Machine Learning (2023) 112:281–310
https://doi.org/10.1007/s10994-022-06276-8

# Dynamic customer segmentation via hierarchical fragmentation-coagulation processes

**Ling Luo**[1] · **Bin Li**[2] · **Xuhui Fan**[3] · **Yang Wang**[4] · **Irena Koprinska**[5] · **Fang Chen**[4]

Received: 10 November 2020 / Revised: 22 September 2022 / Accepted: 3 November 2022 /
Published online: 2 December 2022


**Abstract**
Understanding customer behavior is necessary to develop efficient marketing strategies or launch tailored programs with social value for the public. Customer segmentation is a critical task for understanding diverse and dynamic customer behavior. However, as the popularity of different products varies, building dynamic customer behavior models for products with few customers may overfit the data. In this paper, we propose a new Bayesian nonparametric model for dynamic customer segmentation—Hierarchical Fragmentation-Coagulation Processes (HFCP), which allows sharing behavior patterns across multiple products. We conduct comprehensive empirical evaluations using two real-world purchase datasets. Our results show that HFCP can: (i) determine the number of groups required to model diverse customer behavior automatically; (ii) capture the changes such as split and merge of customer groups over time; (iii) discover behavior patterns shared among products and identify products with similar or different purchase behavior impacted by promotion, brand choice and change of seasons; and (iv) overcome overfitting problems and outperform previous customer segmentation models on estimating behavior for unseen customers. Hence, HFCP is a flexible and accurate segmentation model that can be used by stakeholders to understand dynamic customer behavior and compare the purchase behavior for different products.

**Keywords** Customer segmentation · Purchase behavior analytics · Temporal data analysis · Bayesian temporal models

## 1 Introduction

Modeling customer purchase behavior is a critical task for successful business operation and marketing. Understanding customer purchase behavior allows businesses to identify the target customers that are most likely to buy their products, so that they can reach the right customers at the right time (Wang and Zhang, 2013). An accurate purchase behavior

Editor: Ulf Brefeld.

✉ Ling Luo
ling.luo@unimelb.edu.au

Extended author information available on the last page of the article

model is necessary to develop cost- and time-efficient marketing strategies, or launch tailored programs with social value for the public. In this paper we focus on dynamic customer segmentation—identifying groups of customers with similar purchase behavior and tracking their evolution over time. Dynamic customer segmentation is a critical part in customer behavior models.

The main challenges of building an accurate customer behavior model are posed by the dynamics and diversity of purchase behavior data. The customer purchase behavior can change dynamically due to various factors, such as the popularity of products, promotional campaigns and changes of seasons. For example, the availability and price of fruits such as strawberries are significantly impacted by seasonal changes, so that their sale volume fluctuates through a year. This requires a flexible temporal model to track the dynamic purchase behavior over time. Moreover, as different customers have diverse shopping habits, different preferences for products and receptivenesses to price changes, analyzing all customers together may overlook important patterns. Some customers may buy a product only when it is on sale, while another group of customer may purchase regularly without waiting for promotions. It is beneficial to segment customers into different groups and capture the behavior pattern of each customer group.

The two main customer segmentation approaches include rule-based methods and mixture models. The rule-based methods assume that the customers with different geographic, demographic or cultural characteristics have different preferences, so the customers are segmented based on these attributes (Song et al., 2001; Dong and Kaiser, 2008; Böttcher et al., 2009). The mixture modeling approach is data-driven, which identifies different groups based on their historical purchase records. The behavior of an individual customer is modeled as a mixture of different prototypes of behavior weighted by the membership in those groups (Bucklin et al., 1998). To model the dynamic behavior, the temporal segmentation models were designed, and the techniques range from a mixture of stochastic processes (Kim et al., 2017), dynamic topic modeling (Iwata et al., 2009), collaborative filtering over time (Li et al., 2011) to Hidden Markov Model (HMM) (Netzer et al., 2008). Recently, a random partition model Fragmentation-Coagulation Processes (FCP) was adapted to conduct dynamic segmentation of customer behavior (Luo et al., 2017). FCP uses a sequence of partitions to track the evolution of customer groups—it can split one group into smaller groups when its members start to have divergent behavior, and merge several groups when their members have similar behavior.

Existing techniques such as those proposed in (Kim et al., 2017; Iwata et al., 2009; Luo et al., 2017) consider a single product for each model they build. There are two weaknesses of single-product analysis in this way: (1) business analysts often desire to explore and compare the purchase behavior of multiple products, which can support them to design effective promotion strategies across products and optimize product placement (Datta et al., 2010; Du et al., 2017); and (2) from the data availability perspective, although the size of the whole purchase dataset can be large, the amount of records for each product often forms a long-tail distribution, since many products only have small amount of records (Clemons and Nunes, 2011). When the data of a single product is used to train an advanced temporal model, the model may overfit the small amount of records for less popular products, making it difficult to find useful patterns or generalize to unseen customers. If the records of multiple products are simply mixed together, the model may lose the behavior patterns associated with some minor products and it is also hard to match the identified patterns with products after getting the model. In this case, a hierarchical model, which learns shared behavior patterns across multiple products, is a better choice.

Through the hierarchical structure, a model can describe a customer group using a pattern selected from a shared set of patterns, which is learned given the records of multiple products, but not simply mixing them together. For example, the hierarchical model can learn that overall there are three patterns A, B and C considering their purchase rates and cycles, shared by the customers purchasing apples and bananas. The customers buying apples present patterns A and B, while the customers buying bananas present patterns B and C. In this way, analyzing the shared patterns across multiple products can support stakeholders to easily compare the purchase behavior of these products, e.g. detecting whether the sales of different products are negatively correlated due to the competition among them. In addition, learning shared patterns is more robust, which overcomes overfitting problems for some products with few transaction records.

Therefore, we are motivated to propose a novel model named *Hierarchical Fragmentation Coagulation Processes (HFCP)* for the dynamic customer behavior segmentation of multiple products. Given the purchase records of a set of customers, HFCP can not only track the split and merge of customer groups, but more importantly, it can learn the shared behavior patterns across products and avoid overfitting the records of a single product. In addition, HFCP can provide stakeholders with a comprehensive understanding of the dynamic behavior patterns of multiple relevant products, including identifying the similarities and differences of the products.

It is worth noting that HFCP is not a straightforward extension of FCP, and the design and inference of HFCP involves significant technical challenges. We have tackled the critical problems to ensure HFCP could have consistent marginal distribution over time and align customer groups across products. More details about the design of HFCP will be introduced in Sect. 4.

We conduct empirical evaluations of HFCP using two real-world supermarket purchase transaction datasets. The experimental results demonstrate the main innovations and strengths of the proposed HFCP model:

1. It is a Bayesian nonparametric model, which can determine and adjust the number of groups in the segmentation to model customer behavior for multiple products.
2. It can capture the dynamics of purchase behavior by splitting and merging customer groups over time.
3. It can discover the behavior patterns shared across multiple products via the hierarchical approach, which can identify products with similar or different purchase behavior. We examine the purchase patterns of more than 100 products from two supermarket datasets using HFCP and compare each product with their counterparts from the same product category. In the case studies, we explore the impact of promotions, brand choice and change of seasons on the purchase behavior.
4. It outperforms three other customer segmentation models—the mixture of Homogeneous Poisson Processes (HomoPP), the mixture of Non-Homogeneous Poisson Processes (NHPP) and FCP, on estimating the purchase behavior of unseen customers. It demonstrates that sharing patterns across relevant products can overcome the overfitting problems and improve the model's generalization capability to handle unseen data.

## 2 Related work

Customer segmentation can be used to identify diverse behavior patterns in the market (Sarkar et al., 2018; Carnein and Trautmann, 2019). Conventional segmentation techniques include rule-based models and mixture models. The rule-based segmentation models (Dong and Kaiser, 2008) studied purchase behavior based on various criteria, such as income, race, age and education level. For example, Taylor et al. (2015) developed a scoring system named Healthy Trolley Index to examine dietary quality. They grouped different customers by their gender, age and living arrangements and compared the proportions of food expenditure on different product categories with the benchmark provided in the official guide to healthy eating. However, the previous analysis found that it may not be helpful to segment customers based on demographic and psychographic variables for frequently purchased products such as food and drinks (Bucklin and Gupta, 1992), whereas using behavioural variables is a sensible approach to segmenting customers (Kotler and Armstrong, 2010). The segmentation based on mixture modeling is data-driven, which infers the latent customer groups using purchase data. Bucklin et al. (1992; 1998) proposed mixtures of logit models to segment customers based on their brand choice, purchase events and quantity bought.

Apart from the diverse behavior patterns, the dynamics of customer behavior is another challenge for modeling. Stochastic processes can be used to capture the dynamics of data (Ross, 1996; Kim et al., 2014; Costa et al., 2015; Ren et al., 2008; Elliott and Teh, 2016). Kim et al. (2014) proposed a hierarchical time-rescaling point processes for modeling temporal patterns such as periodic, bursty, self-exciting and sale-effect patterns. However, this work only modeled the behavior of individual customers, which may overfit sparse records. In order to deal with both diversity and dynamics, temporal components can be integrated to the mixture models. For example, the mixture of NHPP (Luo et al., 2016) was proposed to group customers based on their behavior patterns, and each group was described by a Poisson process with an intensity function over time, which combines a polynomial term and a periodic term. Then a customer can be modeled by their soft membership in customer groups and the behavior patterns of those groups. Iwata et al. (2009) proposed a topic tracking model to identify a set of latent topics for customer preferences and detect the changes of these topics. However, the number of groups needs to be predefined in (Luo et al., 2016) and (Iwata et al., 2009), and the customer group membership remains unchanged over time. Bayesian nonparametric dynamic models such as (Ren et al., 2008; Elliott and Teh, 2016) can generate flexible number of groups in clustering. Ren et al. (2008) proposed dynamic Hierarchical Dirichlet Process to model time-evolving data. The clustering at each time step is modeled by Dirichlet Process (DP), which can have unbounded number of groups. The DPs at consecutive time points are linked via a parameter to control their similarity. These three models have limitations in tracking the changes in both customer group membership and group-level behavior. For instance, when a promotion starts, the customers of a group can have different responses to the price change, which leads to a split of one group into multiple groups.

To capture the dynamics of customer group membership, the models in (Netzer et al., 2008) and (Xing and Sohn, 2007) were proposed to describe behavior changes via the transition of latent states of HMM. More specifically, Netzer et al. (2008) proposed a non-homogeneous HMM with time-varying covariates in the transition matrix to track latent customer-business relationship states, such as dormant, transitory or active. The number of states was determined by model selection measures, like log-marginal density and deviance

information criterion. Xing and Sohn (2007) and Elliott and Teh (2016) segmented data using HMM, where the different states of an HMM were used to describe the groups of a partition. Despite its flexibility, this may generate large amount of states, which makes it hard to learn the transition probabilities between possible states from sparse observations. The random partition model FCP (Bertoin, 2006) did not match latent states with groups, but it utilized a sequence of partitions to capture the evolution of groups. FCP conducted fragmentation and coagulation operations directly on group members, so that it can easily track how members switch from one group to another. The idea of FCP has been successfully adapted in different applications such as genetic analysis (Teh et al., 2011; Elliott and Teh, 2012), financial markets (Eguiluz and Zimmermann, 2000) and customer segmentation (Luo et al., 2017). However, the limitation of FCP in (Luo et al., 2017) is that it modeled the purchase behavior of a single product, which makes it hard to analyze behavior patterns across multiple products and may overfit the purchase records of a single product.

In contrast to above, our HFCP is an innovative Bayesian nonparametric model to conduct dynamic customer segmentation, identify purchase behavior patterns shared among products, which can overcome the overfitting problem for modeling individual customers or a single product, and also facilitate business analysts to compare the purchase patterns of different products.

## 3 Preliminaries

The proposed model HFCP is a hierarchical FCP, and the snapshot (i.e. marginal distribution) of HFCP at any time step can be interpreted as an HDP, so the basics of FCP and HDP are introduced in this section as preliminaries.

### 3.1 Fragmentation-coagulation process

FCP is a random partition process defined based on Chinese Restaurant Processes (CRP) (Teh et al., 2011). The partition at any time $t$ in FCP follows CRP. CRP is a random partition model, which can be described by the analogy of customers choosing tables in a Chinese restaurant (Pitman, 2002a).

Each table of CRP corresponds to a customer group in a partition. For a partition $\rho \sim \text{CRP}(A, \alpha, \delta)$[1], $A$ is the set of all customers, $\alpha$ is the *strength* parameter ($\alpha > -\delta$), which controls the number of groups in partition $\rho$; and $\delta \in [0, 1)$ is the *discount* parameter, which controls the probability of generating more tables with few customers (Pitman and Yor, 1997). For a new customer, the probabilities of joining an existing group (or called block) $b$ of the partition $\rho$ or starting a new group are defined in Eqs. 1 and 2 respectively,

$$P(\text{ join group } b) = \frac{n_b - \delta}{\alpha + \sum_{b \in \rho} n_b} \tag{1}$$

[1] In our model, the two-parameter version of CRP is used, which has Pitman-Yor Process as the de Finetti measure (Pitman and Yor, 1997).

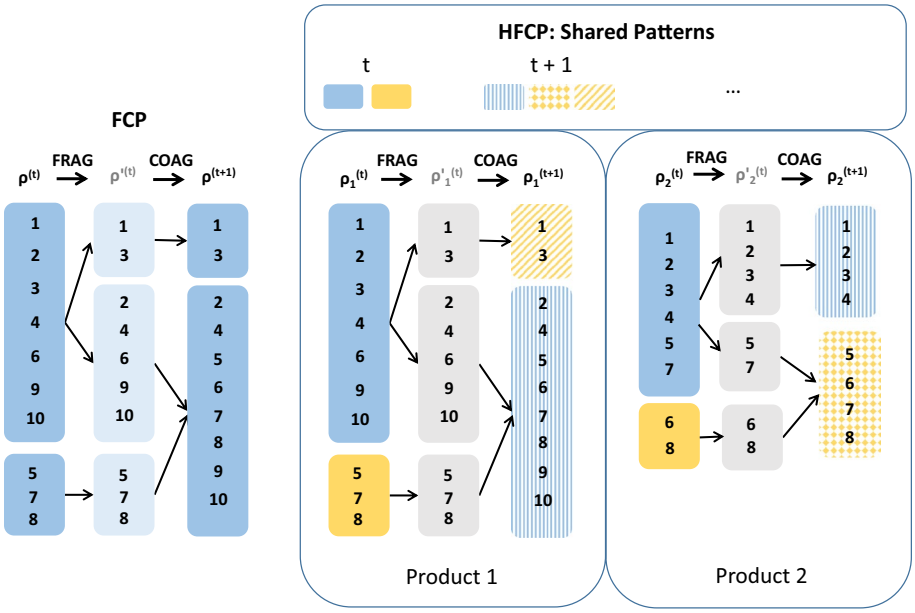


**Fig. 1** Illustration of FCP (left) and HFCP for two products (right) at two time steps (Color figure online)

$$P(\text{start a new group}) = \frac{\alpha + \delta|\rho|}{\alpha + \sum_{b\in\rho} n_b} \tag{2}$$

where $b$ is a group in partition $\rho$, $n_b$ is the size of $b$, and $|\rho|$ is the number of groups in $\rho$.

CRP is a nonparametric model, which does not require specifying the number of groups in a partition, and the number of groups can increase with the amount of data.

FCP contains a sequence of random partitions following CRP, with temporal dependencies among the partitions defined by the fragmentation and coagulation operations. The process FCP$(A, \alpha, \epsilon)$ contains the partitions of dataset $A$, the *strength* parameter $\alpha$ controls the number of groups, and $\epsilon \in [0, 1)$ controls the temporal dependency between consecutive partitions. When $\epsilon = 0$, the partition at $t + 1$ is fully dependent on the partition at $t$, which means that the partition remains the same over time. When $\epsilon \to 1$, the partitions at $t$ and $t + 1$ are independent.

In discrete-time FCP$(A, \alpha, \epsilon)$, the partition at $t = 1$ is $\rho^{(1)} \sim \text{CRP}(A, \alpha, 0)$. The partition $\rho^{(t+1)}$ $(t \in \{1, \ldots, T\})$ is generated by splitting (fragmentation operations, FRAG) and merging (coagulation operations, COAG) its previous partition as follows:

$$\textit{fragmentation}: \quad \rho'^{(t)}|\rho^{(t)} \sim \text{FRAG}(\rho^{(t)}, 0, \epsilon) \tag{3}$$

$$\textit{coagulation}: \quad \rho^{(t+1)}|\rho'^{(t)} \sim \text{COAG}(\rho'^{(t)}, \alpha/\epsilon, 0) \tag{4}$$

The illustration of discrete-time FCP in Fig. 1 (left). The numbers in the boxes are customer indices. In the *FRAG* operation, each block $b$ in $\rho^{(t)}$ is further split based on CRP$(b, 0, \epsilon)$ to get $\rho'^{(t)}$. Each block in $\rho'^{(t)}$ either exists in $\rho^{(t)}$ or is a subset of a block $b$ in $\rho^{(t)}$, so that $\rho'^{(t)}$ is a finer partition of $A$ than $\rho^{(t)}$.

In the *COAG* operation, all the blocks of $\rho'^{(t)}$ are treated as elements, and they are partitioned based on CRP($\rho'^{(t)}, \alpha/\epsilon, 0$) to get $\rho^{(t+1)}$. This operation is performed at the block level, so the customers in $b$ at $t$ will stay together, joining other blocks or remaining as a separated block. After the *COAG* operation, each block in $\rho^{(t+1)}$ either exists in $\rho'^{(t)}$ or it contains multiple blocks from $\rho'^{(t)}$, so that $\rho^{(t+1)}$ is a coarser partition of $A$. The FRAG and COAG operations are conducted alternately, and generates a sequence of partitions.

It has been proved that FCP is consistent and the marginal distribution of $\rho^{(t)}$ remains CRP($A, \alpha, 0$) over time (Elliott and Teh, 2012). The marginal distribution of intermediate partition $\rho'^{(t)}$ between two operations is CRP($A, \alpha, \epsilon$).

### 3.2 Hierarchical Dirichlet Processes

The Dirichlet Processes (DP) have been widely applied to clustering data with an infinite number of groups (Sethuraman, 1994). As discussed in (Teh et al., 2006), CRP is one of the construction approaches of DP. The Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006) use a set of DP to cluster multiple sets of data such as a collection of documents. Each DP of HDP learns the topic of a document based on its content, and these topics are shared across the collection of documents.

Given a collection of random measures $\{G_1, \dots G_J\}$ for $J$ datasets, an HDP model is defined as follows:

$$G_0 \sim \mathrm{DP}(\gamma, H) \tag{5}$$

$$G_j | G_0 \sim \mathrm{DP}(\alpha, G_0) \tag{6}$$

$$\theta_{ji} | G_j \sim G_j, \quad x_{ji} | \theta_{ji} \sim F(\theta_{ji}) \tag{7}$$

where $\gamma$ and $\alpha$ are concentration parameters. $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ is a discrete measure with infinite atoms $(\phi_k)_{k=1}^{\infty}$. Each $\phi_k$ is drawn from the base distribution $H$ and $(\delta_{\phi_k})_{k=1}^{\infty}$ is a probability measure concentrated at $(\phi_k)_{k=1}^{\infty}$. $(\beta_k)_{k=1}^{\infty}$ can be interpreted as the weights of mixture components. The random measures $\{G_j\}_{j=1}^{J}$ are conditionally independent given the base measure $G_0$. Most importantly, the atoms $(\phi_k)_{k=1}^{\infty}$ of $G_0$ are shared among $\{G_j\}_{j=1}^{J}$.

To get a partition of the dataset $A_j = \{(ji) | i = 1, 2, \dots\}$, each element $(ji)$ is associated with a latent factor $\theta_{ji}$. Each latent factor $\theta_{ji}$ draws an atom $\phi_k$ of $G_j$. Since $G_j$ is a discrete distribution, all $\theta_{ji}$ with the same value $\phi_k$ are members of group $k$. Drawing a sequence of factors $(\theta_{ji})_{i=1}^{\infty}$ based on Eqs. (5)–(7) naturally forms a partition of $A_j$. The observed data $x_{ji}$ is drawn from a distribution $F$, given the latent factor $\theta_{ji}$.

In the context of *customer behavior analysis*, $J$ datasets are the transaction records of $J$ products; $x_{ji}$ is the purchase behavior observation for customer $i$ of product $j$; $(\phi_k)_{k=1}^{\infty}$ are behavior patterns of different customer groups. The behavior patterns are prototypes of clusters, and these patterns are shared across $J$ products. $\theta_{ji}$ is the latent factor for customer $(ji)$ ; and $F$ is the likelihood function of the behavior observation, e.g. the probability density function of distributions. For example, $F$ can be Bernoulli distribution if the observations are binary; $F$ can be Poisson distribution, if the observations are the counts of event occurrence.

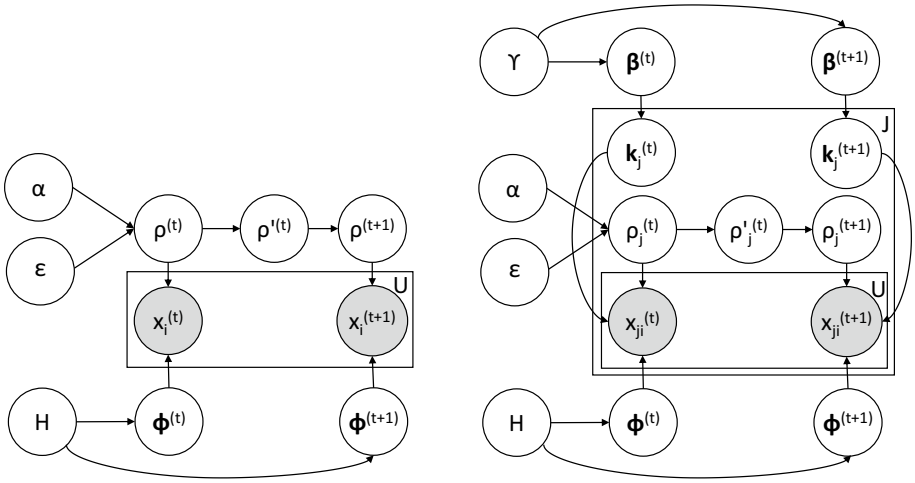


**Fig. 2** The graphical models of FCP (Luo et al., 2017) (left) and HFCP (right) at time steps $t$ and $t+1$

## 4 Customer segmentation with HFCP

We focus on dynamic customer purchase behavior segmentation task, which can be *formally defined* as: given the transaction records of $J$ products, segment customer set $A_j$ for each product $j$ based on their purchase behavior $\boldsymbol{x}_j^{(1:T)}$, identify the behavior pattern $\boldsymbol{\phi}_k^{(t)}$ of each customer group and track the evolution of customer groups over time.

Although FCP can track the split and merge of customer groups, constructing an independent FCP for each product could overfit the purchase records of less popular products, due to the long-tail distribution of purchase records (Clemons and Nunes, 2011). We are inspired by the idea of hierarchical approach, by which the behavior patterns can be jointly learnt and shared among multiple products, to avoid overfitting the records of individual product. However, adding a layer across multiple FCPs is not straightforward, as the model has to allow all FCPs to evolve flexibly and ensures that any behavior pattern in a snapshot of an FCP is selected from a set of patterns shared by multiple FCPs. Our proposed model Hierarchical Fragmentation-Coagulation Processes (HFCP) has HDP as the marginal distribution at any time, which is designed to tackle these challenges, overcome the overfitting problem and enable users to analyze shared dynamic behavior patterns to compare different products.

An illustration of HFCP, which segments customers of two products, is in Fig. 1 (right). At time $t$, customers in $A_j$ are segmented into blocks by $\rho_j^{(t)}$ based on FCP. Each block of $\rho_j^{(t)}$ has a behavior pattern $\boldsymbol{\phi}_k^{(t)}$, such as the number of purchases in one month. A pattern $\boldsymbol{\phi}_k^{(t)}$ is drawn from a set $\boldsymbol{\phi}^{(t)} = (\boldsymbol{\phi}_k^{(t)})_{k=1}^{\infty}$, which is shared across all products. Different blocks can select the same pattern, any block can add a new pattern to the pattern set, and the size of the pattern set is not limited. For example, in Fig. 1 (right), at $t$ and $t+1$, Product 1 (left box) and Product 2 (right box) share two and three patterns (as shown in the top box), respectively.

The formal graphical model of HFCP at time steps $t$ and $t+1$ is shown in Fig. 2 (right), and the graphical model of discrete-time FCP as defined in (Luo et al., 2017) is on the left for comparison. The corresponding generative process is presented in Algorithm 1. In more details, $\boldsymbol{\beta}^{(t)}$ contains the weights of all behavior patterns, which is drawn from

GEM($\gamma$) (GEM stands for Griffiths, Engen and McCloskey) (Pitman, 2002b), and the hyperparameter $\gamma$ determines the probability of generating a new pattern. Then, it starts the main loop over $J$ products. For each product $j$, the model generates a sequence of partitions $\rho_j^{(1:T)}$ based on FRAG and COAG operations. For each block $b_{jm}^{(t)} \in \rho_j^{(t)}$, we draw a pattern index $k_{jm}^{(t)}$ according to the weight $\boldsymbol{\beta}^{(t)}$. Subscript $m$ is the block index, $m \in \{1, \ldots, |\rho_j^{(t)}|\}$, where $|\rho_j^{(t)}|$ is the total number of blocks in this partition. For the pattern index, all the patterns $\boldsymbol{\phi}^{(t)} = (\phi_k^{(t)})_{k=1}^{\infty}$ are drawn from the base distribution $H$, and they are shared across $J$ products. For tractable inference, the base distribution $H$ is often defined as conjugate to the observation distribution $F$.

For each observation $x_{ji}^{(t)}$, it is drawn based on the block allocation $c_{ji}^{(t)}$ of the customer $(ji)$ , and the pattern $\phi_k^{(t)}$ of that block. The block allocation $c_{ji}^{(t)}$ refers to a block in $\rho_j^{(t)}$ in which $(ji)$ belongs to, so we have $c_{ji}^{(t)} = b_{jm}^{(t)}$ for $(ji) \in b_{jm}^{(t)}$.

**Algorithm 1** The generative process of HFCP

**Input:** customer set $A$ for $J$ products, hyperparameters $\gamma, \alpha, \epsilon$,

draw pattern weights $\boldsymbol{\beta}^{(t)} = (\beta_k^{(t)})_{k=1}^{\infty} \sim \text{GEM}(\gamma)$
draw patterns $\phi_k^{(t)} \sim H, k = 1, 2, \ldots$
**for** each product $j$ **do**
  **for** $t = 1$ **to** $T$ **do**
    **if** $t = 1$ **then**
      draw initial partition $\rho_j^{(1)} \sim \text{CRP}(A_j, \alpha, 0)$
    **end if**
    draw partitions
    ${\rho'}_j^{(t)} | \rho_j^{(t)} \sim \text{FRAG}(\rho_j^{(t)}, 0, \epsilon)$
    $\rho_j^{(t+1)} | {\rho'}_j^{(t)} \sim \text{COAG}({\rho'}_j^{(t)}, \alpha/\epsilon, 0)$
    **for** each block $b_{jm}^{(t)} \in \rho_j^{(t)}$ **do**
      draw a pattern index $k_{jm}^{(t)} \sim \text{Discrete}(\boldsymbol{\beta}^{(t)})$
    **end for**
    **for** each customer $(ji) \in b_{jm}^{(t)}$ **do**
      assign block allocation $c_{ji}^{(t)} = b_{jm}^{(t)}$, draw $x_{ji}^{(t)} \sim F(\phi_{k_{jm}}^{(t)})$
    **end for**
  **end for**
**end for**

For the observation data $x_{ji}^{(t)}$, the forms of $F(\phi_k^{(t)})$ and base distribution H are determined by the requirement of certain applications. For example, if we desire to distinguish customers by monthly purchase times, $x_{ji}^{(t)}$ can be drawn from Poisson($\lambda$), where $\lambda$ represents the expected number of purchases per month. When customer $(ji)$ has pattern $k^{(t)}$, $x_{ji}^{(t)}$ is drawn from Poisson($\lambda_k^{(t)}$). In this case, the base distribution $H$ can be Gamma($\alpha_\gamma, \beta_\gamma$), which is a conjugate prior for Poisson distribution. The hyperparameters of the Gamma distribution are *shape* $\alpha_\gamma$ and *scale* $\beta_\gamma$. We could replace Gamma-Poisson by other distributions. For example, if the behavior data is binary purchase indicator, it could be sampled from a Bernoulli distribution with a Beta prior.

The marginal distribution of HFCP at $t$ is designed as an HDP. More specifically, the behavior patterns shared across products correspond to atoms $(\phi_k)_{k=1}^{\infty}$ of $G_0$, with the weight $\boldsymbol{\beta}^{(t)} = (\beta_k^{(t)})_{k=1}^{\infty} \sim \text{GEM}(\gamma)$, and they are the prototypes of clusters (i.e. customer groups).

For the customers of product $j$, the partition $\rho_j^{(t)}$ has the marginal distribution CRP($A_j, \alpha, 0$). As CRP is one of the construction approaches of DP (Teh et al., 2006), the generative scheme of CRP is equivalent to the generative procedure of DP on $G_j$ level in Eqs. (6)–(7). Therefore, the partition of $A_j$ and the allocation of shared behavior patterns to all the blocks in $\rho_j^{(t)}$ correspond to constructing a two-level HDP.

The main advantages of HFCP are:

1) It can automatically determine the number of customer behavior groups, and increase or decrease it based on data;
2) It can track the changes of customer groups by splitting and merging partitions, providing information about the size, duration, ancestors and descendants of each customer group;
3) It can identify behavior patterns shared by different products, which can avoid overfitting for individual products and support users to compare purchase behavior of different products.

## 5 Inference

The inference techniques for HFCP are the forward-backward algorithm (Frühwirth-Schnatter, 1994) and the posterior sampling with an augmented representation (Teh et al., 2006). We use Gibbs sampling to infer the group and behavior for each customer $(ji)$ at $t \in \{1, \ldots, T\}$ given the other customers.

Overall, our inference framework samples the following variables iteratively:

1) the block allocation $c_{ji}^{(t)}$ of each customer $(ji)$ ;
2) the pattern index $k_{jm}^{(t)}$ of each block $b_{jm}^{(t)}$;
3) the weight $\beta_k^{(t)}$ and the parameter $\phi_k^{(t)}$ of each pattern $k$.

The notations used in the inference are summarized as follows: $\rho_{\neg ji}^{(t)}$ represents the projection of $\rho_j^{(t)}$ on $A_j \setminus \{ji\}$, which refers to the set $A_j$ excluding customer $(ji)$ ; $|\rho_j^{(t)}|$ is the number of blocks in $\rho_j^{(t)}$ ; $k_{jm}^{(t)}$ is the pattern index for block $b_{jm}^{(t)}$ ; and the variables with a prime $(')$ are for the intermediate partition $\rho_j'^{(t)}$ after fragmentation operations. We provide a table of notation in supplementary material, which contains the variables, subscript and superscript and hyperparameters used in the paper.

The following subsections describe the detailed sampling process of each variable in the above list. The full inference procedure is illustrated in Algorithm 2 at the end of this section.

### 5.1 Sampling block allocation $c_{ji}^{(t)}$

Since the temporal dependency among partitions follows FCP, we adopt forward-backward (F-B) algorithm (Frühwirth-Schnatter, 1994) to infer the sequence of partitions $\rho_j^{(1:T)}$ of $A_j$ over time. The F-B algorithm is a commonly used inference framework for dynamic latent variable models. Our F-B algorithm includes a forward phase for *sampling* using Eqs. (9)–(10), with a backward phase for *smoothing* using messages in Eqs. (14)–(15). In this step, we assume the pattern allocation $\boldsymbol{k}_j^{(1:T)}$ and $\boldsymbol{\beta}^{(1:T)}$ are known and given as conditions for all customers except $(ji)$ .

### 5.1.1 Posterior distribution

When $t = 1$, we first sample block allocation $c_{ji}^{(t)}$ based on the following posterior probability:

$$\begin{aligned} &P\left(c_{ji}^{(1)} = a | x_{ji}^{(1:T)}, \rho_{\neg ji}^{(1:T)}, \rho'^{(1:T-1)}_{\neg ji}, \boldsymbol{k}_j^{(1:T)}, \boldsymbol{\beta}^{(1:T)}\right) \\ &\propto \underbrace{P\left(c_{ji}^{(1)} = a | \rho_{\neg ji}^{(1)}\right)}_{\text{conditional probability}} \underbrace{P\left(x_{ji}^{(1)} | c_{ji}^{(1)} = a, k_a^{(1)} = k\right)}_{\text{likelihood}} \underbrace{mc^{(1)}(a)}_{\text{message}} \end{aligned} \tag{8}$$

where $\rho_{\neg ji}^{(1:T)}$ refers to the sequence of partitions $\rho_j^{(t)}$ projected on $A_j \setminus \{ji\}$ from $t = 1$ to $T$. In the second line, the first term is the *conditional probability* of the block allocation given the partition of the other customers at time $t$; the second term is the *likelihood* of behavior data $x_{ji}^{(1)}$ given block $a$ and pattern $k$; and the last term is the *message*, which is the conditional probability of the observations after $t$ given $\rho_{\neg ji}^{(t:T)}$ and $\rho'^{(t:T-1)}_{\neg ji}$.

Then, we sample block allocation for fragmentation and coagulation steps iteratively. For the *fragmentation step* at $t$, the posterior distribution of allocating $(ji)$ to $a'$ is as follows:

$$\begin{aligned} &P\left(c'^{(t)}_{ji} = a' | c_{ji}^{(t)} = a, x_{ji}^{(1:T)}, \rho_{\neg ji}^{(1:T)}, \rho'^{(1:T-1)}_{\neg ji}, \boldsymbol{k}_j^{(1:T)}, \boldsymbol{\beta}^{(1:T)}\right) \\ &\propto \underbrace{P\left(c'^{(t)}_{ji} = a' | c_{ji}^{(t)} = a, \rho_{\neg ji}^{(t)}, \rho'^{(t)}_{\neg ji}\right)}_{\text{conditional probability}} \underbrace{mf^{(t+1)}(a')}_{\text{message}} \end{aligned} \tag{9}$$

where the two terms on the second line are *conditional probability* and *message*.

After that, we sample $c_{ji}^{(t+1)}$ for the *coagulation step* based on the following posterior distribution:

$$\begin{aligned} &P\left(c_{ji}^{(t+1)} = a | c'^{(t)}_{ji} = a', x_{ji}^{(1:T)}, \rho_{\neg ji}^{(1:T)}, \rho'^{(1:T-1)}_{\neg ji}, \boldsymbol{k}_j^{(1:T)}, \boldsymbol{\beta}^{(1:T)}\right) \\ &\propto \underbrace{P\left(c_{ji}^{(t+1)} = a | c'^{(t)}_{ji} = a', \rho_{\neg ji}^{(t+1)}, \rho'^{(t)}_{\neg ji}\right)}_{\text{conditional probability}} \underbrace{mc^{(t+1)}(a)}_{\text{message}} \\ &\times \underbrace{P\left(x_{ji}^{(t+1)} | c_{ji}^{(t+1)} = a, k_a^{(t+1)} = k\right)}_{\text{likelihood}} \end{aligned} \tag{10}$$

where the term on the last line is the *likelihood* of behavior data $x_{ji}^{(t+1)}$ given block $a$ and pattern $k$.

The conditional probabilities, messages and likelihood used during the inference of block allocation $c_{ji}^{(t)}$ are defined as follows.

### 5.1.2 Conditional probabilities

The conditional probability of allocating a block for customer $(ji)$ at $t = 1$ is defined as:

$$P(c_{ji}^{(1)} = a | \rho_{\neg ji}^{(1)}) = \begin{cases} n_{ja_{\neg ji}}^{(1)} / (n_{j\cdot}^{(1)} - 1 + \alpha) & \text{if } a \in \rho_{\neg ji}^{(1)} \\ \alpha / (n_{j\cdot}^{(1)} - 1 + \alpha) & \text{if } a = \phi \end{cases} \tag{11}$$

where $n^{(1)}_{ja_{\neg ji}}$ is the number of customers in $a$ excluding $(ji)$ , and $n^{(1)}_{j\cdot}$ is the total number of customers of product $j$. The condition $a \in \rho^{(1)}_{\neg ji}$ means that the selected block $a$ for customer $(ji)$ exists in $\rho^{(1)}_{\neg ji}$. The second case $a = \phi$ means that customer $(ji)$ will start a new block, without any members yet. Therefore, the possible space for $c^{(t)}_{ji}$ is $\{\phi \cup \rho^{(t)}_{\neg ji}\}$.

The conditional probabilities of allocating $c'^{(t)}_{ji}$ and $c^{(t+1)}_{ji}$ for the fragmentation and coagulation steps are:

$$P\left(c'^{(t)}_{ji} = a' | c^{(t)}_{ji} = a, \rho^{(t)}_{\neg ji}, \rho'^{(t)}_{\neg ji}\right) = \begin{cases} 1 & \text{if } a = a' = \phi \\ \epsilon |F^{(t)}(a)| / n^{(t)}_{ja_{\neg ji}} & \text{if } a \in \rho^{(t)}_{\neg ji}, a' = \phi \\ (n^{(t)}_{ja'_{\neg ji}} - \epsilon) / n^{(t)}_{ja_{\neg ji}} & \text{if } a \in \rho^{(t)}_{\neg ji}, a' \in F^{(t)}(a) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$P\left(c^{(t+1)}_{ji} = a | c'^{(t)}_{ji} = a', \rho^{(t+1)}_{\neg ji}, \rho'^{(t)}_{\neg ji}\right) = \begin{cases} \alpha / (\alpha + \epsilon |\rho'^{(t)}_{\neg ji}|) & \text{if } a = a' = \phi \\ \epsilon |C^{(t)}(a)| / (\alpha + \epsilon |\rho'^{(t)}_{\neg ji}|) & \text{if } a \in \rho^{(t+1)}_{\neg ji}, a' = \phi \\ 1 & \text{if } a \in \rho^{(t+1)}_{\neg ji}, a' \in C^{(t)}(a) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $F^{(t)}(a)$ in Eq. (12) refers to the set of blocks in $\rho'^{(t)}_{\neg ji}$ which are split from $a$; it is formally defined as $F^{(t)}(a) = \{s | s \in \rho'^{(t)}_{\neg ji}, s \subseteq a, s \neq \phi\}$. Similarly, $C^{(t)}(a)$ in Eq. (13) refers to the set of blocks in $\rho'^{(t)}_{\neg ji}$ which are merged into $a$ at time $(t+1)$; it is defined as $C^{(t)}(a) = \{s | s \in \rho'^{(t)}_{\neg ji}, s \subseteq a, s \neq \phi\}$. For any empty block (i.e. $a = \phi$), it means that $(ji)$ will start a new block.

### 5.1.3 Messages

Messages are the conditional probabilities of observations $x^{(t+1:T)}_{ji}$ if $x^{(t)}_{ji}$ is in block $a$ or $a'$. In this way, allocating $(ji)$ at $t$ considers the whole observation period, rather than just the condition at $t$. The messages for fragmentation and coagulation steps are denoted as $mf$ and $mc$, respectively. Given $mc^{(T)}(a) = 1$ for all blocks at $T$, the messages at other time intervals can be computed in a backward manner recursively according to the following equations:

$$\begin{aligned} mf^{(t)}(a') &= P\left(x^{(t+1:T)}_{ji} | c'^{(t)}_{ji} = a', \rho^{(t:T)}_{\neg ji}, \rho'^{(t:T-1)}_{\neg ji}, \boldsymbol{k}^{(t+1:T)}_{j}\right) \\ &= \sum_{a \in \rho^{(t+1)}_{\neg ji} \cup \phi} \Big( mc^{(t+1)}(a) \underbrace{P\left(x^{(t+1)}_{ji} | c^{(t+1)}_{ji} = a, k^{(t+1)}_{a} = k\right)}_{\text{Eq. (16)}} \\ &\quad \times \underbrace{P\left(c^{(t+1)}_{ji} = a | c'^{(t)}_{ji} = a', \rho^{(t+1)}_{\neg ji}, \rho'^{(t)}_{\neg ji}\right)}_{\text{Eq. (13)}} \Big) \end{aligned} \tag{14}$$

$$
\begin{aligned}
mc^{(t)}(a) =& P\Big(x_{ji}^{(t+1:T)} | c_{ji}^{(t)} = a, k_a^{(t)} = k, \rho_{\neg ji}^{(t:T)}, \rho'^{(t:T-1)}_{\neg ji}, \boldsymbol{k}_j^{(t+1:T)}\Big) \\
=& \sum_{a' \in \rho'^{(t)}_{\neg ji} \cup \phi} mf^{(t)}(a') \underbrace{P\Big(c'^{(t)}_{ji} = a' | c_{ji}^{(t)} = a, \rho_{\neg ji}^{(t)}, \rho'^{(t)}_{\neg ji}\Big)}_{\text{Eq. (12)}}
\end{aligned} \tag{15}
$$

It means that $mf^{(t)}(a')$ is computed based on $mc^{(t+1)}(a)$ for all possibilities, including all blocks in $\rho_{\neg ji}^{(t+1)}$ and an empty block. Then $mc^{(t)}(a)$ is computed based on $mf^{(t)}(a')$ of all blocks in $\rho'^{(t)}_{\neg ji}$ and an empty block. These two equations are used in turn until we get messages for customer $(ji)$ at all time steps and different cases.

#### 5.1.4 Likelihood

We use the Poisson distribution to describe the number of purchases in a period of time. Given the Poisson distribution, the likelihood of $x_{ji}^{(t)}$ with pattern $k$ is defined as:

$$
P\Big(x_{ji}^{(t)} | c_{ji}^{(t)} = a, k_a^{(t)} = k\Big) = \frac{\Big(\lambda_k^{(t)}\Big)^{x_{ji}^{(t)}} e^{-\lambda_k^{(t)}}}{x_{ji}^{(t)}!} \tag{16}
$$

We use Gamma$(\alpha_\gamma, \beta_\gamma)$ as the prior for $\lambda_k^{(t)}$, so based on the pattern allocation $\boldsymbol{k}^{(t)}$, we can estimate the intensity parameter $\lambda_k^{(t)}$ by maximum *a posteriori* (MAP):

$$
\lambda_k^{(t)} = \begin{cases} \dfrac{\sum_{ji' \in A_k} x_{ji'}^{(t)} + \alpha_\gamma - 1}{|A_k| + (1/\beta_\gamma)} & \text{if pattern } k \text{ exists} \\ \dfrac{\sum_{ji' \in A \setminus \{ji\}} x_{ji'}^{(t)} + \alpha_\gamma - 1}{|A| - 1 + (1/\beta_\gamma)} & \text{if pattern } k \text{ is new} \end{cases} \tag{17}
$$

It means that when $k$ is an existing pattern selected by previous customers, we compute $\lambda_k^{(t)}$ based on the other customers $(ji') \neq (ji)$ in $A_k$ who have pattern $k$. Otherwise, if $k$ is a new pattern, we compute $\lambda_k^{(t)}$ based on the prior knowledge and all the other customers in $A \setminus \{ji\}$.

### 5.2 Sampling pattern allocation $k_{jm}^{(t)}$

The pattern allocation starts after completing the block allocation of a sampling round. Given the block allocation $\boldsymbol{c}^{(t)}$ and the pattern weights $\boldsymbol{\beta}^{(t)}$, the pattern allocation for each block can either select an existing pattern or generate a new pattern. The conditional probabilities of the pattern allocation for block $b_{jm}^{(t)}$ are:

$$
\begin{aligned}
& P\Big(k_{jm}^{(t)} = k | \boldsymbol{c}_j^{(t)}, \boldsymbol{k}_{\neg jm}^{(t)}, \boldsymbol{\beta}^{(t)}\Big) \\
& = \begin{cases} \beta_k^{(t)} \prod_{(ji) \in b_{jm}^{(t)}} P(x_{ji}^{(t)} | k_{jm}^{(t)} = k) & \text{if pattern } k \text{ exists.} \\ \beta_u^{(t)} \prod_{(ji) \in b_{jm}^{(t)}} P(x_{ji}^{(t)} | k_{jm}^{(t)} = k) & \text{if pattern } k \text{ is new.} \end{cases}
\end{aligned} \tag{18}
$$

where $\beta_u = 1 - \sum_{k=1}^{K} \beta_k$ is the weight of a new pattern. As changing the pattern allocation of a block would affect all the elements in that block, we should consider the joint probabilities of all $\{ji\} \in b_{jm}^{(t)}$ during the selection of pattern $k$.

**Algorithm 2** The inference procedure of HFCP

**Input:** purchase behavior $\boldsymbol{X}^{(1:T)}$ of $U$ customers $J$ products, $N$ sampling iterations
**Output:** partition sequence $\rho_j^{(1:T)}$, purchase intensity $\boldsymbol{\lambda}$

**for** iteration = $1 : N$ **do**
  **for** product $j = 1 : J$ **do**
    **for** customer $i = 1 : U$ **do**
      reset all messages $\boldsymbol{mf}$ and $\boldsymbol{mc}$ and intensity $\boldsymbol{\lambda}$
      *// backward filtering: compute messages*
      **for** $t = T$ **to** 1 **do**
        compute $\boldsymbol{mf}^{(t)}$ and $\boldsymbol{mc}^{(t)}$ based on Eq. (14)–(15)
      **end for**
      *// forward sampling: block allocation* $c_{ji}^{(t)}$
      **for** $t = 1$ **to** $(T-1)$ **do**
        **if** $t == 1$ **then**
          sample based on Eq. (8), update $\rho_j^{(1)}$
          if $c_{ji}^{(1)}$ is new, sample $k_{jm}^{(1)}$ and $\beta_k^{(1)}$ for it
        **end if**
        *// fragmentation step*
        sample based on Eq. (9), update $\rho_j^{(t)}$
        if $c_{ji}^{(t)}$ is new, sample $k_{jm}^{(t)}$ and $\beta_k^{(t)}$ for it
        *// coagulation step*
        sample based on Eq. (10), update ${\rho'}_j^{(t)}$
        if ${c'}_{ji}^{(t)}$ is new, sample $k_{jm}^{(t)}$ and $\beta_k^{(t)}$ for it
      **end for**
    **end for**// *for customer* $i$
  **end for**// *for product* $j$
  update pattern allocation $\boldsymbol{k}^{(1:T)}$ based on Eq. (18)
  update weight $\boldsymbol{\beta}^{(1:T)}$ based on Eq. (19)
**end for**// *for one sampling iteration*
compute $\boldsymbol{\lambda}$ based on segmentation for each product

### 5.3 Sampling weight $\boldsymbol{\beta}_{\boldsymbol{k}}^{(\boldsymbol{t})}$

Before the end of each sampling round, the pattern weights $\boldsymbol{\beta}^{(t)}$ are sampled based on:

$$\left(\beta_1^{(t)}, \ldots, \beta_K^{(t)}, \beta_u^{(t)}\right) \sim \text{Dirchlet}\left(r_{\cdot 1}^{(t)}, \ldots, r_{\cdot K}^{(t)}, \gamma\right) \tag{19}$$

where $r_{\cdot k}^{(t)}$ denotes the number of blocks with pattern $k$ across all products at time $t$, and the probability of generating a new pattern is controlled by the hyperparameter $\gamma$.

The full inference procedure is summarized in Algorithm 2. The theoretical time complexity of a sampling round is $O(JUKT)$, which increases linearly with the number of customers $U$, the number of products $J$ and the expected number of patterns $K$ of each partition at time $t$. Compared to FCP which has a time complexity $O(UKT)$ for one sampling round, HFCP cannot parallel the modeling of multiple products of a category. This is due to the hierarchical design of HFCP, which constructs models for multiple products simultaneously by sharing patterns across them.

## 6 Experiment

We conduct empirical evaluations on two real-world purchase datasets to demonstrate the capabilities of HFCP from different perspectives. We first analyze the varying number of customer groups learned from these two datasets. Then, we examine the customer groups of different products discovered by HFCP, impacted by promotions, brand choice and change of seasons. Finally, we evaluate the performance of HFCP and compare it with three customer segmentation methods: HomoPP, NHPP (Luo et al., 2016) and FCP (Elliott and Teh, 2012; Luo et al., 2017), regarding their generalization capabilities of grouping unseen customers.

### 6.1 Experimental setup

#### 6.1.1 Supermarket dataset

The first dataset is from an Australian national-wide supermarket chain, collected through the supermarket loyalty cards between January 1 and December 31, 2014. There are 931 customers in this dataset. Each transaction contains a unique customer id, product metadata (id, category, brand and name), timestamp, purchased quantity and cost. We select 38 most popular products based on the number of customers who bought these products at least 10 times during the observation period. The selected products are from 9 categories, including 4 categories for fresh products, and the other 5 for products like soft drinks, snacks and chilled desserts.

#### 6.1.2 Dunnhumby dataset

The second dataset is collected and published by *Dunnhumby*. This set contains more than 2 million transaction records of 2500 households over two years in multiple branches of a retailer. Each transaction has a similar set of attributes as the supermarket dataset. We use the same criteria to select 66 products, which involves 1877 customers. The selected products of the Dunnhumby dataset are from 12 categories, including 4 categories for fresh products, and the other 8 for packed products.

The full lists of the product names and categories for these two datasets are provided in the supplementary material. This information can help readers to get a better understanding of the experimental datasets and result discussion.

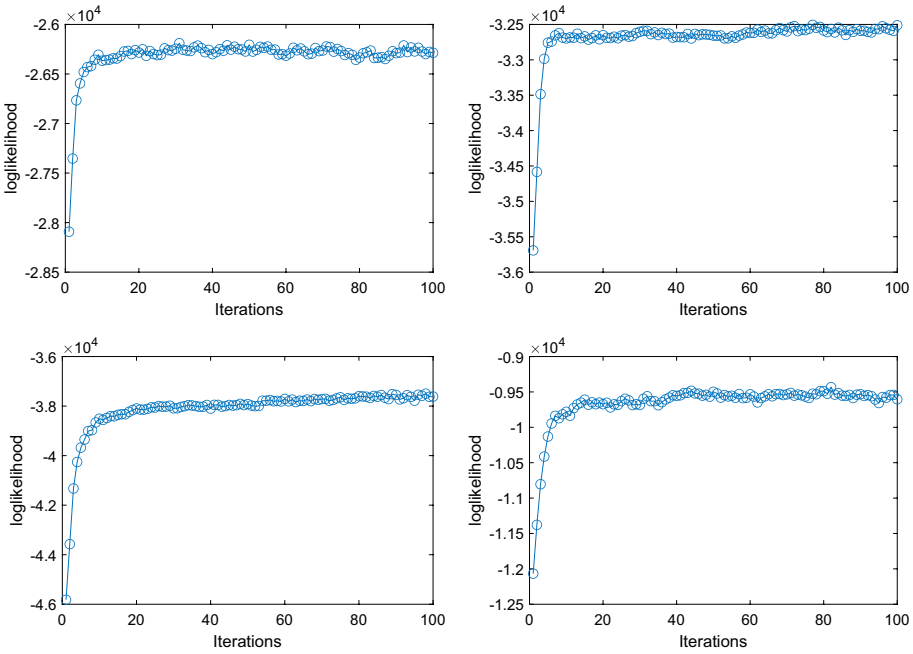


**Fig. 3** Log likelihood of each sampling iteration for 4 categories of products from supermarket dataset (top) and Dunnhumby dataset (bottom)

### 6.1.3 Hyperparameter configuration

To set the hyperparameters of HFCP, we selected the purchase records of 5 products from supermarket dataset from $t = 1$ to 3 to create a validation set. We performed grid search of $\gamma \in [0.2, 2]$, $\alpha \in [0.2, 1]$, $\epsilon \in [0.1, 1]$. We examined the fitness of HFCP model on the data based on log likelihood, and the number of groups generated after 100 sampling iterations. Based on the convergence of log likelihood and number of groups, we set the parameters as follows: $\gamma = 0.5$, $\alpha = 0.8$, $\epsilon = 0.1$. The meaning and impact of these hyperparameters are given in Sects. 4 and 3.1. For the Gamma prior of purchase intensity $\lambda_k^{(t)}$, we set shape $\alpha_\gamma = 2$ and scale $\beta_\gamma = 0.5$. The reasons is that the mode values of the number of purchase events per time unit for both datasets are 0 based on observation. To get observation data with mode 0, the number of purchases $x_{ji}^{(t)}$ should be drawn from Poisson($\lambda$) with $\lambda \in (0, 1]$. To satisfy this condition, we use Gamma(2; 0.5), so that the mode of $\lambda$ is $(\alpha_\gamma - 1)\beta_\gamma = 0.5$. The unit of time is 2-week for the supermarket dataset and 4-week for the Dunnhumby dataset.

The number of sampling iterations $N$ is 100 for HFCP. We examined the convergence of HFCP based on the log likelihood of the observed purchase records given clustering results after each sampling iteration. Figure 3 shows the log likelihood of 100 iterations for 2 categories of supermarket data and 2 categories of Dunnhumby dataset. The log likelihood converges within 40 iterations, so $N$ is set as 100 iterations in the experiment.

For comparison, the hyperparameters of FCP are set as $\alpha = 0.4$, $\epsilon = 0.1$, $\alpha_\gamma = 2$, $\beta_\gamma = 0.5$. The number of groups produced by FCP is mainly controlled by $\alpha$, so this setting

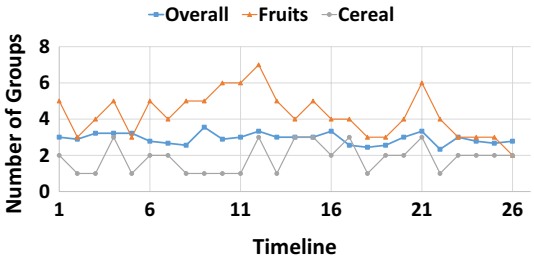


**Fig. 4** Number of customer groups for all products, fruits and cereal from supermarket dataset over time (time unit is 2-week) (Color figure online)

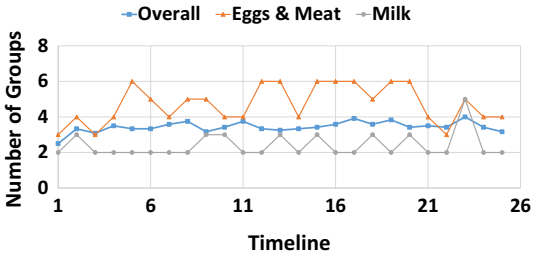


**Fig. 5** Number of customer groups for all products, eggs & meat and milk from Dunnhumby dataset over time (time unit is 4-week) (Color figure online)

makes the prior probability of generating a new customer group for one product equivalent to HFCP with $\gamma = 0.5, \alpha = 0.8$.

### 6.2 Customer segmentation results using HFCP

Our HFCP model is constructed using a category of products, and outputs the sequence of behavior patterns shared by all the customers of different products in this category. We analyze the HFCP's capabilities of tracking the changes of customer groups and learning behavior patterns shared across multiple products.

#### 6.2.1 Dynamic number of groups

As HFCP is a Bayesian nonparametric method, it can learn the number of groups required for modeling the data without model selection, which improves the flexibility of the model.

For the supermarket dataset, Fig. 4 shows the average numbers of groups at all time steps over all categories of products (blue squares), fruits (orange triangles) and cereal products (gray circles). The average numbers of groups are 2.93 for all products, 4.26 for fruits and 1.84 for cereal.

For the Dunnhumby dataset, the numbers of groups for all categories (blue squares), eggs & meat (orange triangles) and milk (gray circles) are shown in Fig. 5. The average numbers of groups are 3.4 overall, 4.72 for eggs & meat, which is the highest among all categories, and 2.4 for milk.

The main reason of having more customer groups for fruits and meat may be that the customer behavior of fruits and meat are more diverse than the other categories, which leads to higher number of distinct patterns than the other categories. On the contrary, the customers purchasing different brands of cereal or milk have similar and stable behavior. In addition, the prices of fruits vary more frequently than the other products due to the change

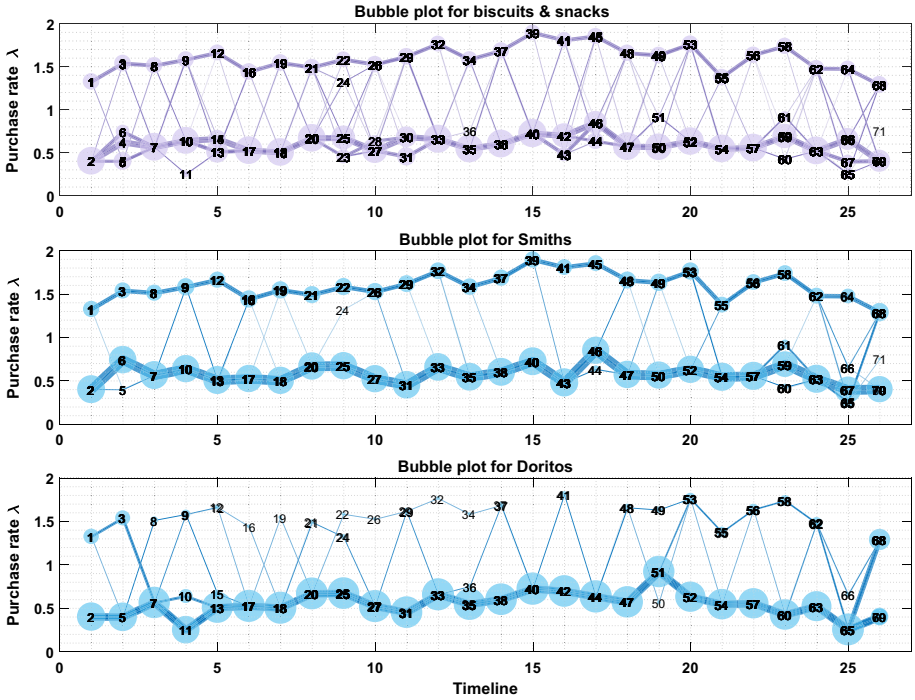


**Fig. 6** Evolution of customer groups for biscuits & snacks using HFCP. The bubble plots are for the category (top), the products Smiths potato chips (middle) and Doritos potato chips (bottom) (Color figure online)

of seasons. Previous research has found that the price elasticity of fruits is higher than cereals and milk (Andreyeva et al., 2010). When the demand of fruits of some customers is more responsive to the price changes than others, it can generate different behavior groups, so the number of groups fluctuates significantly during the observation period for fruits.

### 6.2.2 Evolution of customer behavior groups

In this section, we explore the customer groups in more depth, including the size and purchase rate of a group, how groups evolve over time, and the differences among products of a category.

We present case studies for four categories—biscuits & snacks, soft drinks, chilled desserts and fruit, to demonstrate the capabilities of HFCP in: (1) capturing purchase behavior at both product and category levels and (2) comparing purchase behavior of multiple products from different aspects, such as customers' receptiveness to promotions, brand choice and the impact of seasonal changes.

We visualize the size, purchase rate and trajectory of customer groups in bubble plots such as Fig. 6. Each bubble represents a customer group labeled by a group ID. The size of a bubble is determined by the proportion of customers in a group. The customer groups of a product at any time are exclusive and non-overlapping. The weight and transparency of the links between bubbles denote the numbers of customers switching from the left group

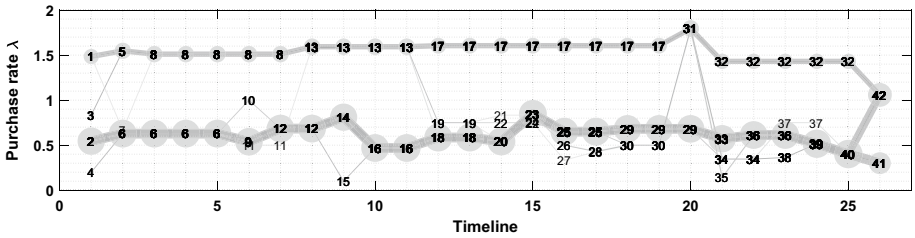


**Fig. 7** Customer segmentation for the Smiths potato chips using FCP generates minor groups that overfit the data

to the right one (the thicker and less transparent line means that more customers switch via that path).

**6.2.2.1 Biscuits and snacks—stable behavior** Figure 6 shows the trajectories of the customer groups for biscuits & snacks category in the supermarket data. There are 5 products in this category. This figure shows the customer groups for the category (top, purple) and two brands of potato chips—Smiths (middle, blue) and Doritos (bottom, blue) as examples. The top plot for the category contains all behavior patterns shared by 5 products. From the category-level plot, we find that about 75% of customers had lower purchase rates at about 0.5, while the other 25% of customers had higher purchase rates, between 1.5 and 2.

The Smiths product mainly has two types of customers with high or low purchase rates, and the customers have stable behavior with few of them switching between groups. Comparing the FCP result for the Smiths (in Fig. 7) with HFCP result (middle in Fig. 6), we notice that FCP model has many minor groups appearing around the major groups, such as groups 3, 4, 21 and 35, which are less general patterns, implying overfitting problems. As for Doritos, it has different pattern distributions from the Smiths, with more than 90% customers having lower purchase rates.

**6.2.2.2 Soft drinks—impact of promotions** The category-level trajectory of the customer groups for soft drinks in the supermarket data is shown in Fig. 8 (top, purple). There are 3 products in this category, and we show Coca-Cola (middle, blue) and Schweppes (bottom, blue) as examples in Fig. 8.

For Coca-Cola, there are three types of behavior patterns, with purchase rates at about 0.5, 2 and 5, respectively. The proportions of customers with three types of patterns are about 75%, 20% and 5% over time. We find that the customer groups with higher purchase rates appear regularly, such as the groups 1, 7, 13 and 18. To evaluate the composition of these groups quantitatively, we analyzed the intersection of every two consecutive groups to check if they contain the same group of customers. For example, groups 1 and 7, groups 7 and 13, groups 13 and 18, ..., groups 69 and 74, and there are 16 pairs of them in total. We defined intersection rate as the size of intersection of two groups over the size of the first group of the pair. The average intersection rate of 16 pairs of consecutive groups with higher purchase rates is 0.897. We also checked the intersection of the members of group 1 and groups 7, 13, 18, 25, 27, 34, and the average intersection rate is 0.892. The groups with higher purchase rates appeared about every two time steps, which is one month. This is consistent with the promotion period of this product, based on the price information. This means that these customers are receptive to the promotions, forming customer groups with higher purchase rates during promotions.

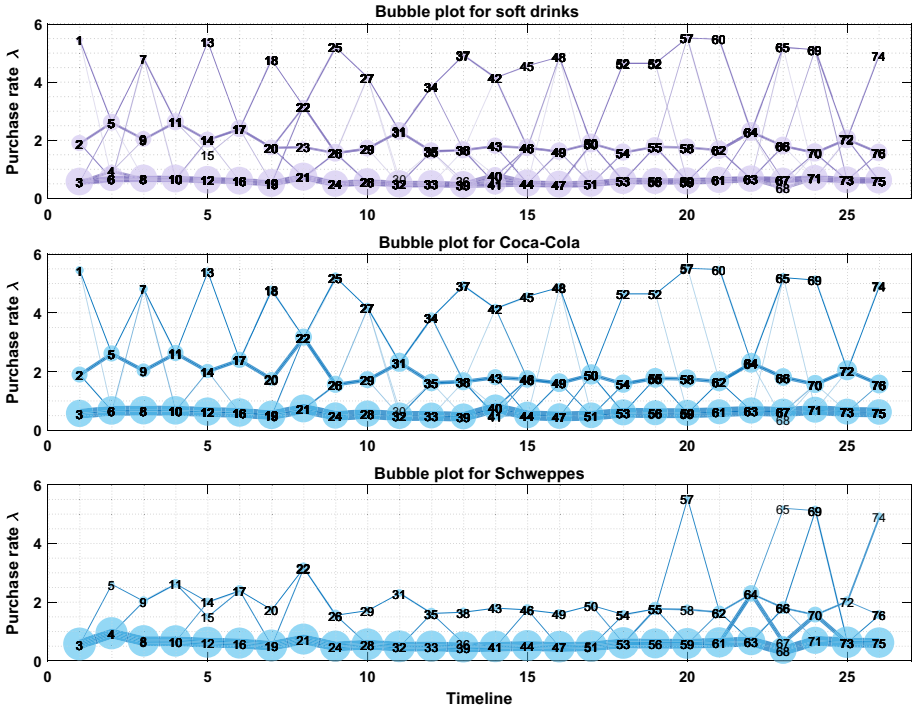


**Fig. 8** Evolution of customer groups for soft drinks using HFCP. The bubble plots are for the category (top), the product Coca-Cola (middle) and the product Schweppes (bottom) (Color figure online)

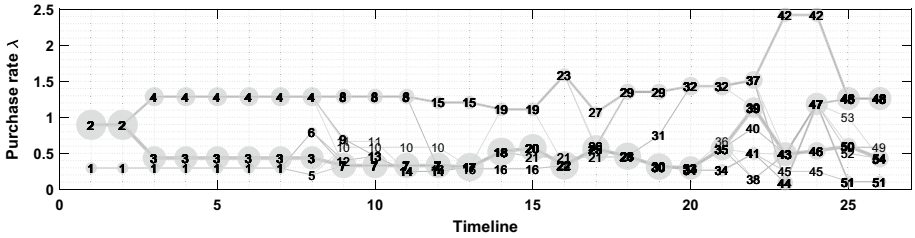


**Fig. 9** Customer segmentation for the product Schweppes using FCP generates minor groups that overfit the data

For Schweppes (bottom in Fig. 8), more than 90% of customers have lower purchase rates and less than 10% of the customers have medium purchase rates. The groups with higher purchase rates appear only after $t = 20$. The groups discovered by HFCP (bottom in Fig. 8) change smoothly over time, while the FCP (in Fig. 9) generates more minor groups at a time step, especially after $t = 21$, which implies that FCP overfits the purchase records.

**6.2.2.3 Chilled desserts—brand choice** The category-level trajectory of the customer groups for chilled desserts of Dunn-humby data is shown in Fig. 10 (top, purple). Besides ice cream, this category involves two types of yogurt, which are referred as Yogurt A and Yogurt B. These two products are from different manufacturers—Yogurt A is a national

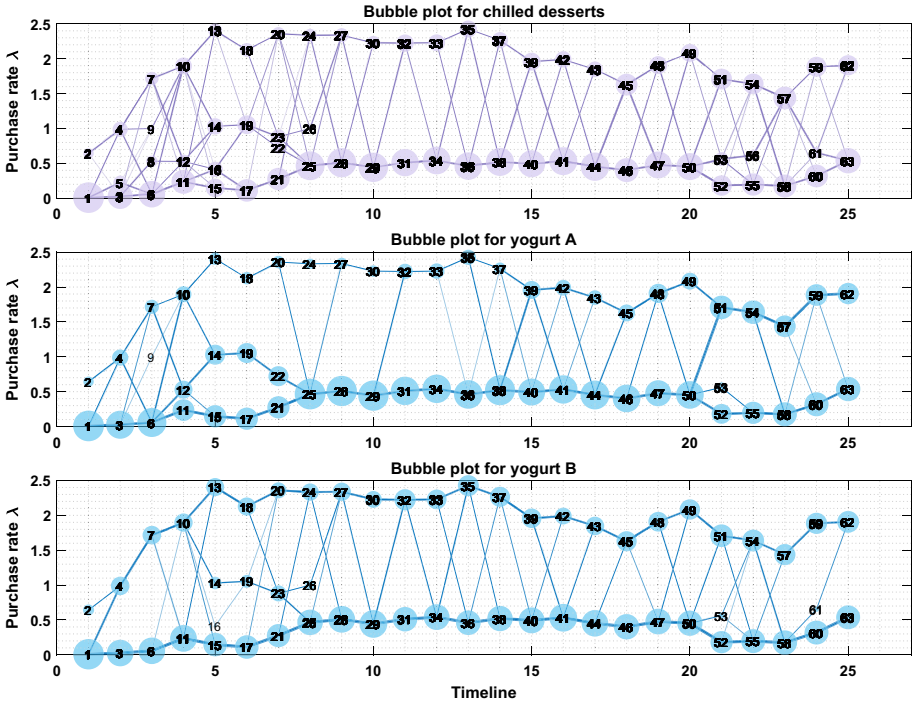


**Fig. 10** Evolution of customer groups for chilled desserts using HFCP. The bubbleplots are for the category (top), yogurt A (national brand) (middle) and yogurtB (private brand) (bottom) (Color figure online)

brand, but Yogurt B is a supermarket private brand. Comparing these two products (middle and bottom plots in Fig. 10), we notice that the customers of both products can be divided into two major groups, with high and low purchase rates, respectively. Yogurt B has a larger proportion of customers with higher purchase rates ($> 2$), and there are more links between two groups than Yogurt A. These patterns imply the impact of price and promotions on purchase behavior, as supermarket private brand usually has lower price than the national brand. When FCP is applied to customer segmentation, Fig. 11 shows that the customers with high purchase rates have been identified, with similar patterns as shown in the bottom plot of Fig. 10, but those with low purchase rates have been split into multiple groups, which may overfit the observations.

**6.2.2.4 Fruit—change of seasons** For the fruit category in the Dunnhumby data, the category-level segmentation result is shown in the top plot of Fig. 12. As there is a mixture of different purchase patterns in this category, the category-level segmentation result contains more customer groups than the other three cases as shown above. For individual product, HFCP can effectively distinguish the patterns of different products. For example, the purchase of strawberries demonstrate seasonal changes (as shown in Fig. 12, middle), with higher purchase rates in Spring and Summer (from $t = 12$ to 19). As for bananas (as shown in Fig. 12, bottom), their popularity remains stable through the whole observation periods, comparing to that of strawberries. However, as these two products have few purchase patterns in common, learning shared patterns may lead to a compromise between them. For example, when modeling strawberries using FCP (as shown in Fig. 13), the purchase rates

are about 2 for the customers who bought more strawberries (such as groups 5, 6, 11 and 24 in Fig. 13), whereas HFCP models the behavior of these customers using purchase rates that are greater than 3 (such as groups 10, 18 and 24 in the middle plot of Fig. 12), which are lifted by other popular products like bananas.

Hence, HFCP can support us to compare the purchase behavior of multiple products effectively and analyze the impact of promotions, brand choice and seasonal changes. It is also important to avoid mixing products with significantly different purchase behavior or imbalanced amount of data when building a model, as the model may underfit one or some of the products due to the compromise among products or imbalanced datasets.

#### 6.2.3 Distribution of behavior patterns

The customers of a product are split into exclusive and non-overlapping behavior groups at any time step. The distribution of customers (i.e. proportions of the customers in each group) for multiple products from the same category can be examined to understand the similarities and differences of purchase behavior of these products. For example, for the biscuits & snacks category, there are 2 groups at $t = 1$ and 4 groups at $t = 2$. The distribution of customers of all 5 products in this category at these two time steps are shown in Fig. 14. At $t = 1$, the overall distribution (the 1st row) shows that, there are 26% of the customers in group 1 (lower purchase rate) and 74% of the customers in group 2 (higher purchase rate). However, the distribution of customers buying the supermarket's own biscuits (the 5th row) is 5% and 95% in these two groups, which is quite different from overall distribution. At $t = 2$, we can discover that (1) Doritos and the supermarket's own snacks have very similar customer distributions and (2) the supermarket's own biscuits are less popular than the other products, as over 90% of the customers are in groups with low purchase rate (group 5).

In order to extend the comparison to multiple categories quantitatively, we computed the correlation between the product-level and its corresponding category-level customer distributions. Figures 15 and 16 show the correlations of all products for the supermarket dataset and Dunnhumby dataset, respectively. The range of a correlation value is between $-1$ and 1, and a higher value means that the customer distribution of that product is more positively correlated with its category-level distribution. The similarity of a product with the other products in the category can be inferred from the correlation value. For example, the product 21—grapes in the supermarket dataset has a negative correlation value, which indicates the different customer behavior of this product compared with the other products in this category. We examine the average correlation value of each category. For the supermarket dataset, as shown in Fig. 17, the lowest one is for fruits, 0.48, and the highest one is for soft drinks, 0.92. For the Dunnhumby dataset, as shown in Fig. 18, the lowest correlation is for vegetables, 0.70, while the highest two are for soft drinks and cereal, both 0.86. Therefore, different fruits and vegetables have more distinct behavior patterns, while for soft drinks and cereal, the customer behaviors of different brands are similar. This is consistent with the result that there are more groups for fruits but less groups for cereal in Sect. 6.2.1.

### 6.3 Generalizability of HFCP

The generalizability of customer segmentation models is a critical factor to consider. As it is difficult to observe the behaviour of whole population in practice, the segmentation

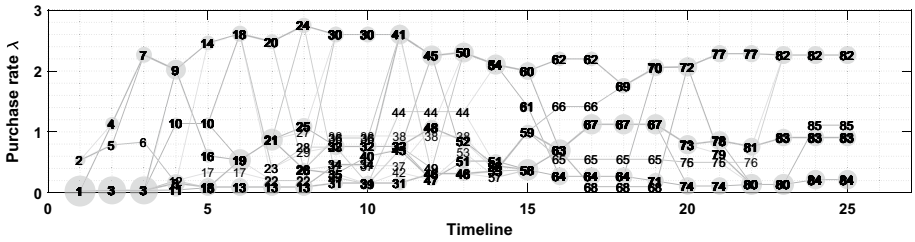


**Fig. 11** Customer segmentation for Yogurt B (private brand) using FCP generates minor groups that overfit the data

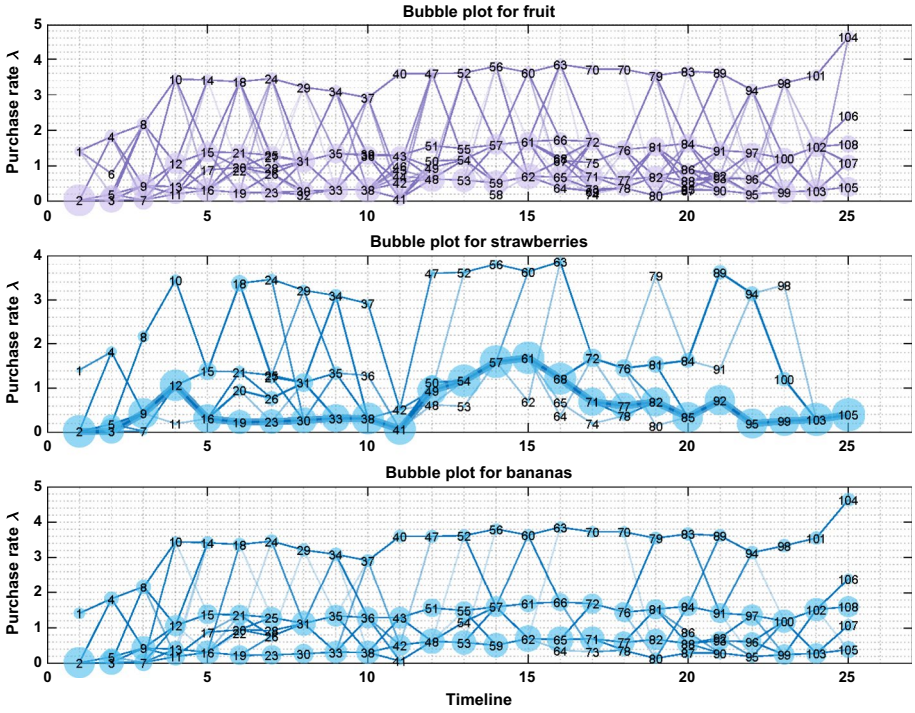


**Fig. 12** Evolution of customer groups for fruits using HFCP. The bubble plots are for the category (top), strawberries (middle) and bananas (bottom) (Color figure online)

model built on a sample of the population should be generalizable to unseen customers. Given a set of groups learned based on observed customers, future customers can be assigned to those existing groups by a generalizable model, with their behaviours accurately represented by the patterns of the allocated groups. We hold out a proportion of customers for a product during the training, and examine if the patterns learned from the remaining data of this product together with the other relevant products could group unseen customers accurately.

We compare the performance with three customer segmentation models HomoPP, NHPP and FCP. The HomoPP and NHPP are mixture of Poisson processes. Specifically, HomoPP describes each customer group using a Poisson process with a fixed intensity value. NHPP describes each customer group using a non-homogeneous Poisson process, which has an intensity function with polynomial and periodic components to capture dynamic behavior. The customer group memberships remain unchanged over time for HomoPP and NHPP. FCP is a dynamic segmentation model, but it cannot capture patterns shared by multiple products.

For HomoPP and NHPP, the number of groups is set as 3. The main reason is that the average number of groups in (Luo et al., 2016) was 3.37, which was tuned using 27 products. Moreover, the average numbers of groups discovered by HFCP for two datasets are 2.93 and 3.4, respectively. The degree of polynomial component in NHPP is set to 2 based on (Luo et al., 2016). The hyperparameters of FCP are set as $\alpha = 0.4$, $\epsilon = 0.1$, $\alpha_\gamma = 2$, $\beta_\gamma = 0.5$ as explained above.

For a target product, we randomly select 10% of its customers. The HFCP model will be trained using the records of these customers of the target product and the records of the other products in the same category. After learning the shared patterns of this category, for each customer in the hold-out set, we select an existing sequence of purchase patterns (i.e. purchase intensities of groups) that can best fit the observed sequence and measure the distance between these two sequences. The distance is measured by *Mean Absolute Errors* (MAE), which is calculated based on the absolute difference between the estimated and

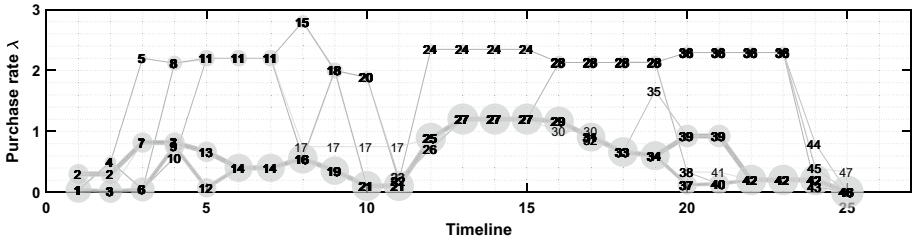


**Fig. 13** Customer segmentation for strawberries using FCP

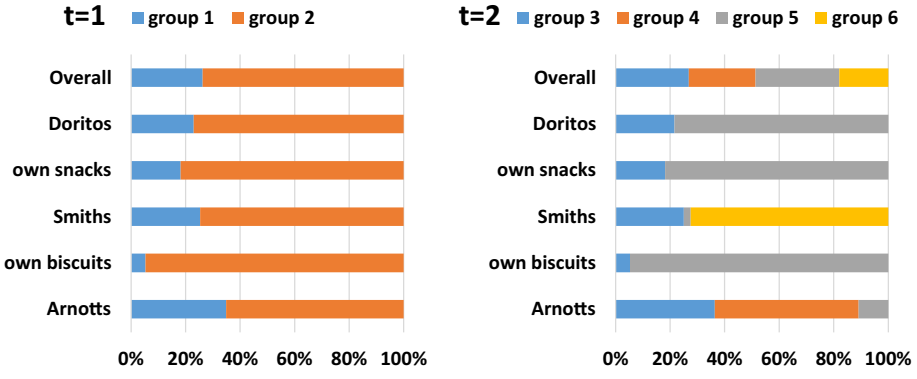


**Fig. 14** Distributions of behavior patterns of all products in biscuits & snacks category at $t = 1$ and $t = 2$ for discovering the similarities and differences of purchase behavior (Color figure online)

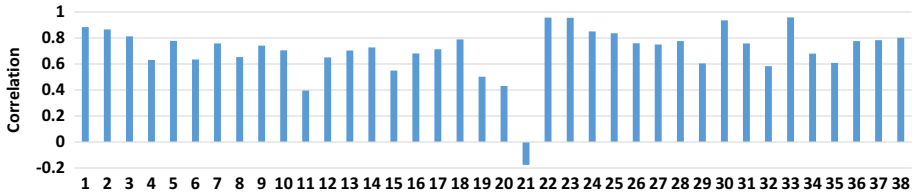


**Fig. 15** Correlation of the customer distributions between a product and the category-level average for the supermarket dataset

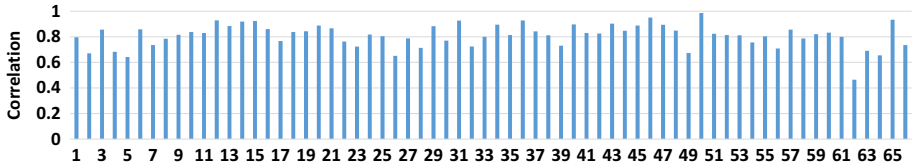


**Fig. 16** Correlation of the customer distributions between a product and the category-level average for the Dunnhumby dataset

actual numbers of purchase events per time step. We use paired t-test to verify if the performance of HFCP is statistically different from a baseline model.

For the supermarket dataset, the average MAE across all hold-out sequences for all products are shown in Fig. 19. The average MAE over all products are 0.61 (std: 0.0037), 0.64 (std: 0.0051), 0.71 (std: 0.0139) and 0.73 (std: 0.0113) for HFCP, FCP, NHPP and HomoPP, respectively. The post hoc tests (t-test using the Bonferroni correction to adjust p) show that the error of HFCP is significantly lower than that of FCP, NHPP and HomoPP ($p < 0.001$ for all three baselines). For the Dunnhumby dataset (as shown in Fig. 20), the average MAE over all products are 0.6 (std: 0.0058), 0.7 (std: 0.0116), 0.79 (std: 0.0317) and 0.83 (std: 0.0267) for HFCP, FCP, NHPP and HomoPP, respectively. The post hoc tests (t-test using the Bonferroni correction to adjust p) show that the MAE of HFCP is significantly lower than the other three methods ($p < 0.001$ for all three baselines).

We also compare the increase of accuracy after using HFCP for each category. The "increase of accuracy" can be interpreted as the gap between the blue line (HFCP) and other lines (corresponding to the other three models) in Figs. 19 and 20. The category information of each product id is provided in the Supplementary Material. For the supermarket dataset, the categories with large increases include confectionery, chilled desserts and soft drinks. For the Dunnhumby dataset, the soft drinks, chilled desserts and snacks have larger increases than categories like cheese, eggs & meat and vegetables. The reason is that the products from a category like soft drinks have similar purchase patterns. Hence, the benefit of using shared patterns from the other products in this category is more significant than the other categories, when estimating the behavior of unseen customers.

Another important aspect to consider is the runtime of different methods. HomoPP and NHPP are mixture models implemented using Expectation-Maximization algorithm. FCP and HFCP are implemented using Gibbs sampling. We collected and analyzed the time spent on modeling customer segmentation of each product from supermarket dataset for HomoPP, NHPP and FCP. As HFCP jointly models a category of products, we collected

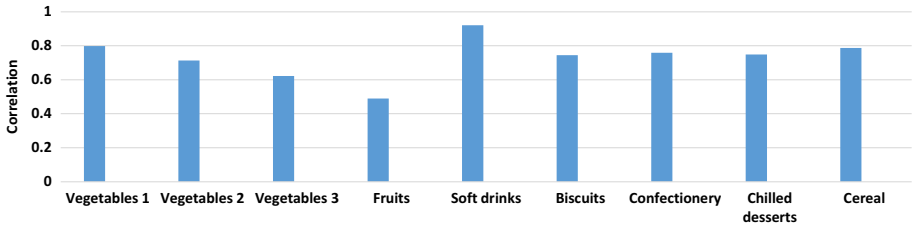


**Fig. 17** Average correlation values for all categories of the supermarket dataset

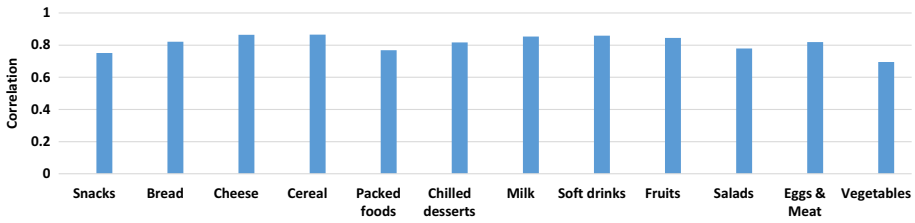


**Fig. 18** Average correlation values for all categories of the Dunnhumby dataset

the time spent on modeling each category of products for HFCP. The results were computed with a 3.6GHz 8-core computer with 32GB of RAM.

The product-level average runtime is 0.06 s (std: 0.07) for HomoPP, 37.2 s (std: 11.6) for NHPP, and 1159.2 s (std:1032.9) for FCP. It shows that HomoPP and NHPP are much more efficient compared to sampling-based method FCP. To compare with HFCP, we aggregated FCP's runtime of each product by category. Figure 21 shows the category-level runtime of FCP and HFCP. FCP is more efficient than HFCP on 8 out of 9 categories. On average, HFCP spent 14.7% more time than FCP. The possible reason is that HFCP requires more time to update category-level behavior patterns and the weight of patterns. In practice, FCP has better scalability than HFCP, because FCP can parallel the modeling of multiple products, but HFCP has to model a category of products at the same time. The longer runtime and more hyperparameters of HFCP are the trade-off of learning shared behavior patterns across multiple products and better generalizability performance.

## 7 Conclusion

Our dynamic customer segmentation model Hierarchical Fragmentation-Coagulation Processes (HFCP) uses a Bayesian nonparametric approach to detect customer purchase behavior groups, model the split and merge of groups over time, and it learns the patterns shared across products through the hierarchical structure. The two important benefits of modeling shared patterns are that it can avoid overfitting the purchase records of a single product, and it can help to compare the purchase behavior of multiple products via the distribution of different patterns.

Through the comprehensive empirical evaluations of HFCP on two real-world transaction datasets: (1) we analyzed the dynamic number of groups discovered by HFCP, which

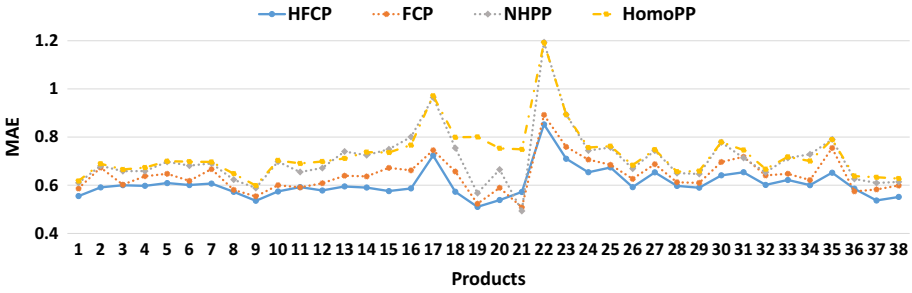


**Fig. 19** MAE of the estimations of hold-out data for all products of supermarket data using HFCP, HomoPP, NHPP and FCP (Color figure online)

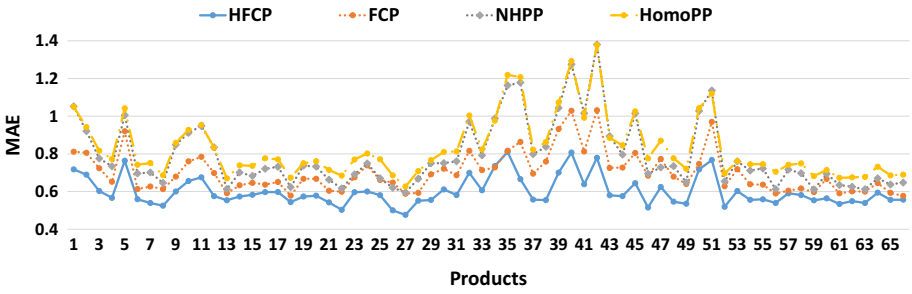


**Fig. 20** MAE of the estimations of hold-out data for all products of Dunnhumby data using HFCP, HomoPP, NHPP and FCP (Color figure online)

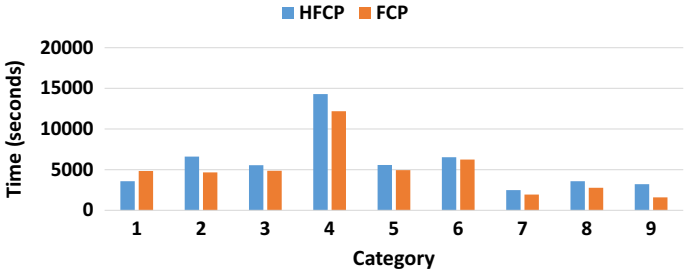


**Fig. 21** Runtime of FCP and HFCP on all products from supermarket dataset (Color figure online)

is determined automatically based on data; (2) we visualized the evolutions of customer groups at the category and product levels and explored various impact factors on customer behavior such as promotions, brand choice and seasonal changes; and (3) we demonstrated that HFCP can outperform HomoPP, NHPP and FCP on estimating the behavior of unseen customers. We found that sharing the behavior patterns learned from relevant products can help modeling a product with fewer records, and the benefit is more significant when the products have similar behavior.

In future work, the application of HFCP can be extended to study purchase behaviors across multiple categories. In addition, we would like to explore how the customers in the store influence each other's behavior by using dynamic segmentation results and flexible relational models (Fan et al., 2021).

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10994-022-06276-8.

**Author contributions** All authors contributed to the study conception and problem formulation. Model design and inference algorithm were completed by LL, BL, XF and YW. Data preprocessing, experimental setup and result analysis were performed by LL. The first draft of the manuscript was written by LL and all authors commented on the manuscript and contributed to the revision.

**Funding** This work was supported in part by the Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01) and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

**Data availability** Dunnhumby dataset is available at https://www.dunnhumby.com/sourcefiles.

**Code availability** The implementation of the proposed HFCP is available at https://github.com/lluo5436/hfcp.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

Andreyeva, T., Long, M. W., & Brownell, K. D. (2010). The impact of food prices on consumption: A systematic review of research on the price elasticity of demand for food. *American Journal of Public Health, 100*(2), 216–222.

Bertoin, J. (2006). *Random fragmentation and coagulation processes* (Vol. 102). Cambridge: Cambridge University Press.

Böttcher, M., Spott, M., Nauck, D., & Kruse, R. (2009). Mining changing customer segments in dynamic markets. *Expert Systems with Applications, 36*(1), 155–164.

Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research, 29*(2), 201–215.

Bucklin, R. E., Gupta, S., & Siddarth, S. (1998). Determining segmentation in sales response across consumer purchase behaviors. *Journal of Marketing Research, 35*(2), 189–197.

Carnein, M., & Trautmann, H. (2019). Customer segmentation based on transactional data using stream clustering. In Q. Yang, Z. H. Zhou, Z. Gong, M. L. Zhang, & S. J. Huang (Eds.), *Advances in knowledge discovery and data mining* (pp. 280–292). Cham: Springer.

Clemons, E. K., & Nunes, P. F. (2011). Carrying your long tail: Delighting your consumers and managing your operations. *Decision Support Systems, 51*(4), 884–893.

Costa, A. F., Yamaguchi, Y., Traina, A. J. M., Traina Jr, C., & Christos, F. (2015). RSC: Mining and modeling temporal activity in social media. In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, pp. 269–278.

Datta, S., Majumder, A., & Shrivastava, N. (2010) Viral marketing for multiple products. In *2010 IEEE international conference on data mining*, pp. 118–127.

Dong, D., & Kaiser, H. M. (2008). Studying household purchasing and nonpurchasing behaviour for a frequently consumed commodity: Two models. *Applied Economics, 40*(15), 1941–1951.

Du, N., Liang, Y., Balcan, M. F., Gomez-Rodriguez, M., & Zha, H. (2017). Scalable influence maximization for multiple products in continuous-time diffusion networks. *Journal of Machine Learning Research, 18*(2), 1–45.

Eguiluz, V. M., & Zimmermann, M. G. (2000). Transmission of information and herd behavior: An application to financial markets. *Physical Review Letters, 85*(26), 5659–5662.

Elliott, L., & Teh, Y. W. (2012) Scalable imputation of genetic data with a discrete fragmentation-coagulation process. In *Proceedings of conference on neural information processing systems*, pp. 2852–2860.
Elliott, L. T., Teh, Y. W., et al. (2016). A nonparametric HMM for genetic imputation and coalescent inference. *Electronic Journal of Statistics, 10*(2), 3425–3451.
Fan, X., Li, B., Luo, L., & Sisson, S. A. (2021). Bayesian nonparametric space partitions: A survey. In *Proceedings of the 30th international joint conference on artificial intelligence—survey track (IJCAI-21)*, pp. 4408–4415.
Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis, 15*(2), 183–202.
Iwata, T., Watanabe, S., Yamada, T., & Ueda, N. (2009) Topic tracking model for analyzing consumer purchase behavior. In *Proceedings of the 22nd international joint conference on artificial intelligence*, AAAI Press, pp. 1427–1432.
Kim, H., Takaya, N., & Sawada, H. (2014). Tracking temporal dynamics of purchase decisions via hierarchical time-rescaling model. In *Proceedings of the 23rd ACM international conference on information and knowledge management*, ACM, pp. 1389–1398.
Kim, H., Takaya, N., & Sawada, H. (2017). Analyzing temporal dynamics of consumer's behavior based on hierarchical time-rescaling. *IEICE Transactions on Information and Systems*, *E100-D*(4), 693–703.
Kotler, P., & Armstrong, G. (2010). *Principles of marketing*. Pearson Education.
Li, B., Zhu, X., Li, R., Zhang, C., Xue, X., & Wu, X. (2011). Cross-domain collaborative filtering over time. In *Proceedings of the 22nd international joint conference on artificial intelligence (IJCAI-11)*, pp. 2293–2298.
Luo, L., Li, B., Koprinska, I., Berkovsky, S., & Chen, F. (2016). Discovering temporal purchase patterns with different responses to promotions. In *Proceedings of the 25th ACM international conference on information and knowledge management*, ACM, pp. 2197–2202.
Luo, L., Li, B., Koprinska, I., Berkovsky, S., & Chen, F. (2017) Tracking the evolution of customer purchase behavior segmentation via a fragmentation-coagulation process. In *Proceedings of the 26th international joint conference on artificial intelligence*, pp. 2414–2420.
Netzer, O., Lattin, J. M., & Srinivasan, V. (2008). A hidden Markov model of customer relationship dynamics. *Marketing Science, 27*(2), 185–204.
Pitman, J. (2002a) Combinatorial stochastic processes. Tech. Rep. 621, Lecture Notes for St. Flour Course, Department of Statistics, UC Berkeley.
Pitman, J. (2002). Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing, 11*(5), 501–514.
Pitman, J., & Yor, M. (1997) The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pp. 855–900.
Ren, L., Dunson, D. B., & Carin, L. (2008). The dynamic hierarchical dirichlet process. In *Proceedings of the 25th international conference on machine learning*, pp. 824–831.
Ross, S. M. (1996). *Stochastic processes* (vol. 2). New York: Wiley.
Sarkar, D., Bali, R., & Sharma, T. (2018). *Customer segmentation and effective cross selling* (pp. 373–405). Berkeley, CA: Apress.
Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica, 4*(2), 639–650.
Song, H. S., Kim, J. K., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications, 21*(3), 157–168.
Taylor, A., Wilson, F., Hendrie, G., Allman-Farinelli, M., & Noakes, M. (2015). Feasibility of a healthy trolley index to assess dietary quality of the household food supply. *British Journal of Nutrition, 114*(12), 2129–2137.
Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566–1581.
Teh, YW., Blundell, C., & Elliott, L. (2011). Modelling genetic variations using fragmentation-coagulation processes. In *Proceedings of conference on neural information processing systems*, pp. 819–827.
Wang, J., & Zhang, Y. (2013). Opportunity model for e-commerce recommendation: Right product; right time. In *Proceedings of the 36th ACM conference on research and development in information retrieval*, ACM, pp. 303–312.
Xing, E. P., & Sohn, K. A. (2007). Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Analysis, 2*(3), 501–527.

## Authors and Affiliations

**Ling Luo[1] · Bin Li[2] · Xuhui Fan[3] · Yang Wang[4] · Irena Koprinska[5] · Fang Chen[4]**

Bin Li
libin@fudan.edu.cn

Xuhui Fan
xuhui.fan@newcastle.edu.au

Yang Wang
yang.wang@uts.edu.au

Irena Koprinska
irena.koprinska@sydney.edu.au

Fang Chen
fang.chen@uts.edu.au

1 School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

2 School of Computer Science, Fudan University, Shanghai, China

3 School of Information and Physical Sciences, University of Newcastle, Callaghan, Australia

4 School of Computer Science, University of Technology Sydney, Sydney, Australia

5 School of Computer Science, University of Sydney, Sydney, Australia