Dirichlet process mixtures of order statistics with applications to retail analytics

James Pitkin,
University College London, UK

Gordon Ross
University of Edinburgh, UK

and Ioanna Manolopoulou
University College London, UK

[Received October 2016. Final revision May 2018]

Summary. The rise of ‘big data’ has led to the frequent need to process and store data sets containing large numbers of high dimensional observations. Because of storage restrictions, these observations might be recorded in a lossy-but-sparse manner, with information collapsed onto a few entries which are considered important. This results in informative missingness in the observed data. Our motivating application comes from retail analytics, where the behaviour of product sales is summarized by the price elasticity of each product with respect to a small number of its top competitors. The resulting data are vectors of order statistics, because only the top few entries are observed. Interest lies in characterizing the behaviour of a product's competitors, and clustering products based on how their competition is spread across the market. We develop non-parametric Bayesian methodology for modelling vectors of order statistics that utilizes a Dirichlet process mixture model with an exponentiated Weibull kernel. Our approach allows us added flexibility for the distribution of each vector, while providing parameters that characterize the decay of the leading entries. We implement our methods on a retail analytics data set of the cross-elasticity coefficients, and our analysis reveals distinct types of behaviour across the different products of interest.

Keywords: Bayesian non-parametrics; Censoring; Cross-elasticity

1. Introduction

The field of retail analytics is concerned with understanding the purchasing behaviour of consumers, for purposes such as sales prediction, inventory management and coupon personalization (Silver et al., 2013; Rudin et al., 2013; Gunawardana and Shani, 2009; Huh and Rusmevichientong, 2009; Bajari et al., 2015; Ferreira et al., 2015). Retail analytics are a particularly challenging area for machine learning, since vast amounts of data are collected at various operational levels, ranging from individual customer transaction data to aggregated sales data across whole sectors. This means that companies are interested in developing efficient summaries of their data, to mitigate storage and computational costs (Akcay, 2013; Intel, 2014).

Address for correspondence: James Pitkin, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK.
E-mail: j.pitkin.12@ucl.ac.uk
A particularly important example involves the price elasticity coefficients that are generated by sales prediction models. Given a set of products which are offered to consumers, the sales of each product typically depend on both its own price as well as the price of each of its competitors. The cross-elasticity of two products is a measure of the dependence that their prices have on their respective sales. In theory, companies would store this information as a matrix which contains the cross-elasticities for each pair of products. However, in practice, computing price elasticities can be computationally prohibitive when there are a large number of products, so companies often instead make use of highly tailored black box sparse regression sales models (Liu et al., 2013; Beheshti-Kashi et al., 2015) and only measure the cross-elasticity for a small number of each product’s competitors, with the remaining entries of the matrix treated as missing or negligible.

Although the resulting output matrix of coefficients can be successful at providing accurate predictions of future sales, the inherent informative missingness implies that a global interpretation and understanding of the behaviour of the market may not be directly available. Here we are particularly interested in clustering groups of similar products together according to the distribution of their competition in the market, but also in identifying products for which potentially important competitors may have been missed out of the observed cross-elasticity matrix.

Formally, the form of 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 to only the top few entries. Mathematically speaking, the data are in the form

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

where $x_i$ is the cross-elasticity vector of dimension $n$ for product $i$, which has $l_i$ uncensored ordered entries, with the remaining being censored. Heterogeneity among products stems both from the rate in which competition decays, as well as from the actual number of uncensored cross-elasticities. Existing price sensitivity analyses focus on reducing these vectors to summary statistics of elasticity coefficients from fitted demand models (Andreyeva et al., 2010; Oliveira et al., 2007). To our knowledge, there is currently no methodology to cluster these entire vectors with minimal information loss, which would allow flexibility in handling the varying lengths of cross-elasticity vectors as well as incorporating the censoring mechanism to provide information about the censored entries.

In this paper we develop non-parametric Bayesian models by interpreting our observed data as realizations of variable length order statistics sequences. We shall show that this succinctly handles partial censoring and enables computationally straightforward inference on the unobserved entries of the cross-elasticity matrix. Our approach uses tools from survival analysis to address inherent censoring mechanisms, together with non-parametric Bayesian Dirichlet process mixture models (DPMMs) that enable products to be clustered into distinct groups for analysis. Using the exponentiated Weibull (EW) distribution as a mixture kernel (Mudholkar and Srivastava, 1993), we can account for both light and heavy tail behaviour that is apparent in the data. As we shall discuss later, the EW distribution has several unique properties which make it ideal for modelling order statistics. We develop efficient sampling mechanisms by adapting algorithm 8 of Neal (2000) and provide interpretation and visualization tools for summarizing and presenting the output. Our approach fully characterizes sales sensitivities by incorporating all the information from the entire cross-elasticity vector, offering two distinct benefits. Firstly, by interpreting these elasticity vectors as order statistic sequences we can directly cluster products in terms of their entire cross-elasticity vectors and conveniently handle their varying length
nature. Secondly, it provides a framework for predicting censored entries which can shed light on potentially important competitors which have been censored.

Although we focus on the retail analytics application, our methodology is general and is relevant in any situation with informative missingness where only the top few order statistics of each observation vector are observed. This includes applications such as sports analytics, where only the top few performances, athletes or teams are observed (Malcata and Hopkins, 2014) or the stylometry analysis of literary texts which often focuses on analysing the frequencies of the top few most common words (Narayanan et al., 2012).

The rest of the paper is organized as follows: Section 2 outlines our motivation for clustering elasticity coefficients in the retail analytics setting. Section 3 covers the properties of uniform order statistics that are relevant to our model and reviews the relevant literature, in particular drawing parallels to survival and reliability analysis. Section 4 provides a background of the pertinent characteristics of the EW distribution and its relevance as a kernel to variable length order statistics sequences. Section 5 covers the theory of DPMMs and further provides the non-parametric mixture model of variable length order statistics sequences along with prior distributions specification. We outline the algorithm that is used in posterior inference in Section 6. Section 7 illustrates our methods on a simulated and a real retail analytics data set of cross-elasticity vectors. dunnhumby ltd, a customer science company, allowed us access to the anonymized cross-elasticity coefficient output of a set of products derived from the loyalty card transactions of a leading UK supermarket retailer. Section 8 provides a summary of our methods with potential extensions and applications of the work.

The programs that were used to analyse the data and some synthetic data can be obtained from

https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets

2. Motivation

It is common in retail analytics to characterize products on the basis of how sensitive their sales are to the prices of their competitors and how customers interact with their product range. Analytics teams are constantly striving to develop models and inference methods that provide insight into understanding how price fluctuations that propagate throughout stores will impact the sales of products whose prices have not changed (Persson, 1995; Ferreira et al., 2015). Clustering products on the basis of their price sensitivity profile can provide a segmentation of a retailer's product range. This ultimately aids store planners in deciding on the value of a given pricing or display combination, as it provides information on how a product’s sales are likely to react to the deviations of prices of other products. For instance, a poor display combination could be one that consists entirely of products that are characterized by their sales being primarily driven by the prices of its competition. This would lead to margin cannibalism—where profit made on one product is offset by the loss of profit of another product. The information of product clusterings would enable better pricing and display optimization.

One approach to such a sales sensitivity analysis is to cluster products in terms of their direct and cross-elasticity coefficients. Existing work on analysing sales sensitivities of retail goods, and using these to partition product ranges, has focused on defining summary statistics which capture many of the important aspects of price sensitivity profiles (Andreyeva et al., 2010). For example, Oliveira et al. (2007) investigated the heterogeneity in direct (but not cross-) elasticities across products and across consumer groups. Similar work investigated variations in category
level summaries of cross-elasticities coefficients and the effect of sales sensitivity across store, demographic and product category levels (Hoch et al., 1995; Guerrero-López et al., 2017). However, much of the important information on price sensitivity profiles lies in the entire cross-elasticity vectors and cannot be captured in summary statistics. The data set that we have access to through dunnhumby ltd comprises cross-elasticity vectors for a set of products from a leading UK supermarket chain. Although the precise mechanics of how these estimates are obtained are highly engineered within their proprietary model, the general form of the model is given by refined versions of the Working–Leser regression (Working, 1943; Leser, 1963):

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

(1)

where, for each product $i$ and discrete time $t$ (in days), $S_{it}$ denotes its sales, $Q_{it}$ its price, $P_{ij,t}$ the price of its $j$th competitor product, $\varphi_i$ its direct elasticity and $\eta_{ij}$ product $j$’s relative cross-elasticity with product $i$ (as a multiple of the direct elasticity). The term $\epsilon_{i,t}$ represents time-dependent error. We use the notation $1:n_i$ to denote the set $1, \ldots, n_i$. The function $f(\cdot)$ involves data aggregation and seasonality patterns that are relevant to retail sales, as well as additional information on display combinations and promotions; our focus here is not on this function, but rather on the post-processing of the output of the regression model. Here $n_i$ is the number of competitor products of product $i$, which are preselected by using expert knowledge that is encoded in an algorithm, to avoid using the entire set of products which is computationally prohibitive because of the complexity of $f(\cdot)$. For the purposes of this study and to ease notation in later sections, we assume that competitor products are labelled such that product $i$’s cross-elasticity coefficients $\eta_{ij}$ are increasing in magnitude and that all products have the same potential number of competitors, i.e. $n_i = n_i = 1, \ldots, N$. The cross-elasticity coefficients are estimated by using shrinkage methods for sparsity, so that only $l_i \eta$s are non-zero, with the remaining exactly equal to 0. Table 1 provides two toy examples of variable length order statistic sequences in the context of cross-elasticities.

Fitting this model to data provides us with a vector of cross-elasticities for each product, where some of the entries may be 0 because of sparsity. Our goal is then to cluster products

### Table 1. Ordered elasticity output $\varphi$ and $\eta$ for two fictional products, Bobby’s puffs and Lucan’s crisps†

| Relevant competitor       | Elasticities for Bobby’s cheesy puffs |        | Elasticities for Lucan’s salted crisps |        | Relevant competitor               |
|---------------------------|--------------------------------------|--------|---------------------------------------|--------|-----------------------------------|
| $\varphi_i$               | Bobby’s puffs                        | $-1.41$| $-1.86$                               | Lucan’s salted crisps |
| $\varphi_i \eta_{i6}$    | Supermarket puffs                    | $-1.12$| $0.8$                                 | Sussex’s chives crisps  |
| $\varphi_i \eta_{i5}$    | Harry’s puffs                        | $-1.10$| $0.44$                               | Chef’s paprika crisps  |
| $\varphi_i \eta_{i4}$    | Supermarket nuts                     | $-0.80$| $0.44$                               | Supermarket puffs     |
| $\varphi_i \eta_{i3}$    | Bobby’s tortillas                    | $-0.48$| $-0.10$                              | Lucan’s nuts           |
| $\varphi_i \eta_{i2}$    | Tommy’s chips                        | $-0.35$| $-0.40$                              | Harry’s popcorn       |
| $\varphi_i \eta_{i1}$    | Tommy’s puffs                        | $-0.05$| $0$                                  | Chef’s BBQ crisps     |

†For each product we have columns of order elasticity coefficients $\varphi_i$ and $\varphi_i \eta_{ij}$ along with the respective sequences of $\eta_{ij}$, which demonstrates the decreasing nature of data from model (1). 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.
according to these cross-elasticity vectors. Although, in theory, one can perform clustering alongside the regression, this is computationally prohibitive in the current context because of the highly tailored model fitting that is involved, so we treat the regression fitting as ‘black box’ fitting and work with the cross-elasticity vectors directly. To address the fact that, because of computational limitations, competitor products are preselected by using expert knowledge and are subject to error, we treat the 0 entries as missing minor competitors (with smaller cross-elasticity coefficients than the observed coefficients). This results in a clustering framework whereby observation vectors have different numbers of non-missing entries.

Since cross-elasticity vectors arise as the outcome of penalized regression, it is natural to assume that coefficients are shrunk to 0 as the result of a penalization threshold. For example, in the simplest case of best-subset selection with an orthonormal design matrix, non-zero coefficients are exactly equal to the top order statistics of the corresponding ordinary least squares estimates. With this in mind, we treat the observed non-zero cross-elasticity coefficients as the top order statistics of an underlying vector of length \( n \). The \( l \) observed coefficients of the leading product competitors are thus modelled as the top \( l \) order statistics of a set of \( n \) independent and identically distributed (IID) observations from an unknown underlying distribution. To account for the fact that a different number \( l \) of entries may be observed in each vector, we assume that \( l \) also follows a probability distribution, independently of the actual entries.

3. Order statistics of continuous distributions

The order statistics of a random sample are the \textit{reordered} observations in terms of increasing size. More concretely, given a continuous distribution variable \( X \) and observations \( x_1, \ldots, x_n \sim \text{IID } X \), the order statistics \( x_{(1)}, \ldots, x_{(n)} \) are given by

\[
x_{(1)} < x_{(2)} < \ldots < x_{(n)}. \tag{2}
\]

The \( j \)th order statistic of order (2) 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)

\[
f_{(j)}(x) = n f(x) \binom{n-1}{j-1} F(x)^{j-1} (1 - F(x))^{n-j}. \tag{3}
\]

One of our key modelling assumptions is that a partially observed cross-elasticity vector of length \( n \) with \( l \) non-zero entries in fact corresponds to the top \( l \) order statistics of a random sample of size \( n \). We term each of these vectors of the top \( l \) order statistics as reordered variable length order statistics sequences, and denote them as \( \mathbf{x} = (x_{(n)}, \ldots, x_{(n-\l)}) \). We also denote the \( j \)th order statistic of sequence \( \mathbf{x} \) by \( x_{(j)} \). The density of \( \mathbf{x} | l \) is denoted as \( f_{(n):(n-l+1)} \) and is given by

\[
f_{(n):(n-l+1)}(\mathbf{x}|l) = f_{(n):(n-l+1)}(x_{(n)}, \ldots, x_{(n-l+1)} | l) = \frac{n!}{(n-l)!} F(x_{(n-l+1)})^{n-l} \prod_{j=1}^{l} f(x_{(n+j-l)}). \tag{4}
\]

By the independence of \( x_{(n-j)} | x_{(n-j+1)} \perp x_{(n)}, x_{(n-1)}, \ldots, x_{(n-j+2)} \) and by equation (4), the density of the conditional distribution of \( x_{(n-j)} | x_{(n-j+1)}, l \) for \( j < l \) (denoted as \( f_{(n-j):(n-j+1)} \)) is given by

\[
f_{(n-j):(n-j+1)}(x_{(n-j)} | x_{(n-j+1)}, l) = (n-j) f(x_{(n-j)}) \frac{F(x_{(n-j)})^{n-(j+1)}}{F(x_{(n-j+1)})^{n-j}}. \tag{5}
\]
and thus the density of the joint sample $x|l$ can also be expressed in hierarchical format:

$$f_{(n):(n-l+1)}(x|l) = f(x(n)) \prod_{j=1}^{l-1} f_{(n-j)(n-j+1)}(x(n-j)|x(n-j+1), l). \quad (6)$$

Finally, the joint distribution of $x$ can be combined with the $n-l$ zero entries of $x$ through

$$f(x) = p(l) f_{(n):(n-l+1)}(x|l), \quad (7)$$

where $p(l)$ is the probability mass function over the length of the sequence. Here we assume that $l$ and the magnitude of the non-zero entries of $x$ are independent.

Much work has been done in the study of the theoretical properties of order statistics (Beutner and Kamps, 2009) and has been applied to areas such as modelling software reliability (Wilson and Samaniego, 2007), reliability of propulsion systems of aircraft (Warr and Collins, 2014) and recommender systems (Caron and Teh, 2012). A particularly 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 $k$ out of $n$ framework, a standard assumption is that each sequence or system produces the same marginal order statistic, whereas we would like to allow additional flexibility.

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 that is 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 (2006) and Kottas (2006) used gamma and Weibull kernels within a DPMM framework respectively. The EW distribution was shown to be the first 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.

4. 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 EW distribution. A random variable $X$ is distributed according to the EW distribution, denoted as $X \sim \text{EW}(\alpha, \beta, \lambda)$, if its probability density and distribution function are given by

$$f(x) = \alpha \beta \lambda^\beta x^{\beta-1} \{1 - \exp\{-(\lambda x)^\beta\}\}^{\alpha-1} \exp\{-(\lambda x)^\beta\} \quad (8)$$

and

$$F(x) = [1 - \exp\{-(\lambda x)^\beta\}]^\alpha \quad (9)$$

respectively, where $x > 0$, $\lambda > 0$, $\beta > 0$ and $\alpha > 0$. 

Table 2. EW density behaviours for various combinations of $(\alpha, \beta, \lambda)$

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

Fig. 1. EW density for $(\alpha, \beta, \lambda) = (1.2, 0.8, 1.0)$ (——), $(1.55, 0.8, 1.0)$ (— —), $(0.24, 5.0, 1.0)$ (— — —), $(1.8, 1.4, 0.5)$ (— — —)

The EW distribution is an extension to the standard Weibull distribution through the inclusion of the additional parameter $\alpha$, which enables the distribution to have a wide range of tail behaviours. Similarly to the Weibull distribution, $\lambda$ is a scale parameter whereas $\beta$ controls the tail behaviour of the distribution; distributions are heavy tailed for $\beta < 1$ and light tailed otherwise. Furthermore, decreasing $\beta$ monotonically increases the mean and variance, kurtosis and skew of the EW distribution. The effect of $\alpha$ depends on both the value $\alpha \beta$ and whether $\alpha < 1$; increasing $\alpha$ increases symmetry around the mean and mode. These different modal, asymptotic and tail behaviours (Nassar and Eissa, 2003) are summarized in Table 2. Fig. 1 demonstrates various density plots for differing combinations of $(\alpha, \beta, \lambda)$; various asymptotic, modal and tail behaviours are observed.

4.1. Exponentiated Weibull 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 (4) under the EW distribution for fixed order sequences of lengths $l$ is given by
Fig. 2. (a) Realizations of order statistics sequences with EW($\alpha$, $\beta$, $\lambda$) kernel for combinations ($\alpha$, $\beta$, $\lambda$) = (0.2, 0.6, 0.7) (---), (0.5, 1.5, 1.5) (--), (4, 5, 1.5) (---) and (b) density plots of $f_k(x)$ with EW(0.5, 1.5, 1.5) kernel for orders $k=10$ (---), 9 (--), 8 (---).

\[ f_{(n):(n-l+1)}(x|l) = \frac{n!}{(n-l)!}[1 - \exp\left\{-(\lambda x_{n-(l-1)})^{\beta}\right\}]^{\alpha(n-l)} \prod_{j=1}^{l} f(x_{n+j-l}) \]  

(10)

where $f$ is the EW density function (8).

The EW distribution handles censoring naturally, since the censored, joint and conditional densities under the EW distribution belong to the same family. For example, if $x_i \sim \text{IID} \ EW(\alpha, \beta, \lambda)$ for $i = 1, 2, \ldots, n$, then $x_{(j)} \sim \text{EW}(j\alpha, \beta, \lambda)$. Similarly, the conditional distributions $x_{(n-j)}|x_{(n-j+1)} < x_{(n-j)} \sim \text{EW}(j\alpha, \beta, \lambda), 1 \leq j \leq n-1$, 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 distribution can account for both light and heavy tails, enabling us to capture different types of decay behaviours of the elasticity vectors. Fig. 2 provides some examples of order statistics sequences, which demonstrate various decay behaviours and tail behaviours that can be produced under the EW kernel.

5. Non-parametric mixture model of variable length order statistics sequences

As outlined in Section 2, our ultimate goal is to cluster variable length order statistics sequences, here arising through the behaviour of the cross-elasticity vectors of different products. For this, we use the EW distribution as a representation of cross-elasticity decay behaviour. However, to account for different behaviour across products, we additionally cluster products that potentially correspond to the same EW distribution. We opt for a mixture modelling framework to allow fully model-based uncertainty to propagate through the clustering inference. To account for an unknown number of underlying components, we use a Bayesian non-parametric mixture modelling formulation, guiding the number of clusters through a prior on the base distribution of a component. We thus model the entire set of cross-elasticity vectors non-parametrically as a DPMM (Antoniak, 1974).

A DP is a distribution over random probability measures (Ferguson, 1973), parameterized by base distribution $G_0$ over a measurable space $\Theta$, and concentration parameter $\nu$, denoted as $G \sim \text{DP}(\nu G_0)$. Realizations from $\text{DP}(\nu G_0)$ are centered near the base distribution $G_0$, i.e., for any measurable set $A \subset \Theta$, $\mathbb{E}[G(A)] = G_0(A)$. The concentration parameter $\nu$ controls the degree to which realizations from $\text{DP}(\nu G_0)$ are close to $G_0$. Sethuraman (1994) established a convenient formulation of a DP known as a stick breaking construction, which expresses a distribution $G \sim \text{DP}(\nu G_0)$ as
Dirichlet Process Mixtures of Order Statistics

\[ G = \sum_{j=1}^{\infty} \pi_i \delta_{\theta_i}, \]  
\[ \beta_i \sim \text{beta}(1, \nu), \]  
\[ \pi_i = \beta_i \prod_{j=1}^{i-1} (1 - \beta_j), \]  
(11)

where \( \delta_x \) is the Dirac measure of mass centred at \( x \) and \( \theta_i \sim \text{IID} \ G_0 \). A DPMM was proposed by Antoniak (1974) as a mixture model with a DP prior over the random mixing distribution. A DPMM can be expressed hierarchically:

\[ x_i | \theta_i \sim \pi(x_i | \theta_i), \quad i = 1, \ldots, N, \]  
\[ \theta_i | G \sim G, \]  
\[ G | \nu, \omega \sim \text{DP}(\nu G_0), \quad G_0 = G_0(\cdot | \omega), \]  
\[ \nu, \omega \sim \nu G_0 / G, \]  
(12)

where \( \pi \) is the response distribution of \( x \) and \( F_1 \) and \( F_2 \) are independent priors of parameters \( \nu \) and \( \omega \) respectively. The unique values of a vector \( \theta \) are referred to as \( \theta^* \) and we use \( \theta^*_{C_i} \) to denote \( \theta_i \) given an allocation of observation \( i \) to cluster \( C_i \). DPMMs can be used to estimate a density \( p(x) \) by a countably infinite mixture of kernels functions by placing a DP prior over the mixing distribution. Thus given a density \( p(x) \) to estimate, and the family of density \( f(x|\theta) \) parameterized by \( \theta \), then

\[ p(x) = \int f(x|\theta) \, dG(\theta) \]  
\[ = \sum_{i=1}^{\infty} \pi_i f(x|\theta^*_i) \]  
(13)

since \( G \) admits the representation \( G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \) with \( G \sim \text{DP}(\nu G_0) \).

5.1. The model

We now propose a DPMM of variable length order statistics sequences on mixtures of distributions satisfying equation (10). Placing a \( \text{DP}(\nu G_0) \) on the distributions (10) is an attractive approach to handling the complex multimodalities, decay rates and variable lengths that order statistics sequences can exhibit as discussed in Section 4. Thus, the DPMM of variable length order statistics sequences can be expressed in hierarchical format (6) by

\[ \nu \sim \text{gamma}(\tau_1, \tau_2), \]  
\[ G | \nu \sim \text{DP}(\nu G_0), \]  
\[ (\alpha_i, \beta_i, \lambda_i, w_i) | G \sim G, \]  
\[ l_i \sim 1 + \text{binomial}(n - 1, w_i), \]  
\[ x_{i,j} \sim \text{EW}(\alpha_i, \beta_i, \lambda_i), \quad j = 1, \ldots, n, \]  
(14)

where \( i = 1, 2, \ldots, N \) are the numbers of observations for each observation vector \( i \), with all except the top \( l_i \) entries being censored. The final line of expression (14) can also be expressed through the iterative formulation

\[ x_{i,(n-j)} | x_{i,(n-j+1)}, x_{i,(n-j+1)} \sim \text{EW}(n \alpha_i, \beta_i, \lambda_i), \]  
\[ x_{i,n} \sim \text{EW}(n \alpha_i, \beta_i, \lambda_i), \]  
(15)

which follows from equation (6). We treat the lengths \( l \) and observations \( x_{i,j} \) of \( x \) as indepen-
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.

The base distribution \( G_0 \) is a key aspect of the 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 the prior distribution

\[
G_0(\alpha, \beta, \lambda, w) = \text{gamma}(\alpha | \alpha^1, \alpha^2) \text{gamma}(\beta | \beta^1, \beta^2) \text{gamma}(\lambda | \lambda^1, \lambda^2) \text{beta}(w|a, b).
\]  

(16)

The hyperparameters \((a, b, \alpha^1, \alpha^2, \beta^1, \beta^2, \lambda^1, \lambda^2)\) are treated as fixed, chosen depending on the modelling context and reflecting prior expertise. The prior for \( \nu \) is assumed to be \( \gamma(\tau_1, \tau_2) \), allowing the relationship

\[
\mathbb{E}[N^*|\nu] = \nu \log \left( \frac{\nu + N}{\nu} \right)
\]

(Escobar and West, 1995) (where \( N^* \) is the number of occupied clusters) to inform our prior expectation of the number of clusters.

6. Posterior inference

We now present an efficient Markov chain Monte Carlo (MCMC) procedure for obtaining samples from the posterior of \( p(\alpha, \beta, \lambda, w, \nu|X) \) from the model that is proposed by expression (14) with

\[
X = \{x_{i,1:n} : x_{i,n-l_i+1} \leq x_{i,n-l_i+2} \leq \ldots \leq x_{i,n}, \text{with } x_{i,j} \text{ censored to 0 for } 1 \leq j \leq n - l_i \},
\]

where \( x_i \) includes the variable length order statistics sequence of length \( l_i \) (uncensored ordered entries), with the remaining \( n - l_i \) being censored. There are three steps to obtaining samples from \( p(\alpha, \beta, \lambda, w, \nu|X) \) for each MCMC iteration: sampling the atoms \((\alpha, \beta, \lambda, w)\) of the DP(\( \nu G_0 \)) for each order statistics sequence, sampling the clusterwise atoms for each of the unique clusters (as induced by DP(\( \nu G_0 \))) and, finally, sampling the \( \nu \) scale parameter. We initiate by using the Polya urn exposition of a DP (Blackwell and MacQueen, 1973) by taking a Gibbs sample of \( \theta_i = (\alpha_i, \beta_i, \lambda_i, w_i) \) atoms associated with observation \( x_i \) by using

\[
p(\theta_i|\theta_{-i}, \nu, X) = q_0^* H_i + \sum_{k=1}^{N^*_i} q_k^* \delta_{\theta_{ki}}
\]

(17)

where \( q_0^* \propto \nu \int f(x_i|\theta)G_0(d\theta) \) and \( q_k^* \propto N_k^* f(x_i|\theta_k^*, \nu) \) subject to \( \sum_{k=1}^{N^*_i} q_k^* = 1 \). Here \( f(x_i|\theta) = f_{(a);(n-l_i+1)}(x_i|l_i, \alpha, \beta, \lambda)p(l_i|w) \), where \( f_{(a);(n-l_i+1)} \) is specified in equation (10) and the conditional distribution \( p(l_i|w) = (l_{i-1})^{n-1}(1-w)^{n-l_i} \) is the posterior distribution for \( \theta \) based on the prior distribution \( G_0(\alpha, \beta, \lambda, \omega) \) with likelihood \( f(x_i|\theta, \nu) \). Here \( \theta_{-i} \) denotes the vectorized atoms of \( \theta \) excluding the \( i \)th atom \( \theta_i \), \( \{\theta_1^*, \ldots, \theta_{N^*_i}^*\} \) denotes the unique values of \( \theta_i \), \( N^*_i \) the number of unique clusters that are induced by the DP and \( N^*_i \) the number of points that are assigned to atom \( \theta_{ki}^* \). 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, samples of \( \theta^j \sim \text{IID} G_0 \) for \( j = N^* + 1, \ldots, N^* + c \) are drawn, which then reduces equation (17) to taking a sample from the multinomial distribution given by \( P(\theta = \theta^*_j | \theta_{-i}, x_i, \theta_1^*, \ldots, \theta_{N^*_i+c}^*) \), which corresponds to
The number of auxiliary components $c$ that is chosen determines the level to which $q_{0}^{*}$ is approximated. Finally, the $\theta_{k}$ atoms are then updated for each of the unique clusters $k = 1, \ldots, N^{*}$ to avoid inefficiencies that are associated with having to pass through extremely low probability states to reach higher probability states. This is achieved by taking a single sample from the posterior $p(\theta_{k} | \nu, x_{(i:C_i=k)})$ for each $k = 1, \ldots, N^{*}$. As taking exact samples from $p(\theta_{k} | \nu, x_{(i:C_i=k)})$ is intractable for our choice of kernel (10) and prior $G_{0}$ (16), the Metropolis–Hastings algorithm is used to sample from $p(\theta_{k} | \nu, x_{(i:C_i=k)})$ for $k = 1, \ldots, N^{*}$. This involves taking sufficient burn-in samples until convergence to the stationary posterior distribution is satisfactory, at which point $\theta_{k}$ is then taken as the last sample from the Metropolis–Hastings procedure. Finally, $\nu$ is updated in line with Escobar and West’s (1995) auxiliary variables approach. Further details of the posterior inference steps are included in Appendix A.

7. Data examples

We now illustrate how our methodology works in practice by performing two simulation studies, before proceeding to a real world retail analytics data set. In the first example (Section 7.1) we generate data from our model. In the second example (Section 7.2) we generate data by using a gamma distribution as a kernel (rather than an EW distribution). We then fit our model to generate data from our model. In the second example (Section 7.2) we generate data by using vague priors.

7.1. Simulated data 1

We generate data by using parameters for the mixtures (18) 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 500 samples from the following DPMM of variable length order statistics sequences (14):

$$
G = 0.4\delta_{\theta_{1}} + 0.35\delta_{\theta_{2}} + 0.25\delta_{\theta_{3}},
$$

$$
(\alpha_{i}, \beta_{i}, \lambda_{i}, w_{i}) | G \sim G, \quad i = 1, \ldots, 500,
$$

$$
l_{i} \sim 1 + \text{binomial}(19, w_{i}), \quad i = 1, \ldots, 500,
$$

$$
x_{i,j} \sim \text{EW}(\alpha_{i}, \beta_{i}, \lambda_{i}), \quad j = 1, \ldots, 20,
$$

where $\theta = (\alpha, \beta, \lambda, w)$, with $\theta_{1}^{*} = (0.15, 0.8, 0.91, 0.65)$, $\theta_{2}^{*} = (2.5, 3.3, 0.35, 0.75)$ and $\theta_{3}^{*} = (0.64, 1.7, 0.4, 0.9)$, and within each observation vector $i$ with all except the top $l_{i}$ entries being censored.

7.2. Simulated data 2

This simulated example differs from the first simulation study in that the data are simulated from a mixture of gamma distributions rather than a mixture of 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 to 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 $\theta_{1}^{*}$, $\theta_{2}^{*}$ and $\theta_{3}^{*}$ are selected to produce simulated mixtures that imitate the mixtures (18).
We generate 500 samples from the following DPMM of variable length order statistics sequences:

\[
G = 0.4\delta_{\theta_1} + 0.35\delta_{\theta_2} + 0.25\delta_{\theta_3}, \quad \text{and} \quad \theta_i \sim \text{gamma}(\alpha, \beta), \quad i = 1, \ldots, 500, \quad \text{where} \quad \theta^*_1 = (0.15, 0.5, 0.65), \quad \theta^*_2 = (1.7, 1, 0.75) \quad \text{and} \quad \theta^*_3 = (32, 10, 0.9), \quad \text{and} \quad w_i \sim \text{binomial}(19, w_i), \quad i = 1, \ldots, 500, \quad j = 1, \ldots, 20,
\]

(19)

where \(\theta = (\alpha, \beta, w)\), with \(\theta^*_1 = (0.15, 0.5, 0.65)\), \(\theta^*_2 = (1.7, 1, 0.75)\) and \(\theta^*_3 = (32, 10, 0.9)\), and within each observation vector \(i\) with all except the top \(l_i\) entries being censored.

7.3. Prior distributions and posterior sampling

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

\[
(\alpha, \beta, \lambda, w) \sim \text{gamma}(1, 0.1) \quad \text{gamma}(1, 0.1) \quad \text{gamma}(1, 0.1) \quad \text{beta}(1, 1), \quad \nu \sim \text{gamma}(1, 1)
\]

respectively. The priors for \(\alpha\), \(\beta\) and \(\lambda\) 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.

We use the steps that were outlined in Section 6 for parameter inference and perform 10000 MCMC iterations with 200 burn-in, and we thin every 10 samples after the subsequent burn-in samples have been discarded. We present the MCMC output based on the inference methodology of Section 6 on the simulated data of Sections 7.1 and 7.2. Since individual clusters are not identifiable (up to permutations), an additional identifiability criterion is required to perform clusterwise inference. We implement algorithm 2 of Lau and Green (2007) on the posterior samples to derive an ‘optimal’ partition. Lau and Green (2007) selected the partition of the observations into clusters \(C^*\) (with some permutation) which minimize a linear loss function of the posterior expected loss of the posterior marginal coincidence probabilities. This is equivalent to maximizing

\[
l(C^*, K) = \sum_{(i, j) \in M} \mathbb{I}(C^*_i = C^*_j)(\rho_{ij} - K)
\]

(20)

where \(M = \{(i, j) : i < j; i, j \in \{1, \ldots, N\}\}\), \(K = b/(a + b) \in [0, 1]\) where \(b\) is the penalty of misclassifying two points into different clusters (when they should be) and \(a\) is the penalty of misclassifying two points into the same cluster (when they should not be), \(C\) is a given clustering of the observations (up to permutation) and \(\rho_{ij}\) is the posterior coincidence probability between points \(i\) and \(j\). This optimal partition defines cluster assignments \(C^*_i\) for each observation \(x_i\) and is used in future sections to compute clusterwise point estimates of various quantities of interest. As equation (20) needs to be maximized over \(K\), we maximize equation (20) over each \(K \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}\) and select \(K\) and \(C^*\) that give the maximal value.

The partition \(C^*\) that is obtained through the above algorithm determines an optimal number of clusters and an optimal allocation of each observation into a cluster. Posterior distributions of clusterwise parameters are then obtained by sweeping through MCMC samples and, at each iteration, averaging over all parameter values that are associated with each observation within a cluster of \(C^*\).

Figs 3 and 4 present plots of the MCMC output which include the histogram of the number of occupied clusters \(N^*\), a heat map of the posterior marginal coincidence probabilities (cluster co-
Fig. 3. (a) Posterior probability cluster co-membership probability heat map and (b)–(f) histograms and density estimates for $N^*$, $l$, and a few order statistics of simulated data (18).
Fig. 4. (a) Posterior probability cluster co-membership probability heat map and (b)–(f) histograms and density estimates for $N^*$, $l$ and a few order statistics of simulated data (19): these density plots demonstrate that the EW kernel successfully describes the mixture of decay sequences although the data are generated from a mixture of gamma distributions.
Table 3. Posterior means and (2.5%, 97.5%) credible intervals for the parameters $(\alpha, \beta, \lambda, w)$ of each cluster from simulated data (18)†

| Parameter | Results for cluster 1 | Results for cluster 2 | Results for cluster 3 |
|-----------|-----------------------|-----------------------|-----------------------|
| True $(\alpha, \beta, \lambda, w)$ | (0.15, 0.80, 0.91, 0.65) | (2.5, 3.3, 0.35, 0.75) | (0.64, 1.7, 0.40, 0.90) |
| $\alpha$ | 0.14 (0.12, 0.17) | 5.2 (1.4, 19) | 0.61 (0.52, 0.73) |
| $\beta$ | 0.82 (0.71, 0.94) | 2.9 (1.7, 4.5) | 1.7 (1.6, 1.9) |
| $\lambda$ | 0.99 (0.80, 1.2) | 0.41 (0.30, 0.64) | 0.40 (0.36, 0.44) |
| $w$ | 0.66 (0.65, 0.68) | 0.75 (0.68, 0.81) | 0.91 (0.89, 0.92) |
| $N$ | 208 | 166 | 126 |

†The top and bottom rows show the true parameters and number of observations assigned to each cluster respectively.

membership) and density estimates of the order statistics sequences $x_i$. The maximum a posteriori estimate for the number of clusters is indeed 3, with a clear separation of the observations into three clusters, and the marginal density estimates closely match the corresponding histograms of the data.

Table 3 summarizes our MCMC output for the simulated study in section 7.1 using the clusters that are defined by $C^*$. The estimates are very close to the true parameter values, which are also contained within the 95% credibility intervals, indicating that the inference is working effectively.

7.4. Retail analytics data set
We apply our method of order statistics clustering to a retail analytics data set from a leading UK supermarket chain. Access to the anonymized data was provided by dunnhumby. The data set consists of the cross-elasticities for a category of supermarket products of the format that was described in Section 2:

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

where we have observed only the top $l_i$ order statistics of each cross-elasticity vector $\eta_i$. To enable a straightforward interpretation we focus on the snacks category which consists of $N = 275$ (out of thousands) products, so that observations consist 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 onion, 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. Fig. 5 shows summary plots for the snacks category in this study, although other categories will show different behaviour. Specifically, we show histograms of the lengths $l_i$ of $\eta_i$ as well as the top two terms of the sequences. The histogram of the top order statistics demonstrates spikes centred near 0.0 and 1.0, suggesting possible multimodality.

7.4.1. Omitted competitors and aggregate competition
We introduce two statistics that are relevant to the retail analytics setting: omitted competitors and mean aggregate competition. These notions have key interpretations in the retail analytics context and will enable us to assess model fit.
Fig. 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.

7.4.1.1. Definition 1: omitted competitors. As discussed in Section 2, censoring of lower order statistics in the cross-elasticity vector occurs through penalized regression. However, 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 statistic OC 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 $\eta_i$) and assessing whether this predicted value is sufficiently large. Concretely, we say that an elasticity vector contains omitted competitors if its variable length order statistics sequence satisfies

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

for some truncation constant $\epsilon > 0$ and where $\tilde{\eta}_{n-I}$ represents the random quantity of the $(n-l_i)$th order statistic of $n$ IID EW($\alpha, \beta, \lambda$) samples with $I \sim 1 + \text{binomial}(n-1, w)$. In other words, $\tilde{\eta}$ has the same distribution as $\eta$, but without any censoring. Thus the OC-statistic represents the expected value of the first censored term of a cross-elasticity vector $\tilde{\eta}$, were we to have observed it. The value of $\epsilon$ 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 it will be fixed for our subsequent
Dirichlet Process Mixtures of Order Statistics

19

analysis) as it implies that, if log-price deviations of the next competitor are expected to account for more than 5% of equivalent log-price changes of the product’s own cross-elasticity coefficient $\varphi_i$, we conclude that this is 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 that they provide with respect to defining the 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, enables inferences on subsequent entries of these cross-elasticity vectors that flexibly incorporates the rates of decay across the previous entries.

7.4.1.2. Definition 2: aggregate competition. One of the primary interests of the analysis is characterizing products in terms of their sales sensitivities with respect to their competitors’ prices. We introduce the notion of aggregate competition $AC$ to summarize the total effect of competition on a product’s sales through its competitors’ price changes. We achieve this by defining the aggregate competition of product $i$ as the sum of the top $l$ cross-elasticity coefficients. Concretely, $AC$ for a cross-elasticity vector distribution is given by

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

$AC$ 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 price 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 that its competitors’ price changes would have had on the product’s sales. Thus a large $AC$ indicates that a product’s sales are significantly impacted by its competitors’ prices.

7.4.2. Markov chain Monte Carlo output

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

$$(\alpha, \beta, \lambda, w) \sim \text{gamma}(7, 7/10) \text{ gamma}(0.5, 1) \text{ gamma}(1, 1) \text{ beta}(2, 3),$$

$$\nu \sim \text{gamma}(5, 1)$$

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. Specifically, the priors over $(\alpha, \beta)$ are selected to reflect prior knowledge of the modal nature of the coefficients and the expected heavy-tailed nature of the cross-elasticity coefficients. In addition, these priors were chosen to be more restrictive than in the simulated examples, to overcome the strong dependence between $\alpha$ and $\beta$ which, for small data sets such as this, leads to weak identifiability. In particular, the prior for $\alpha$ is centred near 10 as before, but with a variance of about 14. The complementary parameter $\beta$ is then centred at 0.5 (corresponding to a heavy-tailed EW mixture), with a variance of 0.5. In other words, mixtures are ‘shrunk’ towards smaller values of $\beta$, i.e. towards assuming that there are no omitted competitors, unless the data strongly suggest otherwise. The prior $w \sim \text{beta}(2, 3)$ is selected to prefer observed cross-elasticities of length 4–5. The prior for $\lambda$ (purely a scale
Fig. 6. (a) Histogram of \( N^* \) and (b) heat map of cluster co-membership probabilities (regrouped with respect to \( C^* \)).

Table 4. Retail analytics clusterwise 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 of the clusters.

| Parameter | Results for cluster 1 | Results for cluster 2 | Results for cluster 3 |
|-----------|-----------------------|-----------------------|-----------------------|
| \( \alpha \) | 0.16 (0.036, 0.28) | 0.06 (0.030, 0.21) | 5.73 (2.49, 10.19) |
| \( \beta \) | 1.51 (1.10, 2.13) | 1.71 (1.20, 2.39) | 10.88 (5.88, 17.06) |
| \( \lambda \) | 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% |
| \( E(\tilde{\eta}_{10}) \) | 0.287 | 0.304 | 0.984 |
| \( E(\tilde{\eta}_{9}) \) | 0.156 | 0.124 | 0.963 |
| \( E(\tilde{\eta}_{8}) \) | 0.092 | 0.053 | 0.937 |
| \( E(\tilde{\eta}_{7}) \) | 0.055 | 0.022 | 0.926 |

parameter) is uninformatively chosen as before, and the prior for \( \nu \) is chosen such that, a priori, between three and 30 clusters (roughly) are expected.

Fig. 6 presents a histogram of the number of unique clusters \( N^* \) and a heat map 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 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 the optimal clusters.
Fig. 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 Fig. 5 we observe a spike of very small values for the top order statistics $\eta_{10}$ and $\eta_{9}$, which the model can accommodate through a small value of $\alpha\beta$.

Fig. 7 provides heat maps of pairwise posterior distributions of the parameters which demonstrate a neat separation between pairwise atoms. Interestingly, we observe that larger values of $\lambda$ (corresponding to a smaller mean) are associated with larger values of $w$; instead, $\beta$-values are inversely associated with values of $w$, suggesting that the censoring in this case is largely driven by $\beta$.

To assess model fit, we calculate the posterior predictive $p$-values (Meng, 1994) of AC for each of the clusters that are defined by $C^*$. Posterior predictive $p$-values involve generating repetitions $X^{rep}$ from the predictive distribution $p(X^{rep} | \alpha, \beta, \lambda, w)$ for each MCMC sample and calculating the $p$-value $2[1 - p\{T(X^{rep}) > T(X) | X\}]$ for some test statistic $T(X)$: in this case the aggregate competition. Fig. 8 provides predictive posterior $p$-value plots on the observed aggregate competition AC over each cluster, compared against histograms of generated AC-statistics over predictive replicates of $X$. These all comfortably fall within the 95% prediction intervals. Trace plots of the $(\alpha, \beta, \lambda, w)$ atoms across the unique clusters are included in Appendix B.

7.4.3. Retail analytics discussion
Considering the clusters that are given by $C^*$ 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 significantly underaverage 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 $OC_1 = 0.038$ and $OC_2 = 0.031 < \epsilon$ and thus indicate that we do not expect any of the unobserved cross-elasticities to be of any significance. However, the third cluster exhibits competitor omission since $OC_3 = 0.96 > \epsilon$. 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 $\alpha_1 = 0.16$ and $\beta_1 = 1.51$ with $w_1 = 0.34$ and an aggregate competition of $AC_1 = 0.55$, which point to a light-tailed distribution of cross-elasticities whose probability density diverges at 0. 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 $AC_2 = 0.42$, also implying a light-tailed distribution whose probability density diverges at 0.

The third cluster is rather different; its posterior mean parameters $\alpha_3 = 5.73$ and $\beta_3 = 10.88$ suggest a light-tailed distribution with mode away from 0. It largely consists of vectors with only a single cross-elasticity entry (through $w_3 = 0.016$), although the model suggests that an additional competitor may have been missed (or does not exist). Finally, its aggregate competition (despite the missing competitor) is $AC_3 = 1.11$, so price changes of these leading competitor products can account for 1.11 of equivalent prices changes of the product’s own price changes. These parameters suggest that these products are substitutes, i.e. products only bought as an alternative because other equivalent products are unavailable or too expensive.
Fig. 7. Heat maps of pairwise posterior distributions of the parameters.
Fig. 8. Histograms of AC-samples with the 2.5% and 97.5% quantiles (i) and AC (j) for each cluster in $C^*$ (we observe that our AC test statistic falls comfortably in the credible range for each cluster): (a) cluster 1; (b) cluster 2; (c) cluster 3

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 food items which traditionally have a higher number of competitors than in the second cluster. Similarly to 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 understand the optimal pricing and display combinations better. For example, products in the third cluster are highly sensitive to specific competitor products but otherwise are unaffected by the bulk of products around them. In contrast, 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.

8. Summarizing remarks

We have presented a Bayesian non-parametric mixture model for censored ordered data, using the EW distribution as a kernel. Our approach enables flexible modelling of cross-elasticity coefficients without the need to specify the number of components and lends itself to meaningful interpretation. We implemented our methods on a data set of cross-elasticities, focusing on quantities of interest in the retail analytics context, such as aggregate competition and potential omitted competitors. Our model could 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 enable 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.

Acknowledgements

This work has been carried out with the financial support of the Engineering and Physical Sciences Research Council, the Alan Turing Institute and dunnhumby ltd, our industrial partner. Access to the anonymized data servers was granted to us by dunnhumby ltd. The anonymized data set covers a wide range of sales categories and has sales going back over 5 years.

Appendix A: Posterior sampler

Here we present the details of the posterior inference procedure. We provide more exact expositions of each of the three components of the MCMC algorithm; sampling \( \theta = (\alpha, \beta, \lambda, w) \) atoms of the \( \nu G_0 \) for each of the order statistics sequences, samples from \( p(\alpha, \beta, \lambda, w|\nu, x_{\{c \in \mathcal{C}_i\}}) \) for each cluster and sampling the \( \nu \).

A.1. Sample from \( p(\theta_i|\theta_{-i}, \nu, x_j) \)

As discussed during the posterior inference, we use the algorithm by Neal (2000), and we sample \( \theta_i = (\alpha_i, \beta_i, \lambda_i, w_i) \) by sampling from the multinomial distribution of degrees of freedom of order \( N^* + c \) with entries

\[
\theta_k^{* \text{ HD}} \sim G_0, \quad k = N^* + 1, \ldots, N^* + c,
\]

\[
G_0 = \text{beta}(w|a, b) \text{ gamma}(\alpha|\alpha^1, \alpha^2) \text{ gamma}(\beta|\beta^1, \beta^2) \text{ gamma}(\lambda|\lambda^1, \lambda^2)
\]
with probabilities \( P(\theta_i = \theta_i^* | \theta_{-i}, x_i, \theta_i^*, \ldots, \theta_{N^*+c}^*) \), which is equivalent to

\[
P(C_i = k | C_{-i}, x_i, \theta_i^*, \ldots, \theta_{N^*+c}^*) \propto \begin{cases} \frac{N_i^*}{N - 1 + \nu} f(x_i | \theta_i^*) & \text{for } 1 \leq k \leq N^*, \\ \frac{\nu}{N - 1 + \nu} f(x_i | \theta_i^*) & \text{for } N^* < k \leq N^* + c \end{cases}
\]

where

\[
f(x_i | \theta_i^*) = \left( \frac{n_i - 1}{l_i - 1} \right) \left( w_i^* \right)^{l_i - 1} (1 - w_i^*)^{n_i - l_i - 1} F(x_{i,n_i-l_i-1} | \alpha_k^*, \beta_k^*, \lambda_k^*)^{n_i - l_i - 1} \prod_{j=1}^{l_i} f(x_{i,n_i-j-1} | \alpha_k^*, \beta_k^*, \lambda_k^*)
\]

where \( F(x | \alpha, \beta, \lambda) = 1 - \exp\{-\lambda x^\beta \} \) and \( f(x | \alpha, \beta, \lambda) = \alpha \beta \lambda x^{\beta - 1} [1 - \exp\{-\lambda x^\beta \}]^{\alpha - 1} \exp\{-\lambda x^\beta \} \).

### A.2. Sample from \( p(\alpha^*, \beta^*, \lambda^*, w^* | \nu, X_{i:C_i=k}) \)

To ease notation, we suppress the asterisks from the exponents in this subsection. For \( t = 1, \ldots, T \) iterations, for each unique cluster \( k = 1, \ldots, N^* \), we draw new parameters by using an exponentiated normal distribution proposal for \( \alpha_k^*, \beta_k^*, \lambda_k^* \) centred at the points \( (\log(\alpha_k^*), \log(\beta_k^*), \log(\lambda_k^*)) \) with standard deviations \( \sigma_\alpha, \sigma_\beta \) and \( \sigma_\lambda \), and introducing an auxiliary variable \( \omega_k \) with probabilities

\[
\pi(\alpha_k', \beta_k', \lambda_k', w_k' | x_{i:C_i=k}) \sim [\log(\omega_k') \log(\sigma_\alpha) \log(\sigma_\beta) \log(\sigma_\lambda)] N(\omega_k', \sigma_k^2) \]

This form of the proposals for \( \alpha_k', \beta_k', \lambda_k' \) allows the scale of the proposal to vary according to the magnitude of the parameter values and ensures that proposed values are always positive.

Then, if \( 0 < w_k' < 1 \) (since \( w \) is bounded between 0 and 1), with probability

\[
a = \min \left\{ 1, \frac{\pi(\alpha_k', \beta_k', \lambda_k', w_k' | x_{i:C_i=k})}{\pi(\alpha_k, \beta_k, \lambda_k, w_k | x_{i:C_i=k})} \right\},
\]

set \( (\alpha_k^{t+1}, \beta_k^{t+1}, \lambda_k^{t+1}, w_k^{t+1}) = (\alpha_k', \beta_k', \lambda_k', w_k') \), and otherwise \( (\alpha_k^{t+1}, \beta_k^{t+1}, \lambda_k^{t+1}, w_k^{t+1}) = (\alpha_k, \beta_k, \lambda_k, w_k) \). Here

\[
\pi(\alpha_k, \beta_k, \lambda_k, w_k | x_{i:C_i=k}) \propto \alpha_k^{l_i - 1} \exp(-\alpha_k \lambda_k) \beta_k^{l_i - 1} \exp(-\beta_k \lambda_k) \lambda_k^{-1} \exp(-\lambda_k) w_k^{n_i - l_i - 1} \prod_{j=1}^{l_i} f(x_{i,n_i-j-1} | \alpha_k, \beta_k, \lambda_k)
\]

The scales of the proposal normal distributions \( \sigma_\omega, \sigma_\alpha, \sigma_\beta \) and \( \sigma_\lambda \) should be tuned depending on the data set.

### A.3. Sample from \( p(\nu | \alpha, \beta, \lambda, w, X) \)

Finally we implement the prior of Escobar and West (1995) on \( \nu \) of the DP. By using the fact that specifying \( \nu \sim \text{gamma}(\tau_1, \tau_2) \) and introducing an auxiliary variable \( \gamma \) enables the sampling of \( \nu \) iteratively by a Gibbs sampler. Specifically, we take the samples

\[
(\gamma | \nu, N^*) \sim \text{beta}(\nu + 1, N),
\]

\[
(\nu | \gamma, N^*) \sim \pi_\gamma \text{gamma} \{\tau_1 + N^*, \tau_2 - \log(\gamma)\} + (1 - \pi_\gamma) \text{gamma} \{\tau_1 + N^* - 1, \tau_2 - \log(\gamma)\}
\]

where the weight \( \pi_\gamma \) is defined by \( \pi_\gamma / (1 - \pi_\gamma) = (\theta + N^* - 1) / [N \{\tau_2 - \log(\gamma)\}] \). This concludes the complete iteration of our posterior inference procedure.

### Appendix B: Markov chain Monte Carlo trace plots

Here we assess the convergence of our inferential procedure on the retail analytics data set. Fig. 9 provides traces of the atoms across all unique clusters \( ((\alpha_k^*), (\beta_k^*), (\lambda_k^*), (w_k^*)) \) of DP(\( \nu G_0 \)) samples for the iterations \( t = 1, \ldots, T \) across the unique atoms \( k = 1, \ldots, N_k^* \), where \( N_k^* \) is the number of unique clusters.
Fig. 9. Trace plots of MCMC samples for unique atoms of (a) $\sqrt{\lambda}$, (b) $\sqrt{\beta}$, (c) $\sqrt{\gamma}$, (d) $w$ and (e) $N^*$ on Dunhumby's cross-elasticity data of the snack category.
at iteration $t$ and the trace of $N^t$. We plot the square-root traces of $(\alpha, \beta, \lambda)$ to induce similar scales for graphical convenience. All plots indicate sufficient mixing and satisfactory convergence.

References

Akay, A. E. (2013) Statistical estimation problems in inventory management. PhD Thesis. Tepper School of Business, Carnegie Mellon University, Pittsburgh.

Andreyeva, T., Long, M. W. and Brownell, K. D. (2010) The impact of food prices on consumption: a systematic review of research on the price elasticity of demand for food. Am. J. Publ. Hlth, 100, 216–222.

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist., 2, 1152–1174.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992) A First Course in Order Statistics. Philadelphia: Society for Industrial and Applied Mathematics.

Bajari, P., Nekipelov, D., Ryan, S. P. and Yang, M. (2015) Machine learning methods for demand estimation. Am. Econ. Rev., 105, 481–485.

Barghout, M., Littlewood, B. and Abdel-Ghaly, A. (1998) A non-parametric order statistics software reliability model. Softw. Test. Verific. Reliab., 8, 113–132.

Beheshti-Kashi, S., Karimi, H. R., Thoben, K.-D., Lütjen, M. and Teucke, M. (2015) A survey on retail sales forecasting and prediction in fashion markets. Syst. Sci. Control Engng, 3, 154–161.

Beutner, E. and Kamps, U. (2009) Order restricted statistical inference for scale parameters based on sequential order statistics. J. Statist. Plann. Plunger Inf., 139, 2963–2969.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Pólya urn schemes. Ann. Statist., 1, 353–355.

Barron, F. and Teh, Y. W. (2012) Bayesian nonparametric models for ranked data. In Advances in Neural Information Processing Systems, pp. 1520–1528.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An ANOVA model for dependent random measures. J. Am. Statist. Ass., 99, 205–215.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. J. Am. Statist. Ass., 90, 577–588.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. Ann. Statist., 1, 209–230.

Ferreira, K. J., Lee, B. H. A. and Simchi-Levi, D. (2015) Analytics for an online retailer: demand forecasting and price optimization. Manufct. Serv. Oper. Mangmnt, 18, 69–88.

Ghosh, K. and Tiwari, R. C. (2007) Nonparametric and semiparametric Bayesian reliability analysis. In Encyclopedia of Statistics in Quality and Reliability. New York: Wiley.

Guerrero-López, C. M., Unar-Munguía, M. and Colchero, M. A. (2017) Price elasticity of the demand for soft drinks, other sugar-sweetened beverages and energy dense food in Chile. BMC Publ. Hlth, 17, 180.

Gunawardana, A. and Shani, G. (2009) A survey on retail sales forecasting for fashion retailing service industry. J. Mach. Learn. Res., 10, 2935–2962.

Hanson, T. E. (2006) Modeling censored lifetime data using a mixture of Gammas baseline. Bayesn Anal., 1, 575–594.

Hoch, S. J., Kim, B.-D., Montgomery, A. L. and Rossi, P. E. (1995) Determinants of store-level price elasticity. J. Marking Res., 32, 17–29.

Huh, W. T. and Rusmevichientong, P. (2009) A nonparametric asymptotic analysis of inventory planning with censored demand. Math. Oper. Res., 34, 103–123.

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

Kottas, A. (2006) Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. J. Statist. Plungh Inf., 136, 578–596.

Lau, J. W. and Green, P. J. (2007) Bayesian model based clustering procedures. J. Computnl Graph. Statist., 16, 526–558.

Leser, C. E. V. (1963) Forms of Engel functions. Econometrica, 31, 694–703.

Liu, N., Ren, S., Choi, T.-M., Hui, C.-L. and Ng, S. F. (2013) Sales forecasting for fashion retailing service industry: a review. Math. Probl. Engng., article 738675.

Malacar, R. M. and Hopkins, W. G. (2014) Variability of competitive performance of elite athletes: a systematic review. Sprts Med., 44, 1763–1774.

Meng, X.-L. (1994) Posterior predictive p-values. Ann. Statist., 22, 1142–1160.

Mudholkar, G. S. and Srivastava D. K. (1993) Exponentiated Weibull family for analyzing bathtub failure-rate data. IEEE Trans. Reliab., 42, 299–302.

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R. and Song, D. (2012) On the feasibility of internet-scale author identification. In Proc. Symp. Security and Privacy, pp. 300–314. New York: Institute of Electrical and Electronics Engineers.
Nassar, M. M. and Eissa, F. H. (2003) On the exponentiated Weibull distribution. *Communs Statist. Theory Meth.*, 32, 1317–1336.
Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *J. Computnl Graph. Statist.*, 9, 249–265.
Oliveira, J. M., Foxall, G. R. and Schrezenmaier, T. C. (2007) Consumer brand choice: individual and group analyses of demand elasticity. In *The Behavioral Economics of Brand Choice*, pp. 223–255. Berlin: Springer.
Persson, P.-G. (1995) Modeling the impact of sales promotion on store profits. Foundation for Distribution Research, Economic Research Institute, Stockholm School of Economics, Stockholm.
Rudin, C., Letham, B. and Madigan, D. (2013) Learning theory analysis for association rules and sequential event prediction. *J. Mach. Learn. Res.*, 14, 3441–3492.
Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, 4, 639–650.
Silver, D., Newnham, L., Barker, D., Weller, S. and McFall, J. (2013) Concurrent reinforcement learning from customer interactions. In *Proc. 30th Int. Conf. Machine Learning*, pp. 924–932.
Warr, R. L. and Collins, D. H. (2014) Bayesian nonparametric models for combining heterogeneous reliability data. *Proc. Inst. Mech. Engrs O*, 228, 166–175.
Wilson, S. P. and Samaniego, F. J. (2007) Nonparametric analysis of the order-statistic model in software reliability. *IEEE Trans. Softwr. Engng*, 33, 198–208.
Working, H. (1943) Statistical laws of family expenditure. *J. Am. Statist. Ass.*, 38, 43–56.