The Annals of Applied Statistics
2024, Vol. 18, No. 2, 946–965
https://doi.org/10.1214/23-AOAS1811
© Institute of Mathematical Statistics, 2024

BAYESIAN HIERARCHICAL MODELLING OF SPARSE COUNT PROCESSES IN RETAIL ANALYTICS

BY JAMES PITKIN[2,a], IOANNA MANOLOPOULOU[2,1,b] AND GORDON ROSS[3,c]

[1]*The Alan Turing Institute*
[2]*University College London,* [a]*james.pitkin@cantab.net,* [b]*i.manolopoulou@ucl.ac.uk*
[3]*University of Edinburgh,* [c]*gordon.ross@ed.ac.uk*

The field of retail analytics has been transformed by the availability of rich data, which can be used to perform tasks such as demand forecasting and inventory management. However, one task which has proved more challenging is the forecasting of demand for products which exhibit very few sales. The sparsity of the resulting data limits the degree to which traditional analytics can be deployed. To combat this, we represent sales data as a structured sparse multivariate point process, which allows for features such as autocorrelation, cross-correlation, and temporal clustering, known to be present in sparse sales data. We introduce a Bayesian point process model to capture these phenomena, which includes a hurdle component to cope with sparsity and an exciting component to cope with temporal clustering within and across products. We then cast this model within a Bayesian hierarchical framework, to allow the borrowing of information across different products, which is key in addressing the data sparsity per product. We conduct a detailed analysis, using real sales data, to show that this model outperforms existing methods in terms of predictive power, and we discuss the interpretation of the inference.

1. Introduction. One of the main objectives of retail analytics is to build predictive demand forecasting models for purposes such as inventory management, profit forecasting, assessing the impact of marketing to name but a few. Demand models have been extensively studied in the literature, focusing on forecasting sales of high volumes (Seeger, Salinas and Flunkert (2016b), Ferreira, Lee and Simchi-Levi (2015), Sahu et al. (2014)). However, these forecasting models often struggle to capture the demand dynamics of products with low sales volumes. Such products, known as slow-moving-inventory (SMI), are typically for sale the entire year but are only purchased on 1–5% of days, often with an intermittent pattern. They are usually nonfood merchandise such as technology, fashion, and general household items. The resultant demand data of SMI take the form of a sparse count process per product, largely populated with zeros, with autocorrelation and contemporaneous structure across different products (due to seasonality, promotions, and current trends).

There are three main aspects of a predictive model of SMI which are challenging. First, since these products have low sales volumes, this leads to an inflation of zeros (corresponding to days with no sales), which makes it difficult to estimate the effect of the traditional variables used in forecasting models (prices, promotions, seasonality, etc). Second, SMI demand often occurs in bursts across different products, indicating a dependency either between a product's own sales history and the history of other similar products, or on a common external factor that cannot be accounted for by available covariates. Third, SMI is often stocked and sold for a relatively limited amount of time (short sales cycles), which results in little covariate and demand history.

Received May 2021; revised July 2023.
Key words and phrases. Self-excitation, cross-excitation, hurdle model, Hawkes process, intermittent demand, slow-moving-inventory, demand forecasting.

Previous research dealing with such zero-inflated bursty processes includes exponential smoothing and related methodologies that attempt to forecast future observations as a weighted moving average of past observations over time (Do Croston (1972), Gardner (2006)). Such approaches primarily focus on the temporal burstiness of demand and demonstrated initial success, though they are somewhat heuristic and lack an underlying stochastic process consistent with intermittent demand and fail to provide a framework that naturally accounts for predictors, information borrowing, and uncertainty (Shenstone and Hyndman (2005)). More recent developments have included neural network approaches (Seeger, Salinas and Flunkert (2016a), Rangapuram et al. (2018), Salinas et al. (2020)) that show promise at finding the complex nonlinear interdependencies across multiple intermittent demand series across but suffer from overfitting issues and lack an underlying interpretability (Kourentzes (2013), Pour, Tabar and Rahimzadeh (2008), Mishra et al. (2014)). Two main contributions in the literature explicitly accommodate zero inflation in the context of demand series. The first is Chapados (2014), who implement a Bayesian hierarchical zero-inflated count model with time-varying regression parameters that shares information across intermittent demand series. However, their approach limits the dependency on historical demand to an AR(1) process in the mean of the count distribution and ignores the zero-process altogether, exclude pricing information from their framework and without considering contemporaneous dependence between intermittent demand series. The second is Berry, Helman and West (2020) who developed a dynamic zero-inflated multiscale mixture model of demand time series; the distinction with our method is that they used transaction-level data, whereas we only have aggregate sales data at hand. Though existing approaches have demonstrated a degree of success at forecasting the intermittent demand of SMI, none have developed a unified model that incorporates excitation dynamics, covariates beyond just seasonality, and information pooling between the intermittent demand series in a way that sheds light into additive benefits that each of these components has with respect to forecasting performance.

In this work we develop modelling, inferential, and predictive methods able to learn the dynamics of sparse count processes for SMI products with few to no sales. We build on the self-exciting model for sparse processes of Porter and White (2012) by introducing flexible covariates that relate to product demand and extend their model to include a cross-excitation contribution that allows differing intermittent demand series to excite one another, capturing the process of intertwined contemporaneous excitation dynamics observed in SMI data. We overcome the sparsity of information available for each individual product by integrating the products into a Bayesian hierarchical model that allows for shrinkage and information passing across differing sparse count process.

The layout of this paper is as follows; Section 2 describes the SMI demand data used in this paper. Section 3 describes hurdle models and the Hawkes process. Section 4 outlines our hierarchical Bayesian hurdle model with self- and cross-excitation components to model multiple sparse count processes simultaneously. Section 5 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 nonhierarchical equivalent and models without the self- and cross-excitation terms to highlight the benefits of the information borrowing and excitation components and discuss the implications of these results within the context of retail analytics. Section 6 concludes with a summary of our contributions and a discussion of possible future developments.

2. Data.

The dataset we consider consists of product sales information recorded through the electronic points of sale of a leading United Kingdom supermarket retailer, anonymised for general research purposes so that no individual shoppers could be identified. Access to this anonymised dataset was provided by dunnhumby Ltd. The data consist of 17 longitudinal

TABLE 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	% nonzero 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

SMI sales processes over 464 days of trading between the dates 1 October 2013 to 7 January 2015. For each product the daily count corresponds to the aggregated sales of a touchscreen tablet across five large supermarkets within south London. 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 October 2013 to 29 September 2014 (a full trading year excluding Christmas), and the remaining 100 trading days between 30 September 2014 to 7 January 2015 kept as hold-out test set. These training and test splits give 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. This dataset is challenging since we only have one year to learn seasonality from and thus makes a hierarchical model formulation particularly applicable.

Table 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% nonzero sales days.

These data demonstrate many of the pertinent features of SMI sales processes. Figure 1 contrasts the sales and respective prices of one of the faster-selling tablets against a slower one. The plots illustrate the zero inflation where many days record zero sales. The sales also 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 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 an excitation process not accounted for by covariate information, as sales bursts occur outside the effects explained by covariate data. Figure 2 provides plots suggesting the existence of possible contemporaneous excitation of tablet sales within a particular brand. We see that sales of a tablet in a given brand are often followed by a subsequent sale of another tablet of the same brand.

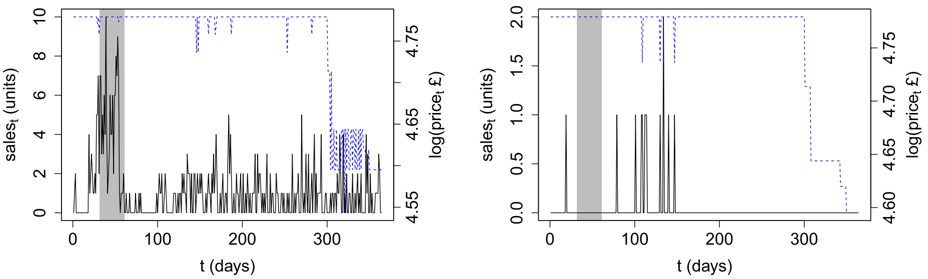

FIG. 1. *Plots of demand series (i.e., the daily number of sales) as a solid black line for two tablets with their respective log prices in \$ (dashed) over* 364 *days of training data. The left panel is a high-volume tablet (Product* 1*) and the right panel is a lower-volume tablet (Product* 7*). The shaded region is the month prior to Christmas.*

3. Background. Our aim is to develop a Bayesian hierarchical model for the sales of each product i on each day t, denoted $y_{it} \in \{0, 1, \ldots\}$. In order to deal with the zero inflation, we will use the hurdle regression model framework of separately modelling the probability of each day having at least one sale and the distribution of the number of sales conditional on there being at least one sale. In other words, we refer to a day as being a "zero" day if no sales occur or a "nonzero" day if there is at least one sale and treat the classification of days as zero vs nonzero as being a Bernoulli (binary) sequence. The number of sales on the nonzero days is then separately modelled.

We will combine this hurdle model framework with self- and cross-excitation components to account for the clustering of events. The remainder of this section reviews the hurdle and excitation models upon which we will build.

3.1. *Hurdle models.* Mullahy (1986) introduced the hurdle regression model to handle an inflation of zeros in count data for which traditional count models (Poisson, negative binomial regression) could not adequately account. The hurdle model defines a distribution over the counts $\{0, 1, \ldots\}$ 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 nonzero counts. Hurdle models, unlike their zero-inflated model counterpart (Lambert (1992)), assume the zero and nonzero processes are separable, as zero observations arise exclusively from the degenerate zero distribution and the count distribution over $\{1, \ldots\}$. We opt for a hurdle model over a zero-inflated model, due to the separability of the zero and count processes (that accommodates efficient inference), so that any occurrences of zero can be directly linked to the zero process.

Within our context of SMI modelling, the inflation of zeros corresponds to days when we observe zero sales, and the count process corresponds to days when we observe nonzero

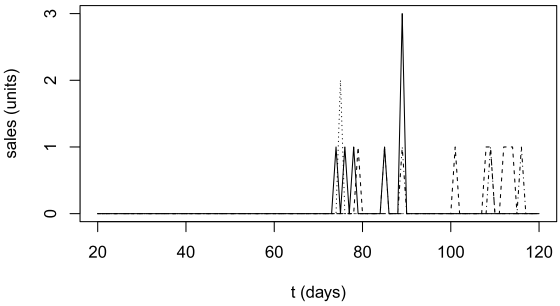

FIG. 2. *Plots of tablet sales for the four products in the GADGET brand over a portion of the training set. The differing colours correspond to the sales of a particular product within the given brand.*

sales. More concretely, let the number of sales of a particular product i be denoted as y_t (we suppress dependence on the product i subscript for notational convenience). The probability density function of the hurdle model, given covariates $\boldsymbol{x}_t$, can be specified as

$$p(y_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}) = \begin{cases} 1 - p(\boldsymbol{x}_t^z, \boldsymbol{\theta}^z) & \text{for } y_t = 0, \\ p(\boldsymbol{x}_t^z, \boldsymbol{\theta}^z) f(y_t \mid \boldsymbol{x}_t^c, \boldsymbol{\theta}^c) & y_t = 1, \dots. \end{cases} \tag{3.1}$$

Here $p(\boldsymbol{x}_t^z, \boldsymbol{\theta}^z)$ is the probability of observing a nonzero count at time t, and $f(\cdot \mid \boldsymbol{x}_t^c, \boldsymbol{\theta}^c)$ is a probability mass function defined on the positive integers. The covariates for the zero process $\boldsymbol{x}_t^z$ and count process $\boldsymbol{x}_t^c$ may overlap. The $\boldsymbol{\theta}^z$, $\boldsymbol{\theta}^c$ are parameters for the zero and count processes, respectively. For notational purposes we let E_t be the indicator for an event day such that $E_t = 1$ if $y_t \geq 1$ (a day t where at least one sales instance is observed) and $E_t = 0$ if $y_t = 0$ (a day t with no sales).

3.2. *Self-exciting processes.* Hawkes (1971) introduced a Hawkes process as a self-exciting temporal point process with conditional intensity function

$$\lambda(t) = \varphi(t) + \sum_{i:t_i<t} \nu(t - t_i), \tag{3.2}$$

where $\varphi(t)$ is the background rate, t_i are the times prior to time t when an event (i.e., nonzero sales) occurred, and $\nu(\cdot)$ a continuous excitation function that controls the extent to which events cluster together. This process effectively describes a count process where events increase the probability of further such events in the short term, leading to clustered events (in our case, days with nonzero sales). In the discrete context, the above can be reexpressed as

$$\lambda(t) = \varphi(t) + \sum_{j<t} \kappa E_j g(t - j), \tag{3.3}$$

where $\varphi(t)$ is, as before, the background rate, E_t is a Boolean indicator indicating event days ($E_t = 1$ for an event day, that is, a day with nonzero sales), $g(\cdot) \geq 0$ is the excitation kernel (a probability mass function) that controls the extent to which events cluster together, and κ is some trigger constant that can be interpreted as the average number of triggered events produced by each event. With a Hawkes process, instances of an event in turn increase ($\kappa > 0$) or decrease ($\kappa < 0$) the probability of further such events occurring in the future. In this work we focus on the case $\kappa > 0$, which represents excitation (rather than inhibition). We denote the history of events up to but not including t as $H_{t-1} = (E_1, \dots, E_{t-1})$. Figure 3 plots two simulated series from a Bernoulli distribution with a Hawkes process term. It illustrates

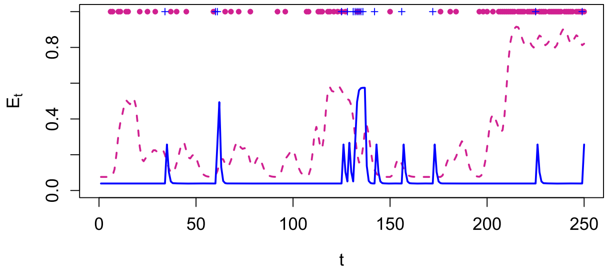

FIG. 3. *Simulated example. Two series of samples are generated from $E_t \sim$ Bernoulli(p_t), with* $\text{logit}(p_t) = \theta + \kappa \sum_{i<t} E_i g(t - i \mid \mu, \tau)$ *for $t = 1, \dots, 364$, where $g(\cdot \mid \mu, \tau)$ is the truncated negative binomial density on the positive integers with mean and scale μ, τ. The crosses are E_t samples generated from* $(\theta, \kappa, \mu, \tau) = (-3.2, 3.1, 1.0, 5.0)$, *and the solid line is the corresponding p_t. The circle dots are E_t samples generated from* $(\theta, \kappa, \mu, \tau) = (-2.5, 5, 5, 60)$, *and the dashed line is the corresponding p_t. We observe how the differing $(\theta, \kappa, \mu, \tau)$ lead to different clustering patterns and underlying shapes of the probability of events.*

the variation in Bernoulli samples depending on the parameters of the excitation kernel and trigger constant. For example, the dashed curve with the higher excitation constant κ shows much stronger excitation as exhibited by the densely clustered events dots and as opposed to the crosses which are mostly isolated events.

3.3. *Cross-exciting processes.* Various extensions to (3.3) have been made to include cross-excitation across related spatial or temporal processes. Lai et al. (2016) proposed a scheme allowing for interexcitation and inhibition across different social media events across both time and space domain. They used a triggering kernel, specified as exponential in time and Gaussian in space, to capture cross-excitation and inhibition in tweets in different topics and geographies. Zhou, Zha and Song (2013) used multidimensional Hawkes process (in the continuous space) to model information spread across sparse low-rank social networks and a triggering function which incorporates excitation from connected individuals in an additive form. Blundell, Beck and Heller (2012) modelled interaction between human relationships using linked Hawkes processes through a kernel trigger function for the cross entries, which are linked via a nonparametric Chinese restaurant process to determine the partitions amongst social groups. Although the aforementioned approaches demonstrate a degree of success within their relevant contexts, they have not been applied to sales forecasting before.

4. 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 introduces a Bayesian hierarchical version of the hurdle model of (3.1) with self- and cross-excitation terms discussed in Section 4.2 in both the zero and count components. Our proposed model makes the following three extensions to existing models: first, we use covariates beyond seasonal information; in particular, we use price along Boolean seasonal variables to assist in forecasting sales. Second, we use cross-excitation in the zero process of (3.1) that aims to capture the contemporaneous nature of sales bursts across the SMI category. Third, we build a Bayesian hierarchical model across the sales y_{it} (the sales at product i at time t) of a SMI category to allow information borrowing, which is key in addressing the sales sparsity per product.

4.1. *Covariate data.* In addition to the excitation exhibited in SMI sales, product level covariates may offer predictive power to SMI forecasting. We introduce covariate data into the model through the background intensity function $\varphi(t)$ of (3.3). 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 the logarithm of its price, along with the indicator functions of week day, month, and Christmas period. We summarise these covariates as

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

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

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. We specify the background intensities $\varphi_i^z(t)$, $\varphi_i^c(t)$ of the zero and count processes of (3.3) as

$$\varphi_i^z(t) = \theta_{i1}^z + \theta_{i2}^z \log(\mathrm{p}_{it}) + \sum_{k=1}^{18} \theta_{i,k+2}^z \mathrm{s}_{kt}, \tag{4.1}$$

$$\varphi_i^c(t) = \theta_{i1}^c + \theta_{i2}^c \log(\mathrm{p}_{it}), \tag{4.2}$$

where $\{\theta^z_{i1}, \dots, \theta^z_{i20}\}$ and $\{\theta^c_{i1}, \theta^c_{i2}\}$ are the parameters associated with the zero and count processes, respectively, for product i. The j index of θ^z_{ij} ranges from $1 - 20$ to include the 1 additive constant, 1 log price variable, 6 week day, 11 month and 1 Christmas indicators. Functions (4.1) and (4.2) 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 overfitting. It is important to note that, for a given product, the count process only exists for t with $E_t = 1$. This reduces the count process data compared to the zero process. The link functions of (4.1) and (4.2) are context-specific and will be specified in the data analysis sections. We now denote these covariates as $\boldsymbol{x}^z_{it} = (\mathrm{p}_{it}, \mathrm{s}_t)$ and $\boldsymbol{x}^c_{it} = (\mathrm{p}_{it})$ for the zero and count processes, respectively, in line with notation of (3.1).

4.2. *Cross-excitation.* SMI sales of different but comparable products may occur in 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 autocorrelatively but also contemporaneously across products. In the SMI context, cross-excitation may occur across all products (i.e., a sale for a product leads to a higher probability of a sale of any other product over the subsequent days), within each brand (i.e., a sale of a product only affects subsequent sales of products of the same brand), or may even be observed across groups of products with no apparent relationship. We refer to these groups as "cross-excitation groups" to include any grouping of the products. Concretely, we define $\tilde{E}_{it}$ as the indicator for a *cross-event day* of product i of some cross-excitation group 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 same group, 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 other sale within the cross-excitation group at time t and 0 otherwise. We denote the history of cross-events up to but not including t as $\tilde{H}_{it-1} = (\tilde{E}_{i1}, \dots, \tilde{E}_{it-1})$.

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

$$S_{it} = \sum_{j<t} \kappa_i E_{it} g(t - j \mid \zeta_i), \tag{4.3}$$

$$\tilde{S}_{it} = \sum_{j<t} \tilde{\kappa}_i \tilde{E}_{it} g(t - j \mid \tilde{\zeta}_i), \tag{4.4}$$

where κ_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 (4.4) 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 cross-excitation group and have shared cross-excitation kernels with shared parameters for products corresponding to the same group. We denote these collections of self- and cross-excitation parameters as $\gamma_i = (\kappa_i, \zeta_i)$ and $\tilde{\gamma}_i = (\tilde{\kappa}_i, \tilde{\zeta}_i)$, respectively.

4.3. *Self- and cross-exciting hurdle model.* We formulate our SMI model by utilising the hurdle model specification of (3.1). In particular, we use a logistic link function to model the zero-process with a background intensity $\varphi^z(t)$ (4.1), including seasonal Boolean covariates, logarithm of price as well as self- and cross-excitation components ((4.3) and (4.4)). Similarly, for the count process, we use a negative binomial distribution with a log-link mean intensity $\varphi^c(t)$ (4.2), which includes logarithm of price as well as the self-excitation term of (4.3). Our model is indexed by 17 longitudinal sales series from the tablets category over 464 (training+test) days of trading between the dates 1 October 2013 to 7 January 2015. We specify the probability mass function of the hurdle model as

$$\begin{aligned} &p(y_{it} \mid \boldsymbol{x}_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i) \\ &\quad = \begin{cases} 1 - p(\boldsymbol{x}^z_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}^z_i) & \text{for } y_{it} = 0, \\ p(\boldsymbol{x}^z_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}^z_i) f(y_{it} \mid \lambda(\boldsymbol{x}^c_{it}, H_{it}, \boldsymbol{\theta}^c_i), \phi) & y_{it} \in \mathbb{N}^+, \end{cases} \end{aligned} \tag{4.5}$$

where $\lambda(\cdot)$ represents a link function and $f(y_{it}|\lambda, \phi) = \binom{y_{it}-2+\phi}{y_{it}-1}(\frac{\lambda-1}{\lambda-1+\phi})^{y_{it}-1}(\frac{\phi}{\lambda-1+\phi})^{\phi}$ and $\phi = 1$, which is the probability mass function of the shifted negative binomial distribution (NB) and H_{it}, $\tilde{H}_{it}$, $\boldsymbol{x}^z_{it}$, and $\boldsymbol{x}^c_{i,t}$ are as defined in Sections 3.2, 4.2, and 4.1, respectively, indexed by product i. We specify the link functions as

$$\begin{aligned} \operatorname{logit}(p(\boldsymbol{x}^z_{it}, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}^z_i)) &= \varphi^z_i(t) + S^z_{it} + \tilde{S}^z_{it}, \\ \log(\lambda(\boldsymbol{x}^c_{it}, H_{it}, \boldsymbol{\theta}^c_i)) &= \varphi^c_i(t) + S^c_{it} \end{aligned}$$

$\varphi^z_i(t)$ and $\varphi^c_i(t)$ are as defined from (4.1) and (4.2), respectively, but indexed by product i. We define $S^z_{it} = \sum_{s<t} \kappa^z_i E_{it} g(t - s \mid \mu^z_i, \tau^z_i)$ and $\tilde{S}^z_{it} = \sum_{s<t} \tilde{\kappa}^z_i \tilde{E}_{it} g(t - s \mid \tilde{\mu}^z_i, \tilde{\tau}^z_i)$ similarly to (4.3) and (4.4), respectively, with $g(t \mid \mu, \tau) = \binom{t-2+\tau}{t-1}(\frac{\mu-1}{\mu-1+\tau})^{t-1}(\frac{\tau}{\mu-1+\tau})^{\tau}$ as the shifted NB distribution. We similarly define $S^c_{it} = \sum_{s<t} \kappa^c_i E_{it} g(t - s \mid \mu^c_i, \tau^c_i)$. We denote the collection of shot parameters as $\tilde{\boldsymbol{\gamma}}^z_i = (\tilde{\kappa}^z_i, \tilde{\mu}^z_i, \tilde{\tau}^z_i)$, $\boldsymbol{\gamma}^z_i = (\kappa^z_i, \mu^z_i, \tau^z_i)$, and $\boldsymbol{\gamma_i}^c = (\kappa^c_i, \mu^c_i, \tau^c_i)$ and collectively denote $\boldsymbol{\theta}^z_i = (\theta^z_{i1}, \dots, \theta^z_{i20}, \boldsymbol{\gamma_i}^z, \tilde{\boldsymbol{\gamma}}^z_i)$ and $\boldsymbol{\theta}^c_i = (\theta^c_{i1}, \theta^c_{i2}, \boldsymbol{\gamma_i}^c)$.

During this work special attention is paid to the specification of hierarchical priors over the collection $\boldsymbol{\theta}^z_i$ and $\boldsymbol{\theta}^c_i$, as they are the mechanism through which we penalise complexity and pool information to combat data sparsity. In particular, we specify $\theta^z_{ij} \sim \mathrm{N}(\rho^z_j, (\sigma^z_j)^2)$ and $\rho^z_j \sim \mathrm{N}(\vartheta^z_j, (\zeta^z_j)^2)$ and fix $(\sigma^z_j)^2$ for $j = 1, \dots, 20$ and similarly specify $\theta^c_{ij} \sim \mathrm{N}(\rho^c_j, (\sigma^c_j)^2)$ and $\rho^c_j \sim \mathrm{N}(\vartheta^c_j, (\zeta^c_j)^2)$ and fix $(\sigma^c_j)^2$ for each $j = 1, 2$. For parameters of the shot function S^z_{it}, we specify $\gamma^z_{ij} \sim \mathrm{Gamma}(\eta^z_j, \nu^z_j)$ with $\eta^z_j \sim \mathrm{Gamma}(\alpha^z_j, \delta^z_j)$ and fix ν^z_j for each $j = 1, 2, 3$. We specify priors on $\tilde{\gamma}^z_{ij}$ and γ^c_{ij} similarly. The full details of hierarchical prior specification are contained in Appendix A.1.

5. Results.

We fit variations of the model (4.5) to the 17 longitudinal SMI sales processes over 364 days of trading between the dates 1 October 2013 to 29 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 September 2014 to 7 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 separable, we will model inference and analysis on these separately.

5.1. *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 (3.1), we implement the following cumulative variations of both the link functions as well as the hierarchical layering used in the modelling for each $i = 1, \dots, 17$:

- *Benchmark (Bench)*: As a simple benchmark to assess our models against, we treat the zero-process as being an independent and identically distributed collection of Bernoulli random variables. The empirical probability of a day being a zero day for product i is hence given by $\gamma_i = \sum_{j=1}^{|T^{\text{train}}|} I(y_{it} = 0)/|T^{\text{train}}|$, where $I(\cdot)$ is the identity function. The test-set is then predicted using the standard Bernoulli(γ_i) likelihood.

- *Baseline model (*Base$_1^z$*)*: We learn the zero process with link function

$$\operatorname{logit}\big(p\big(\boldsymbol{x}_{it}^z, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i^z\big)\big) = \varphi_i^z,$$

  that is, a constant probability per product. This is the Bayesian baseline model, as it estimates the zero-process independent of covariate information. The φ_i^z is estimated using vague priors. The performance of this model is used to verify the relative benefits that covariate information brings to SMI zero-process modelling.

- *Hierarchical Bayesian (*HB^z*)*: We learn the zero process with link function

$$\operatorname{logit}\big(p\big(\boldsymbol{x}_{it}^z, H_{it}, \tilde{H}_{it}, \boldsymbol{\theta}_i^z\big)\big) = \varphi_i^z(t),$$

  with the hierarchical prior formulation discussed in Section 4.3. This model is implemented to establish a benchmark of the simplest regression model, that is, a model that excludes information of previous events and is used to verify the relative benefits of self-excitation and cross-excitation.

- *Bayesian with self-excitation (*BE^z*)*: We learn the zero process of the hurdle model with link function

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

  but exclude the hierarchical prior formulation shown in Section 4.3. More concretely, we fix the parameters ρ_j^z, $(\sigma_j^z)^2$, and η_j^z, ν_j^z across all 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.

- *Hierarchical Bayesian with self-excitation (*HBE^z*)*: We learn the zero process with link function

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

  with the hierarchical prior formulation discussed in Section 4.3. This model is implemented to demonstrate the possible benefits of self-excitation in the standard zero-inflated regression model.

- *Bayesian with self- and cross-excitation (*BEC^z*)*: We learn the zero process with link function

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

  but exclude the hierarchical prior formulation shown in Section 4.3. 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.

- *Hierarchical Bayesian with self- and cross-excitation (*HBEC^z*)*: This is the full model discussed in the Section 4.3. We learn the zero process with link function

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

  with the hierarchical prior formulation discussed in Section 4.3. The hyperpriors are selected to balance borrowing across products and penalising complexity.

Parameter inference is performed by Hamiltonian Monte Carlo sampling algorithm and is implemented using the rstan library (Stan Development Team (2016)). Convergence was confirmed by Heidelberger–Welch statistic across all models and parameters (Heidelberger and Welch (1981)). The specification of hyperpriors is included in Appendix A.1.

5.2. *Zero process fits.* The predictive performance of models Base_1^z, HB^z, BE^z, HBE^z, BEC^z, and HBEC^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} (30 September 2014 to 7 January 2015) for each $i = 1, \dots, 17$, given the history of self- and cross-events H_{it}, $\tilde{H}_{it}$, covariate information $\boldsymbol{x}_{it}^z$, and posterior samples. We denote the sth posterior sample of $\boldsymbol{\theta}_i^z$ of the ith product as $\boldsymbol{\theta}_{is}^z$. The sales occurrence probabilities are based on the posterior samples $\boldsymbol{\theta}_{is}^z$ inferred from the training interval T^{train} (between 1 October 2013 to 29 September 2014). More precisely, we apply the following methodology over the test interval:

1. On given day t on the test interval and sth posterior sample, we compute the full predictive posterior distribution of the probability of a sale occurring based conditioned on $\boldsymbol{x}_{it}^z$, H_{it}, $\tilde{H}_{it}$, $\boldsymbol{\theta}_{is}^z$ for each product $i = 1, \dots, 17$.
2. We observe y_{it+1} (the number of sales of product i on day $t + 1$) for each $i = 1, \dots, 17$ and update the self- and cross-event histories H_{it+1}, $\tilde{H}_{it+1}$ for $i = 1, \dots, 17$.
3. Repeat steps for each t, for each sample s and i over the test period of 30 September 2014 to 7 January 2015.

This builds up a set of daily predictive posterior probabilities p_{its} for each $s = 1, \dots, S$ for the probability of a sale on a given day over T^{test} for each $i = 1, \dots, 17$, based on posterior samples inferred from T^{train} conditioned on $\boldsymbol{x}_{it}^z$, H_{it}, $\tilde{H}_{it}$, $\boldsymbol{\theta}_i^z$.

To evaluate the predictive performance of the models for the zero process we use the log-average log-predictive likelihood, also known as the logarithmic score, computed as

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

where p_{its} is the prediction probability of a sale occurring for product i from posterior sample s for some model of interest, which is averaged over the posterior. Table 2 provides the pl^z scores across products and models.

Table 2 reveals some interesting findings. First, all models beat the simple benchmark model which forecasts based on the empirical distribution of the training set. When it comes to our proposed models, we observe that the model HB^z, the zero process model with covariate information, provides a significant improvement in predictive performance, compared to the baseline model Base_1^z without covariate information. We further see that inclusion of a self-excitation component in (3.1) provides a marked improvement over the model HB^z without self-excitation. Figure 4 demonstrates an example of the benefit of including self-excitation 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 2 further indicates the predictive benefits that hierarchical extensions provide over its nonhierarchical equivalents. Figure 5 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 despite the absence of sales over the Christmas period of the previous year for this product. In this way the hierarchical model benefits from inferring parameter values of other intermittent demand series, which have observed sales over the previous the Christmas period.

Finally, Table 2 indicates that the cross-excitation expositions of models BEC^z and HBEC^z offer an improvement in event day prediction over the test set, compared to their noncross-excitation counterparts (i.e., BE^z and HBE^z). Interestingly, cross-excitation does not offer any benefits in terms of the training set but shows significant predictive gains in the test set.

TABLE 2

Predictive log-likelihoods $pl_i^{z,\text{test}}$ and $pl_i^{z,\text{train}}$ scores of the zero process fits for the models Base$_1^z$, HB^z, BE^z, HBE^z, BEC^z, *and* HBEC^z *and each product along with the benchmark model. The best (i.e., highest predictive likelihood) model for each product is shown in bold. The final two rows show the average pl^z averaged across all products in the test and training sets, respectively*

Product i	$\text{pl}^{z,\text{test}}_{\text{Bench}}$	$\text{pl}^{z,\text{test}}_{\text{Base}_1,i}$	$\text{pl}^{z,\text{test}}_{\text{HB},i}$	$\text{pl}^{z,\text{test}}_{\text{BE},i}$	$\text{pl}^{z,\text{test}}_{\text{HBE},i}$	$\text{pl}^{z,\text{test}}_{\text{BEC},i}$	$\text{pl}^{z,\text{test}}_{\text{HBEC},i}$
1	**−0.0028**	−0.0037	−0.0316	−0.0032	−0.0204	−0.0032	−0.0197
2	−0.7372	−0.7347	−0.6566	−0.6085	−0.5587	−0.6042	**−0.5518**
3	−0.0736	−0.0733	−0.0681	−0.0618	**−0.0556**	−0.0623	−0.0559
4	−0.2946	−0.2944	**−0.2827**	−0.293	−0.2854	−0.29	−0.2835
5	−0.1420	−0.1416	−0.1309	−0.1046	−0.1212	**−0.1027**	−0.1181
6	−0.0364	−0.0367	−0.058	−0.0255	−0.0363	**−0.0254**	−0.0363
7	−0.0693	−0.0692	−0.0742	**−0.0591**	−0.0598	−0.06	−0.0607
8	−0.0673	−0.0674	−0.0895	−0.0647	−0.0691	**−0.0642**	−0.0677
9	−0.0617	−0.0597	−0.0727	**−0.0568**	−0.0598	−0.0569	−0.0593
10	−0.0993	**−0.0991**	−0.113	−0.1076	−0.1045	−0.106	−0.1022
11	−0.1796	−0.1716	**−0.1148**	−0.1401	−0.1179	−0.1397	−0.1180
12	−0.0981	**−0.098**	−0.1186	−0.1048	−0.1053	−0.103	−0.1027
13	−0.1615	−0.1584	−0.1525	**−0.0975**	−0.0999	−0.0981	−0.0991
14	−0.1041	−0.1034	**−0.0866**	−0.1115	−0.0993	−0.1111	−0.0995
15	−0.1041	**−0.1036**	−0.1115	−0.1078	−0.1049	−0.1083	−0.1052
16	−0.0560	**−0.0561**	−0.0747	−0.0612	−0.066	−0.0619	−0.0661
17	−0.1520	−0.1501	−0.1523	−0.136	−0.1309	−0.1366	**−0.1307**
$E[\text{pl}^{z,\text{test}}_{\text{model}}]$	−0.1435	−0.1424	−0.1405	−0.1261	−0.1232	−0.1255	**−0.1221**
$E[\text{pl}^{z,\text{train}}_{\text{model}}]$	**−0.0393**	−0.1145	−0.1131	−0.0985	−0.1071	−0.0983	−0.1071

As a last feature of Table 2 that deserves comment, we note that the predictive likelihood for Product 1 is substantially worse than the others, across all models. This is because this product had a relatively high volume of sales with at least one unit being sold on almost half the days in the sample period. This means that the zero process is high entropy, and hence, the predictive likelihoods will be lower. The forecasting performance for Product 3 is also slightly worse than the other products across all models; this is because the number of nonzero sale days for this product increased substantially from 93% in the training set to 98% in the test set, which may be indicative of some structural change.

5.3. *Cross-excitation groups.* The cross-excitation framework depends on user-defined groups. Brand marketing campaigns can lead to an increase of sales across an entire brand,

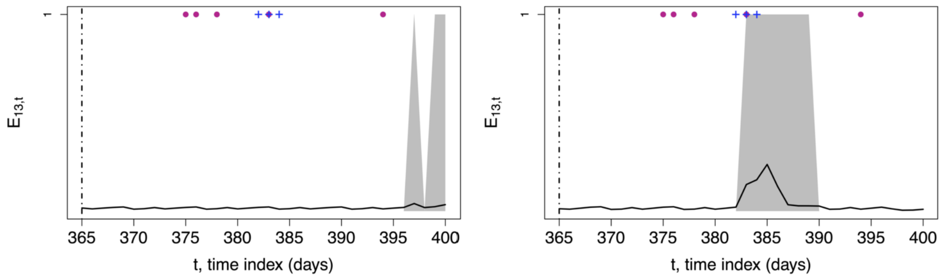

FIG. 4. *Plots of the predictive models* HB^z *(left) and* HBE^z *(right) for product $i = 11$ over a portion of the test set. The crosses and circle dots represent self- and cross-event days, respectively (i.e., E_{it} and $\tilde{E}_{it}$). The black line is the estimated posterior mean of an event day observation (i.e., p_{it}), and the shaded region is the 95% credible interval of these estimates.*

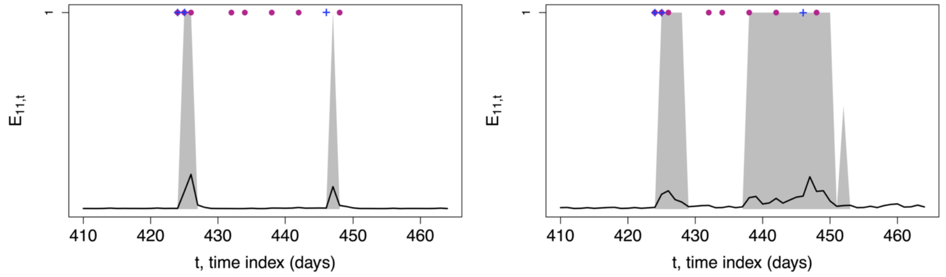

FIG. 5. *Plots of the predictive models* BE^z (*left*) *and* HBE^z (*right*) *for product* $i = 3$ *over a portion of the test set. The crosses and circle dots represent self- and cross-event days, respectively* (*i.e.,* E_{it} *and* $\tilde{E}_{it}$). *The black line is the estimated posterior mean of an event day observation* (*i.e.,* p_{it}), *and the shaded region is the* 95% *credible interval of these estimates.*

which in turn may be captured by the model as cross-excitation; therefore, a natural choice for cross-excitation groups is the product brand. However, other groupings may provide equal good or better model fit. We explore three distinct options for the cross-excitation groups: (a) by product brand, (b) across the entire category (i.e., all products in the category can cross-excite), (c) using data driven preprocessing of products into clusters following the procedure described below:

We create a matrix C to capture cross-excitation from product j to i by computing

$$C_{j,i} = \frac{\sum_{t=t_w}^{T} \sum_{l=t-t_w}^{t-1} \mathbb{1}(y_{it} > 0)\mathbb{1}(y_{jl} > 0)\mathbb{1}(y_{il} = 0)}{\sum_{t=1}^{T} \mathbb{1}(y_{it} > 0) \sum_{t=1}^{T} \mathbb{1}(y_{jt} > 0)}. \tag{5.1}$$

In other words, for each sale event in i we compute the total number of the preceding t_w days which were associated with an event in j but not i. This number is then normalised by the total number of sales days for products i and j to give us the average number of potentially cross-exciting events from j into each sale of i, as a proportion of total sales in j. Here we transform C into a symmetric matrix through $C^{\text{sym}} = C + C^t$. The columns of this matrix are then clustered using complete linkage hierarchical clustering, assuming correlation as distance, visualised in Figure 6. However, note that the model formulation does not require the cross-excitation groupings to be symmetric; one could have product i cross-exciting into product j but not vice-versa; however, this is beyond the scope of this paper.

We cluster the products into three clusters to ensure consistency with the brand groups. Closer inspection shows that product 2, which sells a lot more than the remaining products, has been placed into a cluster by itself; the largest cluster. Which contains all of “GADGET” products as well as some “TECHY,” is largely driven by two products, namely, products 9 and 13, which only sold on two and five days, respectively, but which were associated (either preceeded or followed) by sales of a few other products. Note that the heuristic approximation of cross-excitation groups does not account for excitation, due to simultaneous price reductions, which may be captured through dependence on covariates within the model.

We fit the models using the three types of cross-excitation groups with exactly the same prior distribution specifications as before and show predictive log-likelihoods in Table 3. As can be seen, the best overall model is the one fitted using the empirical clustering described above; notably, the improvement is achieved through products in the larger cluster (products 4, 5, 10, 11, 12, 13). It’s also worth noting that the model allowing “full” excitation across all products performs worst, highlighting the fact that cross-excitations do not happen completely randomly but may be driven by, for example, unobserved targeted promotions.

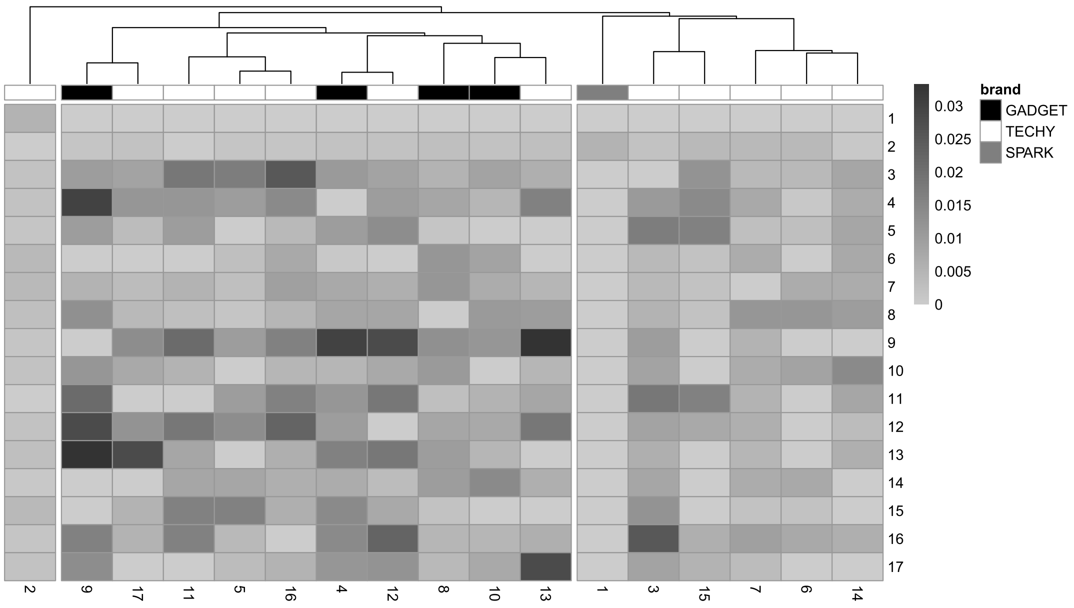

FIG. 6. *Heatmap of approximate cross-excitation matrix, where each column i corresponds to the approximate cross-excitation rate into product i from each row j, given by equation* (5.1). *Row and column labels correspond to product index; the bar above the matrix shows the brand to which each product belongs. The dendrogram shows the full hierarchical clustering using complete linkage and correlation distance. To compare with using brand directly to define cross-excitation groups, we cut the dendrogram into three clusters. Note that the columns have been reshuffled to accommodate visualisation of the dendrogram.*

TABLE 3

Predictive log-likelihoods $pl_i^{z,\text{test}}$ of the zero process fits for the model HBEC^z, *using brand to define cross-excitation groups (left column), our empirical clustering procedure described above (middle column), and cross-excitation across all products (right-hand column). The best (i.e., highest predictive log-likelihood) model for each product is shown in bold. The final row shows the average pl^z averaged across all products in the test set*

Product i	$\mathrm{pl}^{z,\text{test}}_{\text{HBEC},i}$	$\mathrm{pl}^{z,\text{test}}_{\text{HBEC}^c,i}$	$\mathrm{pl}^{z,\text{test}}_{\text{HBEC}^f,i}$
1	**−0.0034**	−0.0051	−0.0051
2	−0.5343	−0.5308	**−0.5303**
3	−0.0603	−0.0643	**−0.0596**
4	−0.2918	**−0.2764**	−0.3243
5	−0.1326	**−0.1259**	−0.1312
6	−0.0249	−0.0273	**−0.0219**
7	−0.0606	−0.0632	**−0.0574**
8	−0.0642	−0.0620	**−0.0607**
9	**−0.0618**	−0.0581	−0.0618
10	−0.0969	**−0.0935**	−0.1028
11	−0.1638	**−0.1628**	−0.1795
12	−0.1032	**−0.0897**	−0.1048
13	−0.1388	**−0.1218**	−0.1375
14	**−0.1062**	−0.1069	−0.1132
15	−0.1072	−0.1094	**−0.1051**
16	−0.0592	−0.0583	**−0.0554**
17	**−0.1502**	−0.1576	−0.1579
$E[\mathrm{pl}^{z,\text{test}}_{\text{model}}]$	−0.1270	**−0.1243**	−0.1299

5.4. *Count process variations.* Similarly to Section 5.2, the benefits of the excitation and hierarchical component to the count process of hurdle model (3.1) are verified by implementing the following cumulative variations in the link functions and hierarchical layerings of the model for each $i = 1, \dots, 17$. These model variations follow the same rationale as with the zero process:

- *Benchmark (Bench)*: As a simple benchmark to assess our models against, we forecast the nonzero test set counts using the empirical density of the nonzero training set counts. For product i, suppose there are n_i training set days with nonzero counts. For each count y_{it} in the test set that also occurs in the training set, the log-predictive likelihood under this benchmark model is

$$\log\biggl(\frac{1}{n_i + 0.5\max_{j,t} y_{jt}}\biggl(0.5 + \sum_{y^*_{it} \in T^{\text{train}}} I(y_{it} = y^*_{it})\biggr)\biggr).$$

  The 0.5 is added to all possible values (here taken as $1, \dots, \max_{j,t} y_{jt}$, the maximum count value observed in the training set) as shrinkage to ensure that test set values, which don't feature in the training set for a particular product, have nonzero probability.
- *Baseline model (*Base^c_1*)*: We learn the count process with link function

$$\log\bigl(\lambda\bigl(\boldsymbol{x}^c_{it}, H_{it}, \boldsymbol{\theta}^c_i\bigr)\bigr) = \varphi^c_i,$$

  that is, a constant rate per product. This is the Bayesian baseline model, as it estimates the zero-process independent of covariate information. The φ^c_i is estimated using vague priors.
- *Hierarchical Bayesian (*HB^c*)*: We learn the count process with link function

$$\log\bigl(\lambda\bigl(\boldsymbol{x}^c_{it}, H_{it}, \boldsymbol{\theta}^c_i\bigr)\bigr) = \varphi^c_i(t),$$

  with the hierarchical prior formulation discussed in Section 4.3.
- *Bayesian with self-excitation (*BE^c*)*: We learn the count process with link function

$$\log\bigl(\lambda\bigl(\boldsymbol{x}^c_{it}, H_{it}, \boldsymbol{\theta}^c_i\bigr)\bigr) = \varphi^c_i(t) + S^c_{it}$$

  but exclude the hierarchical prior formulation shown in Section 4.3.
- *Hierarchical Bayesian with self-excitation (*HBE^c*)*: This is the full model discussed in the Section 4.3. We learn the count process with link function

$$\log\bigl(\lambda\bigl(\boldsymbol{x}^c_{it}, H_{it}, \boldsymbol{\theta}^c_i\bigr)\bigr) = \varphi^c_i(t) + S^c_{it},$$

  with the hierarchical prior formulation discussed in Section 4.3.

Parameter inference is performed by Hamiltonian Monte Carlo sampling algorithm and is implemented using the rstan library (Stan Development Team (2016)). Convergence was confirmed by Heidelberger–Welch statistic across all models and parameters (Heidelberger and Welch (1981)). The specification of these hyperpriors and constant of models HB^c, BE^c, and HBE^c is included in Appendix A.1.

5.5. *Count process fits.* Similarly to the zero processes outlined in Section 5.2, we test the performance of the count variation models Base^c_1, HB^c, BE^c, and HBE^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 September 2014 to 7 January 2015) for each $i = 1, \dots, 17$, given the history of self events H_{it}, covariate information $\boldsymbol{x}^c_{it}$, and posterior samples. We apply the same methodology over the test interval as with the zero process:

1. On event day t (i.e., $E_t = 1$) on the test interval and sth posterior sample, we compute the full predictive posterior distribution of the volume of sales occurring conditioned on H_{it}, $\boldsymbol{x}^c_{it}$, $\boldsymbol{\theta}^c_{is}$ for each $i = 1, \dots, 17$.
2. We observe y_{it+1} (the volume of sales of product i on day $t + 1$) for each $i = 1, \dots, 17$ and update the self-event histories H_{it+1} for $i = 1, \dots, 17$.
3. Repeat steps for each t, for each sample s and i over the test period of 30 September 2014 to 7 January 2015.

This builds up a set of posterior rates λ_{its} for samples $s = 1, \dots, S$ for the probability of the number of sales on a given event day over T^{test} for each $i = 1, \dots, 17$ based on our posterior sample fits inferred from T^{train} conditioned on $\boldsymbol{x}^c_{it}$, H_{it}, $\boldsymbol{\theta}^c_i$.

Similarly to the zero process, we evaluate the predictive performance by calculating the log-predictive likelihood for each of the products $i = 1, \dots, 17$. The log-predictive likelihood for the count process is given by

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

where $\phi = 1$ and λ_{its} is the prediction mean of count sales occurring for product i from the sth posterior sample for some model of interest and $T_i = \{t | y_{it} > 0\}$, that is, T_i are the set time indices corresponding to sales days for product i over some interval of time. Table 4 provides the pl^c scores for across products and models.

Table 4 reveals some interesting findings. First, the extreme sparsity in the number of nonzero counts makes the count density hard to estimate, and so the improvement over the baseline model is lower than for the previous binary forecasting of the nonzero days. Next, we observe that the model variations of HB^c, BE^c, and HBE^c perform significantly better

TABLE 4

Predictive log-likelihoods pl^c_i scores of the count process fits for the models Base$^c_1$, HB^c, BE^c, *and* HBE^c *along with the benchmark model, for each product. The final two rows show the total pl^c across all products in the test and training sets, respectively. Note that products* 1 *and* 6 *did not have any sales in the test set so do not have a predictive likelihood for the count process*

Product i	$\text{pl}^{z,\text{test}}_{\text{Bench}}$	$\text{pl}^{c,\text{test}}_{\text{Base}_0,i}$	$\text{pl}^{c,\text{test}}_{\text{HB},i}$	$\text{pl}^{c,\text{test}}_{\text{BE},i}$	$\text{p}^{c,\text{test}}_{\text{HBE},i}$
1	–	–	–	–	–
2	−0.8572	−0.8619	−0.8943	**−0.6471**	−0.6752
3	−0.8557	−0.9100	−0.5500	−0.6200	**−0.4800**
4	−0.6131	**−0.2286**	−0.2543	−0.2371	−0.2529
5	−0.5978	**−0.0267**	−0.0233	−0.0267	−0.2200
6	–	–	–	–	–
7	−0.2877	−0.0100	**−0.0000**	−0.0400	−0.2200
8	**−3.5264**	−4.9900	−4.1600	−7.9200	−3.1700
9	**−1.3863**	−2.5400	−1.4000	−1.5000	−1.6000
10	−1.7968	−1.9900	−1.9900	−1.9000	**−1.0200**
11	**−1.7525**	−2.3500	−2.3567	−2.4833	−3.6500
12	−1.2321	−0.5100	−0.5450	−0.5150	**−0.3400**
13	−1.5661	−1.1533	−1.1567	−1.1567	**−0.7767**
14	**−1.7996**	−3.0950	−3.2300	−3.2400	−2.6150
15	−1.1632	−1.0200	−1.0250	−0.9750	**−0.3300**
16	−2.8904	**−1.5700**	−2.6400	−1.6300	−1.8000
17	−0.8267	−0.0333	**−0.0267**	−0.0300	−0.1833
$\text{pl}^{c,\text{test}}_{\text{model}}$	−1.0840	−1.0121	−1.0109	−0.9681	**−0.8740**
$\text{pl}^{c,\text{train}}_{\text{model}}$	−1.2234	−1.1614	−1.559	**−1.0641**	−1.1212

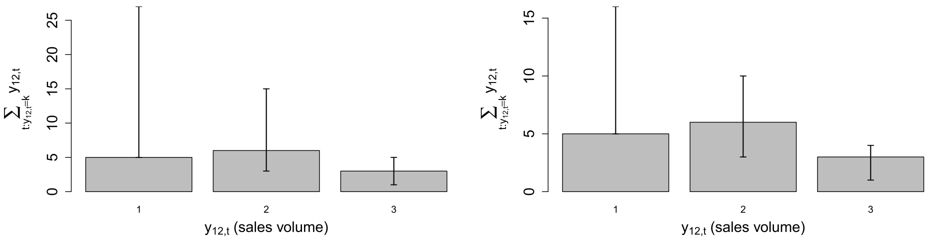

FIG. 7. *Histogram of the observed number of days corresponding to each sale volume for product* $i = 10$, *with the* 95% *credible intervals from the fitted models* HB^c (*left*) *and* HBE^c (*right*). *For example*, *there were exactly five days that had one sale*, *and this is inside the posterior interval for both models. However the* HBE^c *model gives a substantially narrower prediction interval.*

than the Baseline model Base_1^c with no covariates. Similarly to the zero process, Table 4 indicates the count process uniformly benefits from the inclusion of self-excitation in the model variations outlined in Section 5.4.

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 1 indicates, the number of sales that each intermittent demand series has is very small (typically in the order three to 20 sales), and thus it may be expected that information borrowing would particularly benefit the individual models due to the lack of data. An example of this extra strength that comes from the hierarchical model is illustrated by Figure 7. This plot shows a histogram of the number of sales of product 10 in the training set, with corresponding 95% credibility intervals of posterior predictive densities for the models HB^c and BE^c. We observe that the hierarchical model (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, as we can see the aggregate log-predictive likelihood of $\sum_{i=1}^{15} \mathrm{pl}_{\mathrm{model},i}^{c,\mathrm{train}}$ of Table 4 provides more evidence that model HBE^c is the best fitting model.

5.6. *Retail analytics discussion.* The output of models, outlined in Sections 5.1 and 5.4, provides interesting interpretations from a retail analytics perspective. First, we observe that covariate data $\boldsymbol{x}_{it}^z$, $\boldsymbol{x}_{it}$, as specified in 4.1 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.

Though it is hard to tease apart the contribution of each piece of the model to its predictive power for individual products, some patterns emerge. Closer inspection of the results in Table 2 reveals that products with extremely low sales (such as products 10 and 13 which only contain two sales in the test set) essentially do not benefit from a complicated model structure; on the other end, the product with the highest volume of sales (product 1) benefits from the excitation, cross-excitation, and hierarchy of the model. Products with several sales in the test set generally benefit greatly from the excitation component (e.g., products 6, 8 and 11), and a few also take advantage of the cross-excitation component (products 4,5 and 7). Although across the whole dataset the overall best predictive model is the full hierarchical model with both self- and cross-excitation, the choice also depends on specific commercial interest into individual products as well as profit-driven loss functions (Berry, Helman and West (2020)).

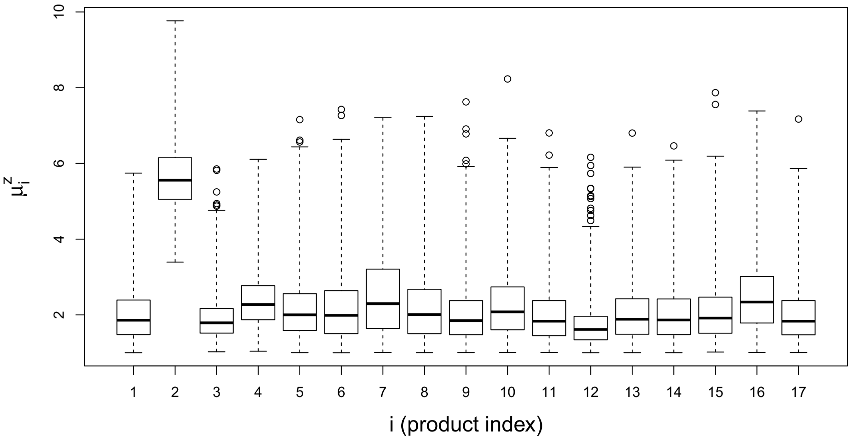

FIG. 8. *Box plots of the posterior distribution of μ_i^c across all products for model* HBE^z. *The μ_i^c estimates being greater than* 2 *indicates the temporal excitation exhibited in that data typically occurs at lags greater than* 1.

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, Ord and Beaumont (2012) and Chapados (2014) where they establish that models incorporating the recent sales history outperform temporally static models. This is important because it ultimately allows retailers to circumvent overstocking 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, Ord and Beaumont (2012) and Chapados (2014). Namely, we establish that the temporal excitation exists even if you condition on the seasonal trends and pricing information of $\boldsymbol{x}_{it}$. 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 8 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 similar) is possibly not enough, compared to the Hawkes process that incorporates the entire history of events.

6. Conclusion.

In this work we introduced 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. Our model specification is interpretable and allows a deeper understanding of the 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. First, 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 into 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 defining empirical cross-excitation groups provided an intuitive and computationally simple method of expressing suspected temporal cross-correlation. Interestingly, the empirical clusters kept all products of "GADGET" brand into the same cluster, pointing to natural interactions between brand and cross-excitation.

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 an a sale occurring within user-defined; it could also be interesting to assess whether the excitation structure could be inferred from the data. Although there are existing models in nonparametric clustering of count time series (Nieto-Barajas and Contreras-Cristán (2014)), cross-excitation signal can only be captured through the full cross-excitation model, which would be computationally prohibitive. A more promising direction may be to relax the assumption of symmetric cross-excitation (i.e., product i may excite product j but not necessarily vice versa) and treating cross-excitation terms through Bayesian variable selection (Tadesse and Vannucci (2021)).

APPENDIX

A.1. Prior formulation. Table 5 specifies the prior structure of the zero process models models Base_1^z, HB^z, BE^z, HBE^z, BEC^z, and HBEC^z. Table 6 specifies the prior structure of the count process models Base_1^c, HB^c, BE^c, and HBE^c.

TABLE 5

Prior formulation of models Base_1^z, HB^z, BE^z, HBE^z, BEC^z, *and* HBEC^z. *We abbreviate* $\mathrm{Normal}(\mu^2, \sigma)$ *and* $\mathrm{Gamma}(\alpha, \beta)$ *to* $N(\mu, \sigma^2)$ *and* $G(\alpha, \beta)$, *respectively*

Parameter	Base_1^z	HB^z	BE^z	HBE^z	BEC^z	HBEC^z
$\varphi_i \sim$	$\mathrm{N}(-3, 3)$					
$\theta_{i1}^z \sim$		$\mathrm{N}(\mu_1^z, 0.05)$	$\mathrm{N}(-3, 0.75)$	$\mathrm{N}(\mu_1^z, 0.05)$	$\mathrm{N}(-3, 0.75)$	$\mathrm{N}(\mu_1^z, 0.05)$
$\theta_{i2}^z \sim$		$\mathrm{N}(\mu_2^z, 0.05)$	$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_2^z, 0.05)$	$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_2^z, 0.05)$
$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\theta_{i20}^z \sim$		$\mathrm{N}(\mu_{20}^z, 0.05)$	$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_{20}^z, 0.05)$	$\mathrm{N}(0, 0.75)$	$\mathrm{N}(\mu_{20}^z, 0.05)$
$\gamma_{i1}^z \sim$			$\mathrm{G}(5, 1)$	$\mathrm{G}(\eta_1^z, 1)$	$\mathrm{G}(5, 1)$	$\mathrm{G}(\eta_1^z, 1)$
$\gamma_{i2}^z \sim$			$1 + \mathrm{G}(1, 2)$	$1 + \mathrm{G}(\eta_2^z, 2)$	$1 + \mathrm{G}(1, 2)$	$1 + \mathrm{G}(\eta_2^z, 2)$
$\gamma_{i3}^z \sim$			$\mathrm{G}(10, 2.5)$	$\mathrm{G}(\eta_3^z, 2.5)$	$\mathrm{G}(10, 2.5)$	$\mathrm{G}(\eta_3^z, 2.5)$
$\tilde{\gamma}_{i1}^z \sim$					$\mathrm{G}(2, 8)$	$\mathrm{G}(\tilde{\eta}_1^z, 8)$
$\tilde{\gamma}_{i2}^z \sim$					$1 + \mathrm{G}(1, 2)$	$1 + \mathrm{G}(\tilde{\eta}_2^z, 2)$
$\tilde{\gamma}_{i3}^z \sim$					$\mathrm{G}(10, 2.5)$	$\mathrm{G}(\tilde{\eta}_3^z, 2.5)$
$\rho_1^z \sim$		$\mathrm{N}(-3, 0.75)$		$\mathrm{N}(-3, 0.75)$		$\mathrm{N}(-3, 0.75)$
$\rho_2^z \sim$		$\mathrm{N}(0, 0.75)$		$\mathrm{N}(0, 0.75)$		$\mathrm{N}(0, 0.75)$
$\vdots$		$\vdots$		$\vdots$		$\vdots$
$\rho_{20}^z \sim$		$\mathrm{N}(0, 0.75)$		$\mathrm{N}(0, 0.75)$		$\mathrm{N}(0, 0.75)$
$\eta_1^z \sim$				$\mathrm{G}(50, 10)$		$\mathrm{G}(50, 10)$
$\eta_2^z \sim$				$\mathrm{G}(10, 10)$		$\mathrm{G}(10, 10)$
$\eta_3^z \sim$				$\mathrm{G}(500, 50)$		$\mathrm{G}(500, 50)$
$\tilde{\eta}_1^z \sim$						$\mathrm{G}(30, 15)$
$\tilde{\eta}_2^z \sim$						$\mathrm{G}(10, 10)$
$\tilde{\eta}_3^z \sim$						$\mathrm{G}(500, 50)$

TABLE 6
Prior formulation of models Base_1^c, HB^c, BE^c, *and* HBE^c

Parameter	Base_1^c	HB^c	BE^c	HBE^c
$\varphi_i^c \sim$	$\mathrm{N}(-4, 4)$			
$\theta_{i1}^c \sim$		$\mathrm{N}(\mu_1^c, 1)$	$\mathrm{N}(1, 0.75)$	$\mathrm{N}(\mu_1^c, 0.05)$
$\theta_{i2}^c \sim$		$\mathrm{N}(\mu_2^c, 1)$	$\mathrm{N}(-1, 0.75)$	$\mathrm{N}(\mu_2^c, 0.05)$
$\gamma_{i1}^c \sim$			$\mathrm{G}(1, 5)$	$\mathrm{G}(\eta_1^c, 5)$
$\gamma_{i2}^c \sim$			$1{+}\mathrm{G}(3, 1)$	$1{+}\mathrm{G}(\eta_2^c, 1)$
$\gamma_{i3}^c \sim$			$\mathrm{G}(4, 1)$	$\mathrm{G}(\eta_3^c, 1)$
$\rho_1^c \sim$		$\mathrm{N}(1, 0.5)$		$\mathrm{N}(1, 0.75)$
$\rho_2^c \sim$		$\mathrm{N}(-1, 0.5)$		$\mathrm{N}(-1, 0.75)$
$\eta_1^c \sim$				$\mathrm{G}(5, 5)$
$\eta_2^c \sim$				$\mathrm{G}(15, 5)$
$\eta_3^c \sim$				$\mathrm{G}(40, 10)$

Acknowledgements. Access to anonymised electronic point of sale data granted by our industrial partner dunnhumby ltd.

Funding. This work has been carried out with the financial support of the EPSRC, the Alan Turing Institute, and dunnhumby ltd, our industrial partner.

REFERENCES

BERRY, L. R., HELMAN, P. and WEST, M. (2020). Probabilistic forecasting of heterogeneous consumer transaction–sales time series. *Int. J. Forecast.* **36** 552–569. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2019.07.007

BLUNDELL, C., BECK, J. and HELLER, K. A. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems* 2600–2608.

CHAPADOS, N. (2014). Effective Bayesian modeling of groups of related count time series. arXiv preprint. Available at arXiv:1405.3738.

DO CROSTON, J. (1972). Forecasting and stock control for intermittent demands. *Oper. Res. Q.* 289–303.

GARDNER, G. S. (2006). Exponential smoothing: The state of the art—part II. *Int. J. Forecast.* **22** 637–666.

FERREIRA, K. J., LEE, B. H. A. and SIMCHI-LEVI, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. *Manuf. Serv. Oper. Manag.*

GHOBBAR, A. A. and FRIEND, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: A predictive model. *Comput. Oper. Res.* **30** 2097–2114.

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410 https://doi.org/10.1093/biomet/58.1.83

HEIDELBERGER, P. and WELCH, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM* **24** 233–245. MR0611745 https://doi.org/10.1145/358598.358630

KOURENTZES, N. (2013). Intermittent demand forecasts with neural networks. *Int. J. Prod. Econ.* **143** 198–206.

LAI, E. L., MOYER, D., YUAN, B., FOX, E., HUNTER, B., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2016). Topic time series analysis of microblogs. *IMA J. Appl. Math.* **81** 409–431. MR3564661 https://doi.org/10.1093/imamat/hxw025

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

SEEGER, M. W., SALINAS, D. and FLUNKERT, V. (2016a). Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, eds.) **29**. Curran Associates, Red Hook.

SEEGER, M. W., SALINAS, D. and FLUNKERT, V. (2016b). Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems* 4646–4654.

MISHRA, P., YUAN, X.-M., HUANG, G. and DUC, T. T. H. (2014). *Intermittent Demand Forecast*: *Robustness Assessment for Group Method of Data Handling*.

MULLAHY, J. (1986). Specification and testing of some modified count data models. *J. Econometrics* **33** 341–365. MR0867980 https://doi.org/10.1016/0304-4076(86)90002-3

NIETO-BARAJAS, L. E. and CONTRERAS-CRISTÁN, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Anal.* **9** 147–169. MR3188303 https://doi.org/10.1214/13-BA852

PORTER, M. D. and WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124. MR2951531 https://doi.org/10.1214/11-AOAS513

POUR, A. N., TABAR, B. R. and RAHIMZADEH, A. (2008). A hybrid neural network and traditional approach for forecasting lumpy demand. *Proc. World Acad. Sci. Eng. Technol.* **30** 384–389.

RANGAPURAM, S. S., SEEGER, M. W., GASTHAUS, J., STELLA, L., WANG, Y. and JANUSCHOWSKI, T. (2018). Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) **31**. Curran Associates, Red Hook.

SAHU, S. K., BAFFOUR, B., HARPER, P. R., MINTY, J. H. and SARRAN, C. (2014). A hierarchical Bayesian model for improving short-term forecasting of hospital demand by including meteorological information. *J. Roy. Statist. Soc. Ser. A* **177** 39–61. MR3158666 https://doi.org/10.1111/rssa.12008

SALINAS, D., FLUNKERT, V., GASTHAUS, J. and DEEPAR, T. J. (2020). Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **36** 1181–1191. ISSN 0169-2070.

SHENSTONE, L. and HYNDMAN, R. J. (2005). Stochastic models underlying Croston's method for intermittent demand forecasting. *J. Forecast.* **24** 389–402. MR2206931 https://doi.org/10.1002/for.963

SNYDER, R. D., ORD, J. K. and BEAUMONT, A. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *Int. J. Forecast.* **28** 485–496.

TADESSE, M. G. and VANNUCCI, M. (2021). *Handbook of Bayesian Variable Selection*. CRC Press, Boca Raton.

ZHOU, K., ZHA, H. and SONG, L. (2013). Learning social infectivity in sparse low-rank networks using multidimensional Hawkes processes. In *AISTATS* **13** 641–649.

STAN DEVELOPMENT TEAM RStan: the R interface to Stan, 2016. R package version 2.14.1. Available at: http://mc-stan.org/.