect of accessibility on sales of wholegrain and white bread in supermarkets. *PloS one*, 11(3):e0151915, 2016.

Angus Deaton and John Muellbauer. An almost ideal demand system. *The American economic review*, 70(3):312–326, 1980.

Fei Deng. Measuring the substitution effects of sales promotions in supermarkets: An analysis based on a dynamic model of differentiated products. *Job Market Paper*, 2005.

Alexandre Dolgui and Maksim Pashkevich. Extended beta-binomial model for demand forecasting of multiple slow-moving inventory items. *International Journal of Systems Science*, 39(7): 713–726, 2008.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Bob Eagle and Tim Ambler. The influence of advertising on the demand for chocolate confectionery. *International Journal of Advertising*, 21(4):437–454, 2002.

Tülin Erdem, Michael P Keane, and Baohong Sun. The impact of advertising on consumer price sensitivity in experience goods markets. *Quantitative Marketing and Economics*, 6(2): 139–176, 2008.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Mark Ferguson and Michael E Ketzenberg. Information sharing to improve retail product freshness of perishables. *Production and Operations Management*, 15(1):57, 2006.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 2015.

Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *arXiv preprint arXiv:1704.04110*, 2017.

Dennis Fok, Philip Hans Franses, and Richard Paap. Econometric analysis of the market share attraction model. In *Advances in Econometrics*, pages 223–256. Emerald Group Publishing Limited, 2002.

Arno Fritsch. mcclust: Process an mcmc sample of clusterings. *R package version*, 1, 2012.

Arno Fritsch, Katja Ickstadt, et al. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian analysis*, 4(2):367–391, 2009.

Everette S Gardner. Exponential smoothing: The state of the art—part ii. *International journal of forecasting*, 22(4):637–666, 2006.

Andrew Gelman. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA, 1991.

Charles J Geyer. Markov chain Monte Carlo lecture notes. *Course notes, Spring Quarter*, 1998.

Adel A Ghobbar and Chris H Friend. Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. *Computers &amp; Operations Research*, 30(14):2097–2114, 2003.

Joydeep Ghosh and Alexander Strehl. Clustering and visualization of retail market baskets. *Advanced Techniques in Knowledge Discovery and Data Mining*, pages 75–102, 2005.

Kaushik Ghosh and Ram C Tiwari. Nonparametric and semiparametric Bayesian reliability analysis. *Encyclopedia of Statistics in Quality and Reliability*, 2007.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice.* CRC press, 1995.

Leslie G Godfrey. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica: Journal of the Econometric Society*, pages 1293–1301, 1978.

Brett R Gordon, Avi Goldfarb, and Yang Li. Does price elasticity vary with economic growth? a cross-category analysis. *Journal of Marketing Research*, 50(1):4–23, 2013.

William H Greene. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. 1994.

Carlos M Guerrero-López, Mishel Unar-Munguía, and M Arantxa Colchero. Price elasticity of the demand for soft drinks, other sugar-sweetened beverages and energy dense food in chile. *BMC public health*, 17(1):180, 2017.

Rafael S Gutierrez, Adriano O Solis, and Somnath Mukhopadhyay. Lumpy demand forecasting using neural networks. *International Journal of Production Economics*, 111(2):409–420, 2008.

Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

GJ Hahn and A Leucht. Managing inventory systems of slow-moving items. *International Journal of Production Economics*, 170:543–550, 2015.

Daniel B Hall and Jing Shen. Robust estimation for zero-inflated poisson regression. *Scandinavian Journal of Statistics*, 37(2):237–252, 2010.

Timothy E Hanson et al. Modeling censored lifetime data using a mixture of Gammas baseline. *Bayesian Analysis*, 1(3):575–594, 2006.

David I Hastie, Silvia Liverani, and Sylvia Richardson. Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and computing*, 25(5):1023–1037, 2015.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Philip Heidelberger and Peter D Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245, 1981.

Stephen J Hoch, Byung-Do Kim, Alan L Montgomery, and Peter E Rossi. Determinants of store-level price elasticity. *Journal of marketing Research*, pages 17–29, 1995.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Mei-Chen Hu, Martina Pavlicova, and Edward V Nunes. Zero-inflated and hurdle models of count data with extra zeros: examples from an hiv-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5):367–375, 2011.

Zhongsheng Hua and Bin Zhang. A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts. *Applied Mathematics and Computation*, 181(2):1035–1048, 2006.

J Stuart Hunter et al. The exponentially weighted moving average. *J. Quality Technol.*, 18(4): 203–210, 1986.

Intel. Getting started with big data analytics in retail. 2014. `http://www.intel.co.uk/content/dam/www/public/us/en/documents/solution-briefs/retail-big-data-analytics-solution-blueprint.pdf`.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Hemant Ishwaran and Lancelot F James. Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankhyā: The Indian Journal of Statistics*, pages 577–592, 2003.

Hemant Ishwaran and Mahmoud Zarepour. Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

Shane T Jensen, Kenneth E Shirley, and Abraham J Wyner. Bayesball: a Bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics*, pages 491–520, 2009.

Dominik Joho, Martin Senk, and Wolfram Burgard. Learning wayfinding heuristics based on local information of object maps. In *ECMR*, pages 117–122, 2009.

Prajakta S Kalekar. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi School of Information Technology*, 4329008:1–13, 2004.

Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

Kirthi Kalyanam. Pricing decisions under demand uncertainty: A Bayesian mixture model approach. *Marketing Science*, 15(3):207–221, 1996.

Mikko Kärkkäinen. Increasing efficiency in the supply chain for short shelf life goods using rfid tagging. *International Journal of Retail &amp; Distribution Management*, 31(10):529–536, 2003.

Kishana R Kashwan and CM Velu. Customer segmentation using clustering and data mining techniques. *International Journal of Computer Theory and Engineering*, 5(6):856, 2013.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Byung-Do Kim, Robert C Blattberg, and Peter E Rossi. Modeling the distribution of price sensitivity and implications for optimal retail pricing. *Journal of Business &amp; Economic Statistics*, 13(3):291–303, 1995.

Byung-Do Kim, Kannan Srinivasan, and Ronald T Wilcox. Identifying price sensitive consumers: the relative merits of demographic vs. purchase pattern information. *Journal of Retailing*, 75 (2):173–193, 1999.

Umay Uzunoglu Kocer. Forecasting intermittent demand by markov chain model. *International Journal of Innovative Computing, Information and Control*, 9(8):3307–3318, 2013.

Andrey Nikolaevich Kolmogorov. Foundations of probability. 1933.

Athanasios Kottas. Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136(3):578–596, 2006.

Athanasios Kottas. Bayesian hierarchical modeling for prediction of extremes of financial indexes. 2013.

Nikolaos Kourentzes. Intermittent demand forecasts with neural networks. *International Journal of Production Economics*, 143(1):198–206, 2013.

Nikolaos Kourentzes. On intermittent demand model optimisation and selection. *International Journal of Production Economics*, 156:180–190, 2014.

Eric Lai, Daniel Moyer, Baichuan Yuan, Eric Fox, Blake Hunter, Andrea L Bertozzi, and Jeffrey Brantingham. Topic time series analysis of microblogs. Technical report, DTIC Document, 2014.

Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

John W Lau and Peter J Green. Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558, 2007.

Richard D Lawrence, George S Almasi, Vladimir Kotlyar, Marisa Viveros, and Sastry S Duri. Personalization of supermarket product recommendations. In *Applications of Data Mining to Electronic Commerce*, pages 11–32. Springer, 2001.

J-H Lee, G Han, WJ Fulp, and AR Giuliano. Analysis of overdispersed count data: application to the human papillomavirus infection in men (him) study. *Epidemiology &amp; Infection*, 140(6):1087–1094, 2012.

Jonathan Lee, Peter Boatwright, and Wagner A Kamakura. A Bayesian model for prelaunch sales forecasting of recorded music. *Management Science*, 49(2):179–196, 2003.

Peter SH Leeflang and Josefa Parreño-Selva. Cross-category demand effects of price promotions. *Journal of the Academy of Marketing Science*, 40(4):572–586, 2012.

Na Liu, Shuyun Ren, Tsan-Ming Choi, Chi-Leung Hui, and Sau-Fun Ng. Sales forecasting for fashion retailing service industry: A review. *Mathematical Problems in Engineering*, 2013.

Albert Y Lo et al. On a class of Bayesian nonparametric estimates: I. Density estimates. *The annals of statistics*, 12(1):351–357, 1984.

Jackie Y Luan and K Sudhir. Forecasting advertising responsiveness for short-lifecycle products. In *Marketing Science Conference at Atlanta*, 2005.

David JC MacKay. *Bayesian methods for adaptive models.* PhD thesis, California Institute of Technology, 1992.

Charles F Manski. The structure of random utility models. *Theory and decision*, 8(3):229–254, 1977.

Ina S Markham and Terry R Rakes. The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. *Computers &amp; operations research*, 25(4):251–263, 1998.

Jeffrey I McGill and Garrett J Van Ryzin. Revenue management: research overview and prospects. *Transportation science*, 33(2):233–256, 1999.

Mario Medvedovic, Ka Yee Yeung, and Roger Eugene Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

Xiao-Li Meng. Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160, 1994.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Cliona Ni Mhurchu, Helen Eyles, Chris Schilling, Qing Yang, William Kaye-Blake, Murat Genç, and Tony Blakely. Food prices and consumer demand: differences across income levels and ethnic groups. *PLoS One*, 8(10):e75934, 2013.

Borislava Mihaylova, Andrew Briggs, Anthony O'hagan, and Simon G Thompson. Review of statistical methods for analysing healthcare resources and costs. *Health economics*, 20(8): 897–916, 2011.

Yongyi Min and Alan Agresti. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19, 2005.

Prerna Mishra, Xue-Ming Yuan, Guangbin Huang, and Truong Ton Hien Duc. Intermittent demand forecast: Robustness assessment for group method of data handling, 2014.

Maryam Mohammadipour. *Intermittent demand forecasting with integer autoregressive moving average models*. PhD thesis, 2013.

Marek Molas and Emmanuel Lesaffre. Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in medicine*, 29(30):3294–3310, 2010.

Govind S Mudholkar and Deo Kumar Srivastava. Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42(2):299–302, 1993.

Francis J Mulhern and Robert P Leone. Implicit price bundling of retail products: a multi-product approach to maximizing store profitability. *The Journal of Marketing*, pages 63–76, 1991.

John Mullahy. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365, 1986.

Peter Müller, Fernando Quintana, and Gary Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749, 2004.

Iain Murray and Zoubin Ghahramani. A note on the evidence and Bayesian occam's razor. 2005.

Manal M Nassar and Fathy H Eissa. On the exponentiated Weibull distribution. *Communications in Statistics-Theory and Methods*, 32(7):1317–1336, 2003.

Jorge Navarro and Moshe Shaked. Hazard rate ordering of order statistics and systems. *Journal of Applied Probability*, 43(2):391–408, 2006.

Peter Neal, Gareth Roberts, et al. Optimal scaling for partially updating mcmc algorithms. *The Annals of Applied Probability*, 16(2):475–515, 2006.

Radford M Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.

Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Brian Neelon, Pulak Ghosh, and Patrick F Loebs. A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):389–413, 2013.

Brian H Neelon, A James O'Malley, and Sharon-Lise T Normand. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, 10(4):421–439, 2010.

Long Ngo, Ira B Tager, and Doris Hadley. Application of exponential smoothing for nosocomial infection surveillance. *American Journal of Epidemiology*, 143(6):637–647, 1996.

Nielsen Company. Get with the program: Card-carrying consumer perspectives on retail loyalty-program participation and perks, 2016.

Yosihiko Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

Jorge M Oliveira, Gordon R Foxall, and Teresa C Schrezenmaier. Consumer brand choice: Individual and group analyses of demand elasticity. In *The Behavioral Economics of Brand Choice*, pages 223–255. Springer, 2007.

Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2011.

Afshin Oroojlooyjadid, Lawrence Snyder, and Martin Takáč. Applying deep learning to the newsvendor problem. *arXiv preprint arXiv:1607.02177*, 2016.

Omiros Papaspiliopoulos and Gareth O Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

Cristian Pasarica and Andrew Gelman. Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, pages 343–364, 2010.

Per-Göran Persson. *Modeling the impact of sales promotion on store profits.* Foundation for Distribution Research, Economic Research Institute, Stockholm School of Economics (EFI), 1995.

Nicholas C Petruzzi and Maqbool Dada. Pricing and the newsvendor problem: A review with extensions. *Operations research*, 47(2):183–194, 1999.

Michael D Porter and Gentry White. Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, 6(1):106–124, 2012.

Joanne M Potts and Jane Elith. Comparing species abundance models. *Ecological Modelling*, 199(2):153–163, 2006.

A Nasiri Pour, B Rostami Tabar, and A Rahimzadeh. A hybrid neural network and traditional approach for forecasting lumpy demand. *Proc. World Acad. Sei. Eng. Technol*, 30:384–389, 2008.

Steven David Prestwich, S Armagan Tarim, Roberto Rossi, and Brahim Hnich. Forecasting intermittent demand by hyperbolic-exponential smoothing. *International Journal of Forecasting*, 30(4):928–933, 2014.

Mohammad Anwar Rahman and Bhaba R Sarker. Intermittent demand forecast and inventory reduction using Bayesian arima approach. In *Proceedings of the 2010 International Conference on Industrial Engineering and Operations Management Dhaka, Bangladesh*, 2010.

Carl ddd Edward Rasmussen. The infinite Gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

Brian J Reich and Montserrat Fuentes. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, pages 249–264, 2007.

Chris Rhodes. The retail industry: statistics and policy. *House of Commons Library*, Number 06186, 2017, 2017.

Chris Rhodes and Philip Brien. The retail industry: statistics and policy. *House of Commons Library*, Number 06186, 2018, 2018.

Timothy Richards, Stephen Hamilton, Koichi Yonezawa, et al. Retail market power in a shopping basket model of supermarket competition. Technical report, 2015.

Martin Ridout, Clarice GB Demétrio, and John Hinde. Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192, 1998.

Gareth O Roberts. Markov chain concepts related to sampling algorithms. *Markov chain Monte Carlo in practice*, 57, 1996.

Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic processes and their applications*, 49 (2):207–216, 1994.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Charles E Rose, Stacey W Martin, Kathleen A Wannemuehler, and Brian D Plikaytis. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics*, 16(4):463–481, 2006.

Jeffrey S Rosenthal et al. Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, 4, 2011.

Peter Rossi. *Bayesian non-and semi-parametric methods and applications*. Princeton University Press, 2014.

Cynthia Rudin, Benjamin Letham, and David Madigan. Learning theory analysis for association rules and sequential event prediction. *Journal of Machine Learning Research*, 14(1):3441–3492, 2013.

SK Sahu and TMF Smith. A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2): 235–253, 2006.

Sujit K Sahu, Bernard Baffour, Paul R Harper, John H Minty, and Christophe Sarran. A hierarchical Bayesian model for improving short-term forecasting of hospital demand by including meteorological information. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(1):39–61, 2014.

Pedro LC Saldanha, Elaine A De Simone, and PF Frutuoso e Melo. An application of non-homogeneous Poisson point processes to the reliability analysis of service water pumps. *Nuclear Engineering and Design*, 210(1):125–133, 2001.

Badrul M Sarwar, George Karypis, Joseph Konstan, and John Riedl. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the fifth international conference on computer and information technology*, volume 1, 2002.

Ida Scheel, Egil Ferkingstad, Arnoldo Frigessi, Ola Haug, Mikkel Hinnerichsen, and Elisabeth Meze-Hausken. A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):85–100, 2013.

Matthias Seeger, Syama Rangapuram, Yuyang Wang, David Salinas, Jan Gasthaus, Tim Januschowski, and Valentin Flunkert. Approximate Bayesian inference in linear state space models for intermittent demand forecasting at scale. *arXiv preprint arXiv:1709.07638*, 2017.

Matthias W Seeger, David Salinas, and Valentin Flunkert. Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems*, pages 4646–4654, 2016.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica sinica*, 4:639–650, 1994.

Lydia Shenstone and Rob J Hyndman. Stochastic models underlying croston's method for intermittent demand forecasting. *Journal of Forecasting*, 24(6):389–402, 2005.

Chris Sherlock, Paul Fearnhead, and Gareth O Roberts. The random walk metropolis: linking theory and practice through a case study. *Statistical Science*, pages 172–190, 2010.

Ya-Yueh Shih and Duen-Ren Liu. Hybrid recommendation approaches: collaborative filtering via valuable content information. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 217b–217b. IEEE, 2005.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, Sigrunn H Sørbye, et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28, 2017.

Ralph D Snyder, J Keith Ord, and Adrian Beaumont. Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting*, 28(2): 485–496, 2012.

Stan Development Team. RStan: the R interface to Stan, 2016. URL `http://mc-stan.org/`. R package version 2.14.1.

Statista. Grocery market share in great britain 2015-2017, 2017.

Thomas J Steenburgh. Measuring consumer and competitive impact with elasticity decompositions. *Journal of marketing Research*, 44(4):636–646, 2007.

Nenad Stefanovic, Dusan Stefanovic, and Bozidar Radenkovic. Application of data mining for supply chain inventory forecasting. *Applications and Innovations in Intelligent Systems XV*, pages 175–188, 2008.

Ron Sun. *The Cambridge handbook of computational psychology*. Cambridge University Press, 2008.

Pete Swabey. Tesco saves millions with supply chain analytics. *Information Age magazine*, 2013.

Aris A Syntetos and John E Boylan. The accuracy of intermittent demand estimates. *International Journal of forecasting*, 21(2):303–314, 2005.

Aris A Syntetos and John E Boylan. Demand forecasting adjustments for service-level achievement. *IMA Journal of Management Mathematics*, 19(2):175–192, 2007.

Aris A Syntetos, John E Boylan, and JD Croston. On the categorization of demand patterns. *Journal of the Operational Research Society*, 56(5):495–503, 2005.

Matthew A Taddy, Athanasios Kottas, et al. Mixture modeling for marked poisson processes. *Bayesian Analysis*, 7(2):335–362, 2012.

Kei Takahashi, Marina Fujita, Kishiko Maruyama, Toshiko Aizono, and Koji Ara. Forecasting intermittent demand with generalized state-space model. In *Operations Research Proceedings 2014*, pages 589–596. Springer, 2016.

Yee Whye Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.

Tesco PLC. Annual report and financial statements 2017, 2017.

Ruud H Teunter, Aris A Syntetos, and M Zied Babai. Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3):606–615, 2011.

Jun Tu and Guofu Zhou. Incorporating economic objectives into Bayesian priors: Portfolio choice under parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 45(4): 959–986, 2010.

Brandon M Turner, Per B Sederberg, Scott D Brown, and Mark Steyvers. A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, 18 (3):368, 2013.

Jason J Turner and Karen Wilson. Grocery loyalty: Tesco clubcard and its impact on loyalty. *British Food Journal*, 108(11):958–964, 2006.

Alejandro Jara Vallejos. *Bayesian semiparametric methods for the analysis of complex data.* PhD thesis, Tese (Doutorado em Ciências)-Faculdade de Ciências, Universidade Católica de Leuven, 2008.

Luis Henrique Rigato Vasconcellos and Mauro Sampaio. The stockouts study: an examination of the extent and the causes in the são paulo supermarket sector. *BAR-Brazilian Administration Review*, 6(3):263–279, 2009.

Sara Wade, David B Dunson, Sonia Petrone, and Lorenzo Trippa. Improving prediction from Dirichlet process mixtures via enrichment. *Journal of Machine Learning Research*, 15(1): 1041–1071, 2014.

Stephen G Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54, 2007.

Rockney G Walters. Assessing the impact of retail price promotions on product substitution, complementary purchase, and interstore sales displacement. *The Journal of Marketing*, pages 17–28, 1991.

Rockney G Walters and Scott B MacKenzie. A structural equations analysis of the impact of price promotions on store performance. *Journal of marketing research*, pages 51–63, 1988.

Lianming Wang and David B Dunson. Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

Colin G Weaver, Pietro Ravani, Matthew J Oliver, Peter C Austin, and Robert R Quinn. Analyzing hospitalization data: potential limitations of poisson regression. *Nephrology Dialysis Transplantation*, 30(8):1244–1249, 2015.

Thomas R Willemain, Charles N Smart, and Henry F Schwarz. A new approach to forecasting intermittent demand for service parts inventories. *International Journal of forecasting*, 20(3): 375–387, 2004.

Simon P Wilson and Francisco J Samaniego. Nonparametric analysis of the order-statistic model in software reliability. *IEEE transactions on software engineering*, 33(3), 2007.

Holbrook Working. Statistical laws of family expenditure. *Journal of the American Statistical Association*, 38(221):43–56, 1943.

Qingzheng Xu, Na Wang, and Heping Shi. Review of Croston's method for intermittent demand forecasting. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pages 1456–1460. IEEE, 2012.

Xiaoming Yan and Yong Wang. A newsvendor model with capital constraint and demand forecast update. *International Journal of Production Research*, 52(17):5021–5040, 2014.

Shuai Yang, Yujie Xiao, and Yong-Hong Kuo. The supply chain design for perishable food with stochastic demand. *Sustainability*, 9(7):1195, 2017.

Shuang-Hong Yang and Hongyuan Zha. Mixture of mutually exciting processes for viral diffusion. In *International Conference on Machine Learning*, pages 1–9, 2013.

Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368, 1962.

XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, and Deepak Agarwal. Glmix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 363–372. ACM, 2016.

Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522. ACM, 2015.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *AISTATS*, volume 13, pages 641–649, 2013.

Mathias Zocher. Multivariate mixed Poisson processes. 2005.