Modelling Customer Behaviour with Topic Models for Retail Analytics

Mariflor Elizabeth Vega Carrasco

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

November 13, 2021

I, Mariflor Elizabeth Vega Carrasco, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Topic modelling is a scalable statistical framework that can model highly dimensional grouped data while keeping explanatory power. In the domain of grocery retail analytics, topic models have not been thoroughly explored. In this thesis, I show that topic models are powerful techniques to identify customer behaviours and summarise customer transactional data, providing valuable commercial value.

This thesis has two objectives. First, to identify grocery shopping patterns that describe British food consumption, taking into account regional diversity and temporal variability. Second, to provide new methodologies that address the challenges of training topic models with grocery transactional data. These objectives are fulfilled across 3 research parts.

In the first part, I introduce a framework to evaluate and summarise topic models. I propose to evaluate topic models in four aspects: generalisation, interpretability, distinctiveness and credibility. In this manner, topic models should represent the grocery transactional data fairly, providing coherent, distinctive and highly reliable grocery themes. Using a user study, I discuss thresholds that guide interpretation of topic coherence and similarity. We propose a clustering methodology to identify topics of low uncertainty by fusing multiple posterior samples.

In the second part, I reinterpret the segmented topic model (STM) to accommodate grocery store metadata and identify spatially driven customer behaviours. This novel application harnesses store hierarchy over transactions to learn topics that are relevant within stores due to customised product assortments. Linear Gaussian Process regression complements the analysis to account for spatial autocorrelation and to investigate topics' spatial prevalence across the United Kingdom.

In the third part, I propose a variation of the STM, the Sequential STM (SeqSTM), to accommodate time sequence over transactions and to learn time-specific customer behaviours. This model is inspired by the STM and the dynamic mixture model (DMM); however, the former does not naturally account for temporal sequence and the latter does not accommodate transactions' dependency on time variables. SeqSTM is suitable for learning topics where product assortment varies with respect to time, and where transactions are exchangeable within time slices.

In this thesis, I identify customer behaviours that characterise British grocery retail. For instance, topics reveal natural groups of products that are used in the preparation of specific dishes, convey diets or outdoor activities, that are characteristic of festivities, household or pet ownership, that show a preference for brands, price or quality, etc. I have observed that customer behaviours vary regionally due to product availability and/or preference for specific products. In this manner, each constitutional country of the UK, the northern and the southern regions of England and London show a preference for different products. Finally, I show that customer behaviours may respond to seasonal product availability and/or are motivated by seasonal weather. For instance, consumption of tropical fruits around summer and of high-calorie foods during cold months.

Impact Statement

This thesis, supported by the ESRC and Dunnhumby, investigates topic models for the analysis of grocery retail data in the UK. My investigation provides insights into customer behaviours that characterise British food consumption while proposing new methodologies to handle large volumes of high-dimensional discrete data. This research is structured to address three problems. First, how to identify genuine customer behaviours through the application of topic models. Second, how to mine spatial behaviours that are perceived through regional demand and regional supply. Third, how to detect temporal patterns that respond to seasonal product availability or seasonal demand. Tackling these three problems has brought the following academic contributions:

- A clustering methodology that fuses multiple posterior samples of latent Dirichlet allocation to quantify topic uncertainty. This method is an alternative to component labelling methodologies for mixture models.

- An evaluation framework for topic models that includes four concepts: the generalisation of the model, topic coherence, *topic distinctiveness* and *topic credibility*. Here, I propose metrics for measuring the last two concepts, topic distinctiveness and topic credibility.

- The demonstration of segmented topic model and linear Gaussian process regression to accommodate store structure over transactions and to identify space-specific customer behaviours.

- A topic model named *sequential segmented topic model* that exploits sequential aggregations of discrete data. This allows the discovery of temporal customer behaviours that respond to seasonal product availability and seasonal demand.

Analysing retail data through topic models retrieves two types of outputs. Customer behaviours are discovered in the shape of product distributions, named topics, where products show different probabilities depending on their relevance to the topics. Only the products that describe a customer need show significant probabilities. Transactions and aggregation of transactions are described as proportions of the identified customer behaviours, where the most relevant behaviours get higher probabilities than those that are less relevant. These two outcomes have major commercial applications in retail analytics:

- Discovering hidden product affinities supports aisle layout, planning replenishment management, and shelf management [1, 2, 3, 4, 5]. For instance, putting highly frequently purchased items at the front of the stores enables quick pick-up and billing [6].

- A closer understanding of when and where products are frequently purchased together improves efficiencies in assortment rotation, maintains optimised stocks, achieves faster return rates, and reduces wastage on food perishables, with a positive impact on a retailer's bottom-line performance [7, 8, 9].

- Describing transactions and profiling customers using customer behaviours supports personalised marketing campaigns and recommender systems [10, 11], developing loyalty programs and tailoring offers such as coupons and promotions [8].

- Understanding how products satisfy customer needs provides insights to develop up-selling and cross-selling campaigns that aim to augment ticket value.

- Identifying changes in customer behaviours can help to establish effective promotion campaigns [12], i.e, price reductions for special events or holidays.

The analysis of retail data through topic models not only brings insights with commercial implications but also provides new venues for sociological, cultural and public health studies. For example:

- Describing the nation's grocery consumption in terms of topic proportions reveals social trends such as vegetarianism or consumption of organic foods. Topic models can provide topic summaries for geographical areas such as regions or middle layer super output areas (MSOAs), providing a spatial comparison of the magnitude of such food preferences.

- Describing British food consumption though topics can also reveal cultural patterns. For instance, people in Britain not only eat 'roast dinner' but also other international dishes such as 'stir fry' or 'fajitas'. These dishes are so popular in the retail data that they are captured as individual customer behaviours.

- Summarising transactional data using topics can help public health investigators to track temporal or spatial changes in the consumption of topics with high content of sugar, alcohol, salt and fat. In comparison to analysis of individual products, topics show the interaction of products from different categories, i.e., a topic with high content of sugar and fat. Topic modelling provides a new way to analyse food consumption, which otherwise, depends on highly subjective and expensive questionnaires.

The customer behaviours identified in this work characterise the British grocery retail industry. However, our methods can be used to analyse other markets, such as the financial market, and grocery or other industries in other nations.

These investigations lead to three paper submissions to statistics journals: Journal of the Royal Statistical Society: Series C, the Annals of Applied Statistics and Public Library of Science (PLOS). The first paper, which compiles findings in Chapter 4, is under the second round of revision. The second paper, which compiles findings in Chapter 5, is ready for submission. I am currently preparing a third paper that contains the work presented in Chapter 6.

Acknowledgements

I immensely thank Dr Ioanna Manolopoulou and Dr Mirco Musolesi for their support and supervision. I would like to thank the ESRC, the Economic and Social Research Council (ESRC), Dunnhumby and the Alan Turing Institute for the financial support that made my PhD possible. In Particular, Rosie Prior and Jason O'Sullivan, who facilitated the partnership between UCL and Dunnhumby.

A special acknowledgement to Dr Ioanna Manolopoulou for her advice, encouragement, and friendship. She has been a role model and an inspiration for my career. Her academic and moral support have been crucial to the completion of my research.

Another special acknowledgement to Robert Dunne for all uncountable cups of tea, for reading all my papers, and for standing by me in the hardest times. Finally, I dedicate this achievement to Flora Carrasco Morales, Edgar Vega Alcazar, Edgar Vega Carrasco and Pompona, my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The grocery retail industry is a vital part of the British economy. In 2019, £167 billion was spent in predominantly food stores, 42.4% of total retail sales in Great Britain [13]. The grocery retail industry keeps growing, the total retail sales have increased by 25% since 1994 [13]; and the average weekly household expenditure in food and drinks increased 1.8% in 2017-2018 compared to 2016-2017 [14]. This industry is highly competitive; companies such as Tesco, Sainsbury's, Asda, and Morrison, known as 'the Big Four', together hold 66.3% per cent of the grocery retail market, while at least 7 grocery retailers share the remaining 33.7% of the market [15]. Tesco, the leading company, has lost approximately 4% of its market in the last 8 years, while discount retailers such as Aldi and Lidl have both been consistently gaining market share [16]. Thus, retailers are constantly under pressure to keep competitiveness while providing value to their stakeholders.

Understanding the motivations and dynamics behind customer behaviours can unlock business opportunities for retailers that aim to keep competitive while delivering improved customer experience and increasing efficiency across business operations. For example, profiling customers and stores through the analysis of customer behaviours supports personalised marketing campaigns and recommender systems [10, 11]. A deep understanding of customer behaviour supports the development of loyalty programmes and tailoring offers such as coupons and promotions. Therein, retailers engage with customers by rewarding loyalty and stimulating voucher redemption [8]. Understanding how products satisfy customer needs provides insights that allow the development of up-selling and cross-selling campaigns, that aim to augment the ticket value by encouraging the purchasing of premium products or new related products. Identifying changes in customer behaviours can also help managers to establish effective promotion campaigns [12], i.e., price reductions for special events or holidays.

Discovering hidden product affinities and identifying the driving motivations behind shopping trips supports planning aisle layouts [1], replenishment management [1, 2, 3], and shelf management [4, 1, 5]. For instance, putting highly frequent items at the front of the stores enables quick pick-up and billing. Financial performance can be improved by strategically positioning high-demand categories and co-occurring items [6]. A closer understanding of customer behaviour improves efficiencies in assortment rotation [7], maintains optimised stock, achieves faster return rates, and reduces wastage on food perishables [8], with a positive impact on a retailer's bottom-line performance [9].

However, identifying the hidden product affinities that satisfy customer motivations is not an easy task. Retail data are characterised by high volumes of transactions generated every day, and thousands of products are available to customers who wish to satisfy one or multiple shopping needs. In 2017, Tesco reported 79 million shopping trips per week [17] while stocking around 40,000 product lines, of which 25,000 are food and beverages. In the same year, Asda reported 18 million customers shopping every week [18] and offered 35,000 products in their superstores on average [19].

The analysis of transactional data involves high-dimensional sparse vectors over thousands of products. For instance, a customer who goes to the supermarket to buy ingredients to make a cake has to choose a few products out of hundreds if not thousands. Say that this customer only buys eggs, flour, and butter; this transaction can be represented by a binary vector where the purchased products are represented by ones while the remaining thousands of products in the product assortment are represented by zeros. Considering millions of transactions, retail transactional data represent an extremely sparse and vast data matrix where almost all elements are zero. Because of sparsity and high dimensionality, linear models are difficult to interpret since features are different for every individual while non-linear models, in general, are difficult to interpret [20].

Typically, the analysis of transactional data and the identification of customer behaviours are contained in the framework of Market Basket Analysis (MBA). MBA uncovers associations between co-occurring items in a 'market basket'. A market basket is a collection of items purchased by a customer given a time interval. Tracking customer purchases is not always possible without a customer identifier such as a loyalty card. Associations Rules (AR) [21, 22] is the most popular technique in MBA. An association rule is, for example, the statement of 90% of the transactions that purchase bread and butter also purchase milk. AR has been extended in many directions to include multi-store environment [22], temporal effects [23], and uncertainty of spatial information [24]. AR is a frequentist technique that only expresses product co-occurrence and does not necessarily convey shopping needs. Moreover, association rules do not convey information about the multiple motivations behind individual transactions.

In this thesis, we investigate topic modelling (TM), a Bayesian framework, that can model customer behaviours while dealing with the complexity of retail data. TM was originally introduced to address the problem of processing, understanding, and summarising documents by identifying the hidden 'topics' in a large collection of text data [25, 26, 27, 28]. Statistically, topics are probability distributions over a fixed vocabulary; in that way, topic-relevant words are reflected by large probabilities. Documents are then summarised as probabilistic mixtures of topics where the document-relevant topics show large probabilities.

Interpreting TM into retail analytics replaces documents and words for transactions and products respectively. Thus, topics are probability distributions over a fixed product assortment and transactions are described as probabilistic topical mixtures. TM describes the highly multidimensional data into a finite number of topics that represent customer behaviours; and at the same time, offers a systematic manner to summarise transactions based on their underlying shopping motivations such as foods for breakfast, goods for a barbecue, buying seasonal produce, ingredients for cooking specific dishes, food for the pet, etc.

Topic models have seldom been applied to retail data. For example, [29, 30, 31, 11], apply latent Dirichlet allocation (LDA), the vanilla topic model, to identify customer behaviours using product categories, dismissing the high resolution of topic models when trained on individual products [32]. Few works have been entirely motivated by retail data. For instance, [33] proposes a model to capture interaction among items and answers counterfactual queries about pricing; [34] combines the correlated topic model with vector autoregression to account for product, customer and time dimensions presented in purchase history data.

In this thesis, we apply topic models to grocery retail transactions without any categorisation; so that, we identify specific products that altogether satisfy specific shopping needs demonstrating customer behaviours. Transactions are not linked to customers or other transactions in our analysis. Despite this, we identify customer behaviours from transactional data.

We discuss the technical challenges of fitting topic models to retail data and propose new methodologies for summarising and evaluating topic models. We also investigate topic models to identify customer behaviours that vary across space and time, proposing and extending topic models to accommodate store-hierarchy or sequential time structure over transactions. As we will show, this investigation provides customer insights with commercial implications and offers new means for social, cultural and dietary research.

1.1 Research questions

In the journey of investigating topic models for retail analytics, we explore the following research questions:

- Is topic modelling a useful framework to identify customer behaviours through the analysis of large volumes of transactional data?

- What are the main customer behaviours that characterise food consumption in the British retail market?

- Are there any customer behaviours that respond to regional supply and regional demand?

- How do customer behaviours vary according to seasonal product availability and seasonal demand?

1.2 Thesis contribution

Aiming to answer the research questions, we have found challenges and opportunities that are discussed in the following four areas.

1.2.1 Summarising topic models

Topic models are powerful techniques that can identify customer behaviours while coping with the sheer volume and high dimensionality of transactional data. Given the complexity of topic models, exact inference is intractable and approximation techniques such as Gibbs sampling are needed to estimate topic distributions. Summarising the posterior distribution of a topic model is challenging because there is no guarantee of correspondence between individual topics across samples, due to inherent posterior variability; therefore, estimates of the topic distributions cannot be combined across samples for any analysis that relies on the content of specific topics [26]. Empirically, we have observed that topics, depending on their variability, may show significant distributional differences across posterior samples. As such, averaging posterior samples is not advisable since topics reflecting different semantic concepts could be merged. We have also found that repetitions of the same topic model may exhibit significant variations; a topic associated with a particular semantic concept (in our case, a customer behaviour) may appear and disappear across posterior samples of different Gibbs samplers, concurring with [35, 36, 37].

In response, we propose a new methodology to summarise topic distributions while exploiting posterior samples of various Gibbs samplers. This methodology follows a hierarchical clustering that merges topics from different posterior samples, and the number of clusters is not contained by the size of the original topic models. This methodology uses cosine distance as a measure of similarity as it has been reported to correlate with human judgement on topic similarity [38]. Up to a cosine distance threshold, clusters gather topics that are assumed to represent the same customer behaviour; thereby, the cluster size provides a measure of uncertainty, i.e., a recurrent topic that appears 20 times out of 20 posterior samples is a highly certain topic, as opposed to, a highly uncertain topic that appears once out of 20 posterior samples. Topic recurrence as a measure of (un)certainty allows users to disregard uncertain topics, which can be less semantically meaningful and may not represent genuine themes [39, 40].

1.2.2 Evaluating topic models

Typically, the assessment of topic models is dominated by measures that quantify the generalisation capability of the inferred topics such as held-out log-likelihood or perplexity [41, 42]. Alternatively, topic models can be assessed by the interpretability of individual topics [43], measured by point-wise mutual information metrics. Complementing model generalisation and topic coherence, we propose to evaluate topics by their distinctiveness and credibility. Topic distinctiveness measures semantic dissimilarity among topics and topic credibility quantifies distributional similarity among runs of the same model. Utilising a tailored survey with experts in retail analytics, we provide thresholds of topic coherence and topic similarity that guide users to interpret the performance of topic models in the domain of grocery retail data.

1.2.3 Learning customer behaviours with spatial patterns

Understanding the customer behaviours behind transactional data has significant commercial value in the grocery retail industry. Topic models have been proven to be a powerful tool in the analysis of transactional data, identifying topics that display frequently-bought-together products, and summarising transactions as mixtures of topics. Applications of topic models to retail data have mainly exploited the vanilla topic model, namely, latent Dirichlet allocation

(LDA) [25]. LDA does not exploit spatial metadata and assumes that all transactions are exchangeable. However, transactions not only happen nationwide but also show customer behaviours that vary geographically due to customised product assortments that are designed to fulfil local demand and local supply. Thus, a fine-grained analysis of transactional data should accommodate spatial variations that are constrained by the stores' product assortments.

In response, we propose the novel application of the segmented topic model (STM) [44] to retail data. STM extends LDA to accommodate hierarchy over transactions, so that transactions are only exchangeable within their store. In this manner, STM learns topics that are not only relevant nationwide but also those that are relevant at specific stores or areas. STM topics are clustered and filtered according to their recurrence. STM learns store-specific topical mixtures that can be used to identify customer behaviours with spatial patterns by mapping topic probabilities at store's locations. Linear Gaussian process regression complements the spatial analysis of topics by modelling geographical prevalence across the UK while accounting for spatial autocorrelation.

1.2.4 Identifying time-varying customer behaviours

Most topic model applications to retail data do not exploit time metadata (transaction purchasing time), dismissing the temporal patterns of grocery consumption that are driven by seasonal product availability or by season-driven customer motivations, i.e., customers buying tropical fruit during summer or consuming high-calorie foods during cold months. While each transaction can be ordered according to their purchased time, a time sequence cannot be assumed because transactions are purchased by independent customers. However, transactions could be partially ordered given time slices. That is to say, transactions are exchangeable within their time slice, i.e, it is the same if a transaction happens at the beginning or at the end of the time slice, but time slices must follow an ordered sequence. Moreover, time-specific topical mixtures are of interest since they summarise time-variant topical composition.

We propose a new topic model called the *sequential segmented topic model*

(SeqSTM), which is a variant of the segmented topic model. SeqSTM handles hierarchy over transactions and learns time-specific topical mixtures, but also assumes a time sequence in a first-order Markov fashion. In other words, SeqSTM modifies the prior distribution of a time slice taking into account the previous time-specific topical mixture, allowing computationally efficient smoothing over time. SeqSTM also assumes that transactions are conditionally independent of each other given their associated time-specific topical mixture. Inference for SeqSTM is solved by a block Gibbs sampler algorithm [45] and a modified Dirichlet prior as in [46]. SeqSTM topics are clustered and filtered according to their recurrence. SeqSTM is capable of identifying time-specific behaviours that are overlooked by other, simpler models such as the STM and LDA. Moreover, the analysis of time-specific topical mixtures reveals customer behaviours with festive, seasonal and periodic patterns. Without a time framework, these temporal patterns would not be evident.

1.3 Thesis overview

This thesis has the following structure:

- **Chapter 2** presents the Dirichlet family and the fundamentals of the Dirichlet distribution, the Dirichlet process, and the Poisson-Dirichlet process. The Dirichlet distribution and its conjugacy property are key elements in the formulation of the latent Dirichlet allocation. The Dirichlet process paves the way to comprehend the Poisson-Dirichlet process. The Poisson-Dirichlet process is a key element in the formulation of the segmented topic model and sequential segmented topic model.

- **Chapter 3** introduces the basis of topic modelling and summarises the most popular topic models. Here, we interpret the latent Dirichlet allocation and the segmented topic model in terms of retail data and describe their inference algorithms based on collapsed Gibbs sampling methods.

- **Chapter 4** presents the application of the latent Dirichlet allocation to retail data. LDA outcomes are evaluated on four aspects: model generalisation, topic coherence, topic distinctiveness and topic credibility. This chapter also presents a clustering methodology that summarises the posterior distribution of LDA while quantifying topic uncertainty. We provide thresholds to interpret quality aspects of topics that are relevant to retail analytics.

- **Chapter 5** presents the application of the segmented topic model to retail data, which can accommodate store hierarchy over transactions and identify customer behaviours with spatial patterns. Linear Gaussian regression is used to quantify the topical prevalence over regions while accounting for spatial autocorrelations.

- **Chapter 6** introduces a new topic model, the sequential segmented topic model, which exploits time hierarchy over transactions taking into account the temporal sequence of time slices. The sequential segmented topic model allows for the identification of customer behaviours with temporal patterns while describing topic variability across time.

- **Chapter 7** summarises this thesis and discusses its limitations and future work.

Chapter 2

The Dirichlet Family

In this chapter, the members of the Dirichlet family are introduced. The Dirichlet distribution and its conjugacy with the multinomial distribution are key elements to model topical mixtures and topic distributions. The Dirichlet process (DP) is an extension of the Dirichlet distribution. DP is not used directly in our analysis of transactional data, but it paves the way to the two-parameter Poisson-Dirichlet Process (PDP). The PDP is fundamental in topic models where topical mixtures derive from other topical mixtures, i.e., transaction-specific topical mixtures that derive from time-specific topical mixtures. The PDP has useful representation through the Chinese restaurant process, which allows the construction of efficient inference algorithms, working as a bridge between the Dirichlet and multinomial distributions.

2.1 The Dirichlet distribution

The Dirichlet distribution [47, 48, 49] defines a probability distribution on a space of all finite probability vectors which has a support over the $k-1$-dimensional probability simplex defined by $\Delta_k = (\theta_1,\ldots,\theta_k) : \sum_{i=1}^{k}\theta_i = 1, \theta_i \geq 0$. The Dirichlet Distribution is reduced to the Beta distribution when $k = 2$.

2.1.1 Definition

Definition 2.1.1 (Dirichlet Distribution)

Let $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k) \in \Delta_k$ be a random vector distributed according to a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)$ and $\alpha_i > 0$ for $i = 1,\ldots,k$, denoted by $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$, if its probability density function is:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{1}{\text{Beta}_k(\boldsymbol{\alpha})} \prod_k \theta_k^{\alpha_k - 1}, \tag{2.1}$$

where $\text{Beta}_k(\boldsymbol{\alpha})$ is a k-dimensional Beta function that normalises the Dirichlet, defined by:

$$\text{Beta}_k(\boldsymbol{\alpha}) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}, \tag{2.2}$$

where $\Gamma(t) = \int_0^\infty x^{t-1} e^x dx$.

2.1.2 Properties

2.1.2.1 Moments, marginal and mode

The mean, variance, covariance, marginal and mode are defined by:

$$\begin{aligned}
\mathbb{E}[(\theta_1, \dots, \theta_k)] &= \left(\frac{\alpha_1}{\alpha_0}, \dots, \frac{\alpha_k}{\alpha_0}\right) \\
\mathbb{V}[\theta_i] &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(1 + \alpha_0)} \\
\mathbb{C}ov[\theta_i, \theta_j] &= \frac{-\alpha_i \alpha_j}{\alpha_0^2(1 + \alpha_0)} \quad i \neq j \\
\theta_i &\sim Beta(\alpha_i, \alpha_0 - \alpha_i) \\
Mode(\theta_1, \dots, \theta_k) &= \left(\frac{\alpha_1 - 1}{\alpha_0 - k}, \dots, \frac{\alpha_k - 1}{\alpha_0 - k}\right)
\end{aligned} \tag{2.3}$$

where $\alpha_0 = \sum_{i=1}^k \alpha_k$.

The $\boldsymbol{\alpha}$ parameters of the Dirichlet distribution can be parametrised by a base measure $\mathbf{m} = \mathbb{E}[\boldsymbol{\theta}]$ and a precision parameter α_0, so $\boldsymbol{\alpha} = \alpha_0 \mathbf{m}$ [50]. A uniform base measure indicates $\theta_i = \frac{\alpha_0}{k} \forall i$. The precision parameter controls the extent to which Dirichlet samples differ from the base measure $\mathbf{m}$, large precision α_0 decreases the variance of the distribution. α_k controls the relative likelihood of the i^{th} component, small $\alpha_1, \dots, \alpha_k$ retrieves sparse random vectors (with few components with high probability).

2.1.2.2 Conjugacy

The Dirichlet distribution is a conjugate prior of the multinomial distribution. The multinomial distribution is a discrete distribution over k mutually exclusive categories with probabilities $\theta_1, ...\theta_k$. Given a sequence of n independent samples, $\mathbf{n}$ is a random vector of category occurrence $n_1, \ldots, n_k$ with $\sum_{i=1}^{k} n_k = n$. Then, $\mathbf{n}$ is multinomial distributed, denoted $\mathbf{n} \backsim \text{Multinomial}(\boldsymbol{\theta})$, with probability mass function:

$$p(\mathbf{n} \mid \boldsymbol{\theta}) = \frac{n!}{n_1! \ldots n_k!} \prod_k \theta_k^{n_k}. \tag{2.4}$$

Let $\boldsymbol{\theta}$ be Dirichlet distributed with hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$. Then the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{n}$ is Dirichlet with hyperparameters $(\theta_1 + n_1, \ldots, \theta_k + n_k)$, denoted by $\boldsymbol{\theta} \mid \mathbf{n} \sim Dir(\boldsymbol{\alpha} + \mathbf{n})$.

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{n}) &\propto p(\mathbf{n} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &\propto \Big(\frac{n!}{n_1! \ldots n_k!} \prod_{n=1}^{k} \theta_i^{n_i}\Big)\Big(\frac{1}{Beta_k(\boldsymbol{\alpha})} \prod_k \theta_i^{\alpha_i - 1}\Big) \\ &\propto \prod_k \theta_i^{n_i + \alpha_i - 1}. \end{aligned} \tag{2.5}$$

2.1.2.3 Aggregation

Aggregating parts of the Dirichlet sample space results in a new partition of the space that is also Dirichlet distributed.

Let $(\theta_1, \ldots, \theta_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$ and $I_{1:m}$ is a partition of $1, \ldots, k$, then

$$\Big(\sum_{i \in I_1} \theta_i, \ldots, \sum_{i \in I_m} \theta_i\Big) \sim Dir\Big(\sum_{i \in I_1} \alpha_i, \ldots, \sum_{i \in I_m} \alpha_i\Big). \tag{2.6}$$

2.2 The Dirichlet process

The Dirichlet process (DP) is not used in our applications of topic modelling to retail data but paves the way to understand the Poisson-Dirichlet process which is used in Sections 5 and 6.

2.2.1 Definition

Definition 2.2.1 (Dirichlet Process)

Let H be a random measure on a measurable space $(\chi,\mathbb{B})$ and b be a positive real number. Then, a random probability measure G on $(\chi,\mathbb{B})$ is a Dirichlet process with base measure H and a concentration parameter b, denoted by $G \sim DP(b,H)$, if for any finite measurable partition $(B_1,\ldots,B_k)$ of χ, the random vector $(G(B_1),\ldots,G(B_k))$ is Dirichlet distributed with parameter $(bH(B_1),\ldots,bH(B_k)))$:

$$(G(B_1),\ldots,G(B_k)) \sim (bH(B_1),\ldots,bH(B_k))). \tag{2.7}$$

The DP is equivalent to the Dirichlet distribution if H is a probability vector over a finite space,

$$DP(b,\text{Discrete}(H)) = \text{Dir}(bH). \tag{2.8}$$

Thus, the DP is an extension of a Dirichlet distribution.

2.2.2 Properties

2.2.2.1 Moments

For any measurable set $B \in \mathbb{B}$, the mean, variance and covariance of $G \sim DP(b,H)$ are given by:

$$\begin{aligned}\mathbb{E}(G(B)) &= H(B)\\ \mathbb{V}(G(B)) &= \frac{H(B)(1-H(B))}{b-1}\\ \mathbb{C}ov(G(B),G(B')) &= \frac{-H(B)(H(B'))}{b-1} \quad \forall B' \cap B = \emptyset.\end{aligned} \tag{2.9}$$

The base measure is the mean of the DP. The concentration parameter b, also called *precision* in Section 2.1, controls the variance between G and H. Thus, large values of b indicate that the DP concentrates more mass around the base measure H.

2.2.2.2 Discreteness

Distributions drawn from a DP are discrete with probability one [47]. Thus, samples from $G \sim DP(b,H)$ have a strictly positive probability of being redrawn. Thus, samples exhibit a clustering property [51].

2.2.2.3 Conjugacy

The posterior distribution of DP is a DP with updated concentration parameters and base measure. Given $\theta_1,\ldots,\theta_n$ be i.i.d samples from $G \sim DP(b,H)$, posterior distribution of the DP is:

$$G \mid \theta_1,\ldots,\theta_n \sim DP\Big(b+n, \frac{b}{b+n}H(\cdot) + \frac{n}{b+n}\frac{\sum_{i=1}^{n}\delta_{\theta_i}(\cdot)}{n}\Big). \tag{2.10}$$

2.3 The Poisson-Dirichlet process

The two-parameter Poisson-Dirichlet process (PDP), also known as Pitman-Yor process [52, 53, 54], is a two-parameter generalisation of the DP. The PDP is a probability distribution over distributions over a measurable space $(\chi,\mathbb{B})$, takes a base distribution $H(\cdot)$ over the measurable space with domain χ, and returns a discrete distribution with a finite or countable infinite subset of χ.

The PDP is parametrised by a the discount parameter, b the concentration parameter and a random base measure $H(\cdot)$ over χ, denoted as $PDP(a,b,H(\cdot))$, where a and b control the amount of variability around $H(\cdot)$. The PDP extends the DP as is the special case when the discount parameter a is 0.

Interestingly, the PDP behaves according to a ‘power law’; the tail of a distribution drawn from the PDP is much longer than that drawn from the DP [55]. Thus, PDP is more suitable than the DP for natural language processing applications in which words follow a ‘power law’ behaviour [56].

2.3.1 Definition

Definition 2.3.1 (The Poisson-Dirichlet Process)

We say random measure G is a Poisson-Dirichlet process with parameters a, b and base distribution $H(\cdot)$ over some measurable space $(\chi,\mathbb{B})$, denoted by $G \sim PDP(a,b,H(\cdot))$, if a probability vector $\mathbf{p} = (p_1, p_2, ...)$ is drawn from a Poisson-Dirichlet distribution with parameters a and b, and unique i.i.d samples $X_1^\star, ..., X_K^\star$ from a base distribution $H(\cdot)$ define a discrete distribution on χ given by:

$$G = \sum_{k=1}^{\infty} p_k \delta_{X_k^\star(\cdot)}, \tag{2.11}$$

where $0 \le p_k \le 1$ and $\sum_k^\infty p_k = 1$. $\delta_{X_k^\star}$ is a discrete measure concentrated at $X_k^\star$.

Definition 2.3.2 (The Poisson-Dirichlet distribution)

For $0 \le a < 1$ and $b > -a$, suppose that independent random variables V_k are distributed with Beta$(1-a, b+ka)$. Let

$$\tilde{p}_k = V_k \prod_{k=1}^{k-1} (1 - V_j), \quad k = 1, 2, ..., \infty. \tag{2.12}$$

Now, let $\mathbf{p} = (p_1, p_2, ...)$, the ranked (sorted) values of $\tilde{p}_1, ... \tilde{p}_k$, be distributed as a Poisson-Dirichlet distribution (PDD) with parameters a, b, denoted by PDD(a,b). Note that $(p_1, p_2, ...)$ are sorted in decreasing order, however, the order does not matter when using the PDD in equation 2.11.

2.3.2 The Chinese restaurant process

The Chinese restaurant process, also known as the Blackwell-Macqueen urn scheme [57], provides a practical representation for incremental sampling from the posterior of the PDP.

Consider a Chinese restaurant with an infinite number of *tables*, each table has infinity capacity to sit customers around. Customers sitting at the same table k share the same dish $X_k^\star$. When the first customer X_1 arrives at the restaurant, the customer sits at the first empty table, then the $(N+1)^{th}$ subsequent customer X_{N+1} chooses to sit in an occupied t^{th} table with probability proportional to $\frac{n_k^\star - a}{N+b}$, or chooses to sit in a new empty table with probability proportional to $\frac{Ka+b}{N+b}$. Mathematically, the conditional posterior distribution of X_{N+1} given a finite sequence of samples $X_1, \ldots, X_N$ from $G \sim PDP(a,b,H)$ and G marginalised is:

$$p(X_{N+1} \mid x_1, \ldots, X_N, a, b, H) = \sum_{k=1}^{K} \frac{n_k^\star - a}{N+b} \delta_{X_k^\star(\cdot)} + \frac{Ka+b}{N+b} H(\cdot), \tag{2.13}$$

where K is the distinct values in $X_1, \ldots, X_N$ ordered as $X_1^\star, \ldots, X_K^\star$ with counts $n_1^\star, \ldots, n_K^\star$. Note that $X_1, \ldots, X_N$ are samples from a base distribution $H(\cdot)$ while $X_1^\star, \ldots, X_K^\star$ are the unique samples among $X_1, \ldots, X_N$.

When the base distribution is non-atomic (continuous), the probability of repeated draws is effectively zero [53]. Thus, tables serve distinct dishes, and no pair of tables serves the same dish. However, when the base distribution is discrete, the probability of the same dish being served by multiple tables is positive with probability one, i.e., a dish can be served on one or more tables. A latent variable $t_k^\star$ is introduced to count the number of tables serving the same dish $X_k^\star$. Then, the conditional posterior distribution of X_{N+1} with G marginalised is:

$$p(X_{N+1} \mid X_1, \ldots, X_N, t_1^\star, \ldots, t_K^\star, a, b, H(\cdot)) = \sum_{k=1}^{K} \frac{n_k^\star - a t_k^\star}{N+b} \delta_{X_k^\star}(\cdot) + \frac{Ta+b}{N+b} H(\cdot), \tag{2.14}$$

where $n_k^\star$ is the number of customers having dish $X_k^\star$ and $\sum_{k=1}^{K} n_k^\star = T$.

The number of tables $t_k^\star$ defines the *multiplicity* of dish $X_k^\star$. In other words, the multiplicity is the frequency of a distinct value drawn from the base measure appearing in the sampled data. Then, the joint posterior distribution of customers $(X_1, X_2, ..., X_N)$ and multiplicities $t_1^\star, t_2^\star, ..., t_k^\star$ is given by:

$$p(X_1, X_2, ..., X_N, t_1^\star, t_2^\star, ..., t_k^\star \mid a, b, H) = \frac{(b \mid a)_T}{(b)_N} \prod_{k=1}^{K} H(X_k^\star)^{t_k^\star} S^{n_k}_{t_m^\star, a}, \tag{2.15}$$

where $(x \mid y)_N$ denotes the Pochhammer symbol and $S^N_{M,a}$ is a generalised Stirling number.$(x)_N$ is Pochhammer symbol with $y = 1$.

The Pochhammer is defined by:

$$(x \mid y)_N = x(x+y)...(x+(N-1)y) = \begin{cases} x^N & \text{if y=0} \\ y^N \times \frac{\Gamma(x/y+N)}{\Gamma(x/y)} & \text{if y>0,} \\ \frac{\Gamma(x+N)}{\Gamma(x)} & \text{if y=1,} \end{cases} \tag{2.16}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

The Stirling number [58] is computed with linear recursion [54] as:

$$\begin{aligned} &S^{N+1}_{M,a} = S^{N}_{M-1,a} + (N - Ma) S^{N}_{M,a} \quad M \le N, \\ &S^{N+1}_{M,a} = 0, \quad M > N, \quad S^{N}_{0,a} = \delta_{N,0}. \end{aligned} \tag{2.17}$$

2.3.3 Properties

2.3.3.1 Moments

For any measurable set $B \in \mathbb{B}$, the mean, variance and covariance of $G \sim PDP(a, b, H)$ are given by:

$$\begin{aligned} \mathbb{E}(G(B)) &= H(B) \\ \mathbb{V}(G(B)) &= \frac{1-a}{1+b} H(B)(1 - H(B)) \\ \mathbb{C}ov(G(B), G(B')) &= -\frac{1-a}{1+b} H(B)(H(B')) \quad \forall B' \cap B = \emptyset. \end{aligned} \tag{2.18}$$

2.3.3.2 Power law

Note that the expectation of p_k from Definition 2.3.2 is of order $O(k^{-1/a})$ if $0 < a < 1$, which indicates the partition size decays according to a 'power law'. For the DP, the expectation of p_k is of order $O((\frac{b}{1+b})^k)$ which decreases exponentially in k.

In the CRP representation, the proportion of tables with N customers scales as $O(N^{-(a+1)})$, and the total number of tables scales as $O(N^a)$. Having the discount parameter causes the tail of a distribution drawn from the PDP to be much longer than that drawn from the DP, indicating a large number of tables with small numbers of customers.

2.4 Summary

In this chapter, we described the Dirichlet distribution, the Dirichlet process and the Poisson-Dirichlet process. We described the Chinese restaurant process which provides a useful sampling scheme for the Poisson-Dirichlet process. As we will show in Chapter 3, the Dirichlet distribution is the key distribution in latent Dirichlet allocation, the vanilla topic model; and the Poisson-Dirichlet process plays a fundamental role in the segmented topic model. The Dirichlet process aids the transition between the Dirichlet distribution and the Poisson-Dirichlet process.

Chapter 3

Topic Modelling for retail analytics

Topic modelling (TM) is a statistical framework that was originally introduced to organise, search, and understand large collections of text documents [25, 26, 27, 28]. Topic models facilitate the finding and discovering of text content in digital libraries through their ability to learn and apply subject tags to documents [59], that otherwise, impossible to do by the human eye [60]. For instance, the task of classifying thousands of books would require a large investment of time and human resources to build a categorisation scheme, to read every single book, and to determine the categories that represent each book. Feasible or not, accurate or not, reading, understanding and processing a large collection of books, or discrete data, may exceed human capacity.

3.1 A brief introduction

In general, topic modelling tackles the problem of processing, understanding, and summarising large collections of discrete data. Here, we briefly discuss some of the most well-known topic models.

Latent Dirichlet allocation (LDA) [61, 25] is the most popular topic model. LDA interprets documents as topical mixtures (distributions over topics) and topics are distributions over words. In this way, each topic represents different subjects that are characterised by co-occurrent words. Bringing topic modelling into the context of retail analytics, grocery transactions are combinations of products that can be summarised as discrete distributions over topics, and topics are discrete distributions over products that exhibit different probability rankings to represent different customer needs. LDA will be described in detail in Section 3.2.

Before LDA, two models were proposed for automated document indexing:

Latent semantic analysis (LSA) [28] is a linear algebra approach to computing a document-term matrix. LSA takes a vector space representation of documents based on term frequencies and applies a dimension reducing linear projection based on a Singular Value Decomposition (SVD). LSA lacks a statistical foundation.

Probabilistic LSA (pLSA) [27] also known as the aspect model, is the very first topic model [60]. pLSA is the probabilistic version of latent semantic analysis (LSA), introduced to decompose documents with a probabilistic approach. pLSA assumes that distributions over topics are document-specific discrete parameters; thereby, pLSA cannot be used to summarise new documents as topical mixtures.

LDA has been extended in several directions to overcome its many limitations. For instance, the hierarchical Dirichlet process (HDP) addresses the problem of determining the number of topics a priori, correlated topic model (CTM) and Packincho allocation model (PAM) model topic correlations.

Hierarchical Dirichlet process (HDP) [62] is the non-parametric version of LDA, in which the number of topics is unknown a priori and is to be inferred from the data. HDP assumes that documents are generated by sampling words from multinomial probability vectors, which are Dirichlet Process distributed with a base measure that is also Dirichlet Process distributed.

Correlated topic model (CTM) [63] assumes that subsets of the underlying latent topics are highly correlated, i.e., if one topic is observed then another correlated topic would have a higher probability of being seen in the same document. CTM directly models correlation between topics by using the logistic normal distribution, which incorporates a covariance structure among the components. The non-conjugacy between logistic normal distributions and multinomial distributions hinders the derivation of an efficient inference algorithm such as Gibbs sampler. Thus, the inference is carried out through a variational inference procedure.

Packincho allocation model (PAM) [64] is a mixture model that uses a directed acyclic graph (DAG) structure to capture arbitrary topic correlations. In PAM, a hierarchical model is presented by leaf nodes and interior nodes. Each leaf node is associated with a word in the vocabulary, and each non-leaf interior node corresponds to a topic, having a distribution over its node children. Some interior nodes may also be children of other interior nodes, thus representing a mixture of topics. PAM, therefore, captures not only correlations among words (as in LDA), but also correlations among topics themselves. The non-parametric Bayes Pachinko allocation [65], the non-parametric version of PAM, learns the number of topics as long as topical correlations.

LDA disregards any structure within documents, i.e., a book is composed by chapters or a novel is composed by paragraphs; thus dismissing the fact that chapters or paragraphs from the same document may discuss different topics. Segmented topic model (STM) and Sequential latent Dirichlet allocation (seqLDA) are models that extend LDA to accommodate document structure. In the context of retail analytics, document structure can be interpreted as store hierarchy over transactions.

Segmented topic model (STM) [44] represents documents as collections of segments (sentences, paragraphs, sections, etc.). STM is a hierarchical model where documents and segments are described as topical mixtures. In STM, the random variables associated with transactions are Poisson-Dirichlet process (PDP) distributed, which through the Chinese restaurant process (CRP) derives an efficient Gibbs sampler algorithm. For structures with more than one layer, a variational algorithm is exposed in [66]. STM and CRP will be described in detail in Section 3.3.

Sequential latent Dirichlet allocation (seqLDA) [67] extends LDA to explicitly model the sequential topic structure of a document, that is how a sub-idea in a segment is closely related to its antecedent and subsequent segments. The progressive topical dependency is captured using a hierarchical two-parameter Poisson–Dirichlet process (HPDP). HPDP is defined on a singly connected network of probability vectors (topical mixtures), where each probability vector is PDP distributed with the antecedent probability vector as base distribution.

LDA assumes that documents are exchangeable, disregarding the temporal sequence between them. Temporal patterns can still be identified by post-processing transaction-specific topical mixtures or by applying topic models such as (continuous) dynamic topic model, dynamic mixture model and topics over time.

As discussed in [26], topical mixtures inferred by a time-unaware LDA are ordered and aggregated by year and average topic probabilities are used to construct linear trends, providing quantitative measures of time prevalence. STM could be also applied over time-aware aggregations of documents. Another post-processing approach involves running LDA models on each time slice; in this manner, the LDA model may capture time-specific topics that otherwise would be overlooked by the time-unaware LDA model. This approach requires aligning topics across periods.

Dynamic topic model (DTM) [68] extended LDA to let topics and topic distributions vary depending on time-variant Gaussian Priors in Markovian style, i.e. the Gaussian Prior at time t depends on the Gaussian Prior at time $t-1$. In detail, DTM chains the natural parameters of multinomial distributions in a state-space model that evolves with Gaussian noise. Given the non-conjugacy between Gaussian and Multinomial distributions, a variational Kalman filtering method is then introduced to facilitate DTM inference. DTM requires time to be discretised (grouped by years, months, days, etc.) which may affect the memory requirements and computational complexity of posterior inference [69].

Continuous dynamic topic model (cDTM)) [69] replaces the state-space model of DTM with its continuous generalisation, the Brownian motion model; so that the only discretisation is the resolution at which the timestamps are measured. The inference of cDTM is addressed with a sparse variational inference method that takes advantage of the data sparsity; i.e., topics at time t are only determined by the words observed at time t, easing memory requirements and computational complexity.

Dynamic Mixture Model (DMM) [46] accounts for sequential data assuming static topics and time-changing topical mixtures. DMM was introduced to analyse data streams and to identify topics of time series. DMM interprets a snapshot as a document and each stream as a word occurrence over time. Thus, document-specific topical mixtures show dependency in a Markovian fashion where the topical mixture at snapshot depends on the topical mixture at the previous snapshot. DMM requires discretised time units.

Topics over time (TOT) [70] is a topic model that assumes static topics, but it does not discretise time nor follow a first-order Markov assumption. Instead, TOT captures topics that are associated with continuous distributions over time. TOT assumes that timestamps are random variables that are sampled from a Beta distribution. TOT is not appropriate for modelling data where topics may be bursty and multi-modal [46].

LDA and all the aforementioned topic models are unsupervised learning machine methods, which aim to identify the hidden semantic structure but do not aim to predict a response. A topic model with a supervised approach is the supervised latent Dirichlet allocation and labelled LDA.

Supervised latent Dirichlet allocation (sLDA) [71] extends LDA to model a response variable associated with each document. sLDA jointly models documents and responses aiming to find latent topics that would best predict the response variables for future unlabelled documents. sLDA uses a maximum-likelihood procedure for parameter estimation, which relies on variational approximations to handle intractable posterior expectations.

Labeled LDA (L-LDA) [72] is a supervised topic model suitable for multiply labelled corpora. L-LDA not only extends LDA by incorporating supervision but also extends Multinomial Naive Bayes by incorporating a mixture model. L-LDA associates each label with one topic in direct correspondence.

LDA also disregards document metadata such as authorship. Author-topic model address this challenge to identifying topics and their associations with authors.

Author-topic model (ATM) [73] extends LDA to include authorship information. ATM assumes that documents are written by several authors, who write about themes that can be summarised as mixtures over topics. In ATM, authors but not documents are described as topical mixtures.

Beyond text mining applications, topic models have been widely adopted in other fields such as image classification [74, 75, 76], audio classification [77], recommender systems [78], social network analysis [79], linguistics [80], biology [81], political science [82], history [80, 83], software engineering [84, 85], social media [86, 87], emotion classification [88], medicine and biology [89, 90, 91] to name a few.

In this thesis, LDA and STM are used to identify topics that represent customer behaviours. We describe both topic models in more detail in the next sections.

3.2 Latent Dirichlet allocation

We interpret LDA [25, 26] in terms of retail data, where transactions (bag of products) can be summarised as topical mixtures and topics are discrete distributions over a fixed assortment of products.

3.2.1 Generative process

LDA assumes a generative process in which transactions are created by a two-step sampling process. First, the generative process assumes that each transaction d has a finite number of products N_d, and for each product w_n there is a topic assignment z_n which is sampled from the transaction-specific topical mixture θ_d. Second, a product is sampled from the topic distribution associated to the topic assignment ϕ_{z_n}. Transaction-specific topical mixtures $\Theta = [\theta_1,...,\theta_D]$ and topic distributions $\Phi = [\phi_1,...,\phi_K]$ are also sampled once from a Dirichlet distribution with hyperparameters $\boldsymbol{\alpha} = [\alpha_1,...,\alpha_K]$ and $\boldsymbol{\beta} = [\beta_1,...,\beta_V]$, respectively. K is the number of topics that is assumed to be known a priori and V is the size of the product assortment. Mathematically,

$$\begin{aligned}\phi_k &\sim \text{Dirichlet}(\boldsymbol{\beta})\\ \theta_d &\sim \text{Dirichlet}(\boldsymbol{\alpha})\\ z_{n,d} &\sim \text{Multinomial}(\theta_d)\\ w_{n,d} &\sim \text{Multinomial}(\phi_{z_{n,d}}).\end{aligned} \tag{3.1}$$

As shown in Figure 3.1, the generative process of LDA describes a collection of random variables where only the products are observed. Note that topics are shared among all transactions. LDA assumes that products are exchangeable, thereby, any order is disregarded. This assumption is known in text mining as *bag-of-words*. In text mining applications, the word order may be relevant to fairly represent the data. In the retail context, this assumption suits the nature of transactional data well as customers do not tend to check out products with an order that expresses their shopping motivations. Transactions are assumed to be exchangeable with each other, disregarding any potential temporal sequence among transactions.

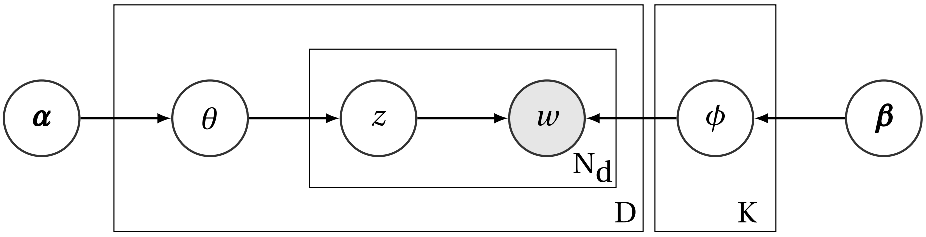


Figure 3.1: LDA graphical model. Nodes denote random variables and edges denote dependencies. Unshaded node denote hidden random variables and shaded nodes denote observed random variables. Plates denote replication.

LDA joint distribution is given by:

$$P(\Phi,\Theta,\mathbf{z},\mathbf{w} \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = p(\mathbf{w} \mid \Phi,\mathbf{z})\, p(\Phi \mid \boldsymbol{\beta})\, p(\mathbf{z} \mid \theta)\, p(\theta \mid \boldsymbol{\alpha}), \tag{3.2}$$

and posterior distribution is given by:

$$P(\Phi, \Theta, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\Phi, \Theta, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}, \tag{3.3}$$

where $\mathbf{z}$ and $\mathbf{w}$ are vectors of topic assignments and observable words, respectively.

Marginal probability, and, consequently, posterior distribution, cannot be computed tractably. There are various approaches for estimating the posterior distribution, such as gradient descent [92], Gibbs sampling [26], variational inference [25], and expectation propagation [93]. We use the collapsed Gibbs sampling algorithm [26] to sample from the posterior distribution and learn topic distributions since this method has shown advantages in computational implementation, memory, and speed.

3.2.2 The collapsed Gibbs sampler

The Gibbs sampling algorithm follows a collapsing strategy in which the topic distributions Φ and topical mixtures Θ are integrated out, reducing the inference problem to $P(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\mathbf{w} \mid \mathbf{z}, \boldsymbol{\beta}) p(\mathbf{z} \mid \boldsymbol{\alpha})$. This is:

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{z}, \boldsymbol{\beta}) &= \left(\frac{\Gamma(\beta_0)}{\prod_v \Gamma(\beta_v)} \right)^K \prod_{k=1}^{K} \frac{\prod_v \Gamma(\beta_v + n_{k,v})}{\Gamma(\beta_0 + n_k)}, \\ p(\mathbf{z} \mid \boldsymbol{\alpha_0}) &= \left(\frac{\Gamma(\alpha_0)}{\prod_k \Gamma(\alpha_k)} \right)^D \prod_{d=1}^{D} \frac{\prod_k \Gamma(\alpha_k + n_{d,k})}{\Gamma(\alpha_0 + n_d)}, \end{aligned} \tag{3.4}$$

where $\alpha_0 = \sum_k \alpha_k$, $\beta_0 = \sum_v \beta_v$,$n_{k,v}$ is the number of times product v is allocated to topics k, $n_{d,k}$ is the number of times topic k is allocated to document d, n_k is the number of times topic k has been allocated, and n_d is the number of products in transaction d.

A Markov chain is constructed, in which the stationary distribution is the posterior distribution of interest. The Gibbs sampling algorithm initialises the chain with random topic assignments. The next state is reached by sequentially sampling all variables from their distribution conditioned on the current values of all other variables and the data. Mathematically:

$$p(z_i = k \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{N_{k,v}^{-i} + \beta_v}{N_k^{-i} + \beta_0} \frac{N_{d,k}^{-i} + \alpha_k}{N_d^{-i} + \alpha_0}, \tag{3.5}$$

where the notation N^{-i} is a count that does not include the current assignment of z_i. This full conditional distribution can be interpreted as the product of two terms: the probability of the product v under topic k and the probability of topic k under the current topic distribution for transaction d. Consequently, the probability of assigning a topic to any particular product in a transaction will be increased once many products of the same type have been assigned to the topic and the topic has been assigned several times to the transaction. Dirichlet hyperparameters smooth the probability of sampling a topic assignment, so the topic is still probable even when the topic count is zero.

After a burn-in period, states of the Markov chain (topic assignments) are recorded within an appropriate lag to ensure low autocorrelation between samples. For a single sample s, Φ and Θ are estimated from the counts of topic assignments and Dirichlet parameters by:

$$\hat{\phi}_{k,v}^s = E(\phi_{k,v}^s \mid \mathbf{w}, \mathbf{z}) = \frac{N_{k,v}^s + \beta_v^s}{N_k^s + \beta_0^s}, \quad k = 1 \ldots K, v = 1 \ldots V, \tag{3.6}$$

$$\hat{\theta}_{d,k}^s = E(\theta_{d,k}^s \mid \mathbf{w}, \mathbf{z}) = \frac{N_{d,k}^s + \alpha_k^s}{N_d^s + \alpha_0^s}, \quad d = 1 \ldots D, k = 1 \ldots K. \tag{3.7}$$

Due to the lack of identifiability, i.e., two topics from different samples that share the same index might not correspond to the same word distribution, LDA samples can only be integrated out for the calculation of statistics that are independent of the content of the topics [26].

In applications to retail data where the discrete space is high-dimensional, the Gibbs sampler needs to allocate a huge number of combinations of products into topics. The convergence of the Gibbs sampler is thus a computational challenge despite its closed-form solution.

3.2.3 Dirichlet priors

As indicated by [94], a Dirichlet asymmetric prior over topic distributions and a Dirichlet symmetric prior over topics improve topic coherence and model generalisation. Optimising Dirichlet parameters provides a good approximation of a fully Bayesian model without increasing computational costs. The optimal hyperparameters $\boldsymbol{\alpha}^\star$ are those which maximise the evidence or probability of the data given prior hyperparameters $P(D \mid \boldsymbol{\alpha})$.

$$P(D \mid \boldsymbol{\alpha}) = \prod_{d=1}^{D} \frac{\Gamma(\alpha)}{\Gamma(N_d + \alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_{d,k} + \alpha_k)}{\Gamma(\alpha_k)}. \tag{3.8}$$

[50] proposed the estimation of $\boldsymbol{\alpha}$ using an optimisation step through a fixed-point iteration method. This method derives the logarithm of the evidence and estimates the optimal values of asymmetric Dirichlet hyperparameters over topic proportions by:

$$\alpha_k^\star = \alpha_k \frac{\sum_{d=1}^{D} \Psi(N_{d,k} + \alpha_k) - \Psi(\alpha_k)}{\sum_{d=1}^{D} \Psi(N_d + \alpha_0) - \Psi(\alpha_0)}, \tag{3.9}$$

where $\alpha_0 = \sum_k^K \alpha_k$ and Ψ is the digamma function. Updates are repeated for a number of iterations or until $\alpha^\star$ converges to the values that maximise $P(D \mid \boldsymbol{\alpha})$. Note that LDA inference alternates between cycles of sampling topic assignments and of hyperparameter optimization. Thus, token counts, i.e., $N_{d,k}$, are obtained from the previous cycle of topic assignments.

Similarly, the optimal symmetric Dirichlet hyperparameters over topic distributions are given by:

$$\beta_0^\star = \beta_0 \frac{\sum_{v=1}^{V} \sum_{k=1}^{K} \Psi(N_{k,v} + \beta_v) - \Psi(\beta_v)}{V \sum_{k=1}^{K} \Psi(N_k + \beta_0) - \Psi(\beta_0)}, \tag{3.10}$$

where $\beta_0 = \sum_v^V \beta_v$ and $\beta_v = \frac{\beta_0}{V}$, Ψ is the digamma function. Again, updating is repeated until convergence of $\beta_0^\star$.

Later, [95] shows that the fixed-point iteration method could be sped up by recording topic frequencies and using the digamma recurrence relation:

$$\alpha_k^\star = \alpha_k \frac{\sum_{n=1}^{\max_d N_{d,k}} C_k(n)[\Psi(n+\alpha_k) - \Psi(\alpha_k)]}{\sum_{n=1}^{\max_d N_d} C_\cdot(n)[\Psi(n+\alpha_0) - \Psi(\alpha_0)]}, \tag{3.11}$$

where $C_k(n)$ is the number of transactions in which topic k has been seen exactly n times and Ψ is the digamma function. $\max_d N_d$ is the maximum transaction size. $\max_d N_{d,k}$ is the maximum number of times topic k has been seen across all transactions.

3.3 Segmented topic model

The segmented topic model (STM) [44] extended LDA to harness document structure so that documents are sets of paragraphs (segments) and each paragraph is a bag-of-words. The intuition behind this model is that paragraphs individually exhibit different themes, but as a collection, they follow the general theme exposed in the document. In the context of grocery retail data, a store can be interpreted as a set of transactions occurring in a specific location. While transactions may individually exhibit different customer needs, as a collection they exposed a store-specific topical mixture. Thereby, stores may summarise and describe specific and shared customer behaviours. Topics are distributions over a fixed assortment of products, which is composed of shared and specific products across store-specific assortments.

3.3.1 Generative process

We interpret STM in retail vocabulary. STM follows a two-step generative process. First, for each product w_n in a transaction d (which has a finite number of products N_d), a topic assignment z_n is sampled from the transaction-specific topical mixture v_d. Second, a product is sampled from the topic distribution associated to the topic assignment ϕ_{z_n}. So far, the generative process of STM follows the same strategy as LDA generative process. However, in STM transaction-specific topical mixtures v are not drawn from a Dirichlet distribution, but from a two-parameter Poisson-Dirichlet Process parametrised by a discount parameter a, a strength parameter b and a store-specific topic mixture. Store-specific topical mixtures Θ are drawn from a Dirichlet distribution with hyperparameter $\boldsymbol{\alpha}$. Topic distributions Φ are drawn from a Dirichlet distribution with hyperparameter $\boldsymbol{\beta}$. Mathematically,

$$\begin{aligned} \phi_k &\sim \text{Dirichlet}(\boldsymbol{\beta}) \\ \theta_d &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \nu_{p,d} &\sim \text{PDP}(a, b, \theta_d) \\ z_{n,p,d} &\sim \text{Multinomial}(\nu_{p,d}) \\ w_{n,p,d} &\sim \text{Multinomial}(\phi_{z_{n,p,d}}). \end{aligned} \tag{3.12}$$

As shown in Figure 3.2, the generative process of STM describes a collection of random variables in which only the products are observed. Note that transaction-specific topical mixtures derive from their corresponding store-specific topical mixture. Thus, transactions and stores share the space of latent topics. STM follows the LDA assumptions: transactions are 'bags-of-products', so product order is disregarded. Transactions are also assumed to be exchangeable from each other, disregarding any potential temporal sequence among transactions or any relationship between stores.

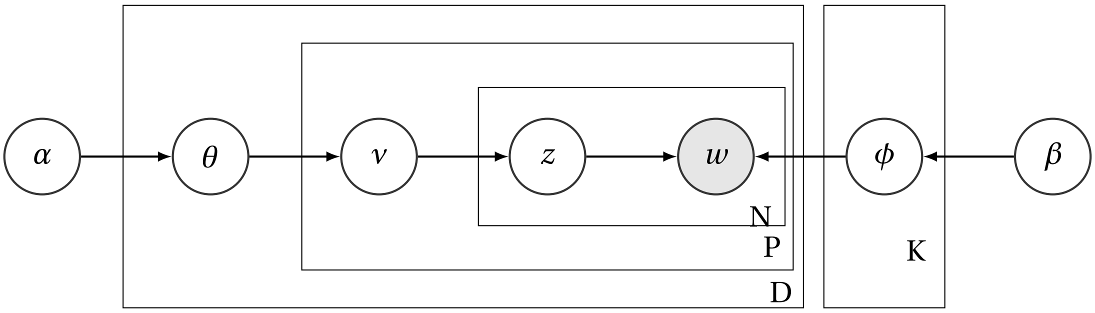


Figure 3.2: STM graphical model. Nodes denote random variables and edges denote dependencies. Unshaded node denote hidden random variables and shaded nodes denote observed random variables. Plates denote replication. The hidden variables are z topic assignments, θ store-specific topical mixtures, ν transaction-specific topical mixtures, ϕ topic distributions, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ Dirichlet hyperparameters. K number of topics, D number of stores, P number of transactions, and N number of products.

PDP is the key distribution in the STM. PDP is used to handle conjugacy between Dirichlet and Multinomial distributions. STM joint distribution is given by:

$$P(\Phi, \Theta, \nu, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\mathbf{w} \mid \Phi, \mathbf{z}) p(\Phi \mid \boldsymbol{\beta}) p(\mathbf{z} \mid \nu) p(\nu \mid \theta) p(\theta \mid \boldsymbol{\alpha}), \tag{3.13}$$

and posterior distribution is given by:

$$P(\Phi, \Theta, \nu, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\Phi, \Theta, \nu, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}, \tag{3.14}$$

where **z** and **w** are vectors of topic assignments and observable words, respectively.

Like LDA, STM's marginal probability, and consequently the posterior distribution, cannot be tractably computed. [44, 45] proposed a collapsed Gibbs sampling method in which latent statistics $\boldsymbol{t}$, that symbolise “table counts" in the Chinese restaurant process (CRP) [96], are introduced to marginalise out the hidden variables ν and to leave the hidden variables θ in conjugate form. This leads to the joint conditional distribution of topic assignments, transactions and table counts:

$$\begin{aligned} & p(\mathbf{z}, \mathbf{w}, \mathbf{t} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) = \\ & \prod_{d} \frac{\mathrm{Beta}_K(\boldsymbol{\alpha} + \sum_p \mathbf{t}_{p,d})}{\mathrm{Beta}_K(\boldsymbol{\alpha})} \prod_{p,d} \frac{(b \mid a)_{\sum_k t_{p,d,k}}}{(b)_{N_{p,d}}} \prod_{p,d,k} S^{N_{k|p,d}}_{t_{p,d,k},a} \prod_{k} \frac{\mathrm{Beta}_V(\boldsymbol{\beta} + \mathbf{N}_k)}{\mathrm{Beta}_V(\boldsymbol{\beta})} \end{aligned} \tag{3.15}$$

where $t_{p,d,k}$ is the table count for store d, transaction p and topic k. $\mathrm{Beta}_K(\boldsymbol{\alpha})$ is K dimensional beta function that normalises the Dirichlet distribution defined in Equation 2.2; $\mathbf{t}_{p,d}$ is a table count vector (i.e. $t_{p,d,1}, ..., t_{p,d,K}$); $(x \mid y)_N$ denotes the Pochhammer symbol defined in Equation 2.16; $N_{p,d}$ size of transaction p in store d; $S^N_{M,a}$ is a generalised Stirling number defined in Equation 2.17. $N_{k|p,d}$ number of topic assignments of topic k in transaction p in store d. $\mathrm{Beta}_V(\boldsymbol{\beta})$ is V dimensional beta function that normalises the Dirichlet distribution; $\mathbf{N}_k$ is a vector of $N_{v|k}$, which is the number of products of type v assigned to topic k.

3.3.2 The block Gibbs sampler

[45] proposes a block Gibbs sampling algorithm that jointly samples topic assignments and table indicators, leading to a more efficient sampling method. Table counts are not sampled, instead reconstructed by summation of the table indicators.

$$t_k = \sum_{n=1}^{N} u_n 1_{z_n=k}, \tag{3.16}$$

Using the table indicator representation, the PDP posterior distribution is:

$$p(z,t \mid a,b,\theta) = \prod_k \frac{n_k!}{t!(n-t)!} p(z,u \mid a,b,\theta), \tag{3.17}$$

responding to $\frac{n_k!}{t!(n-t)!}$ sitting arrangements.

The joint distribution of topic assignments and table indicators can be obtained by placing Equation 3.17 in Equation 3.15 resulting in:

$$p(\mathbf{z},\mathbf{w},\mathbf{t} \mid \boldsymbol{\alpha},\boldsymbol{\beta},a,b) = \prod_d \frac{\text{Beta}_K(\boldsymbol{\alpha} + \sum_p \mathbf{t}_{p,d})}{\text{Beta}_K(\boldsymbol{\alpha})} \prod_{p,d} \frac{(b \mid a)_{\sum_k t_{p,d,k}}}{(b)_{N_{p,d}}} \prod_{p,d,k} S^{N_{k|p,d}}_{t_{p,d,k},a} \frac{t_{p,d,k}!(N_{p,d,k} - t_{p,d,k})!}{n_{p,d,k}!} \prod_k \frac{\text{Beta}_V(\boldsymbol{\beta} + \mathbf{N}_k)}{\text{Beta}_V(\boldsymbol{\beta})} \tag{3.18}$$

The block Gibbs sampling algorithm goes as:

1. Sample a table indicator $u_{z_n} = 1$ or $u_{z_n} = 0$ with probabilities:

   $$p(u_{z_n} = 1 \mid z_n = k) = \frac{t_k}{n_k} \quad p(u_{z_n} = 0 \mid z_n = k) = 1 - \frac{t_k}{n_k}, \tag{3.19}$$

   and discounts the current assignment z_n from $N_{p,d,k}$ and reduces $t_{p,d,k}$ by 1 if $u_{z_n} = 1$.

2. Compute the full conditional distribution taking into account two scenarios: the probability of opening a new table (Equation 3.20) and the probability of choosing an occupied table (Equation 3.21) if $t'_{p,d,k} > 0$.

$$p(z_n = k, u_n = 1 \mid \mathbf{z} - \{z_n\}, \mathbf{u} - \{u_n\}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) \propto$$
$$\frac{\alpha_k + t'_{d,k}}{\alpha + t'_d} \frac{b + a t'_{p,d}}{b + N'_{p,d}} \frac{S^{N'_{p,d,k}+1}_{t'_{p,d,k}+1}}{S^{N'_{p,d,k}}_{t'_{p,d,k}}} \frac{t'_{p,d,k} + 1}{n'_{p,d,k} + 1} \frac{\beta_v + N'_{k,w_{p,d,n}}}{\beta + N'_k}, \tag{3.20}$$

$$p(z_n = k, u_n = 0 \mid \mathbf{z} - \{z_n\}, \mathbf{u} - \{u_n\}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) \qquad \propto$$
$$\frac{S^{N'_{p,d,k}+1}_{t'_{p,d,k}}}{S^{N'_{p,d,k}}_{t'_{p,d,k}}} \frac{1}{b + N'_{p,d}} \frac{n'_{p,d,k} - t'_{p,d,k} + 1}{n'_{p,d,k} + 1} \frac{\beta_v + N'_{k,w_{p,d,n}}}{\beta + N'_k}, \tag{3.21}$$

where the dash indicates statistics after excluding the current assignment.

3. Update the counts of $n_{p,d,k}$ and $t_{p,d,k}$ with the sampled topic assignment z_n and table indicator u_n.

Topics ϕ, store-specific topical mixtures θ and the transaction-specific topical mixtures v are not explicitly sampled, but can, instead, be inferred per-iteration through their conditional posterior mean estimates:

$$\widehat{\theta}^s_{1,k} = E(\theta^s_{1,k} \mid \mathbf{t}^s, \boldsymbol{\alpha}) = \frac{\alpha_k + \sum_p t^s_{p,1,k}}{\alpha_0 + \sum_{p,k} t^s_{p,1,k}}, \tag{3.22}$$

$$\widehat{v}^s_{p,d,k} = E(v^s_{p,d,k} \mid \mathbf{z}^s, \mathbf{t}^s, a, b) = \frac{N^s_{p,d,k} - a \times t^s_{p,d,k}}{b + N^s_{p,d}} + \theta_{d,k} \frac{\sum_k t^s_{p,d,k} \times a + b}{b + N^s_{p,d}}, \tag{3.23}$$

$$\widehat{\phi}^s_{k,v} = E(\phi^s_{k,v} \mid \mathbf{z}^s, \boldsymbol{\beta}) = \frac{\beta_v + N^s_{k,v}}{\beta_0 + N^s_k}, \tag{3.24}$$

where $\alpha_0 = \sum_k^K \alpha_k$ and where $\beta_0 = \sum_v^V \beta_v$. In [44], the discount parameter a is set by the user and the strength parameter b is sampled from $\text{Gamma}\left(\sum_{p,d,k} t_{p,d,k}, \sum_{p,d} \log 1/q_{p,d}\right)$, where $q_{p,d}$ is an auxiliary variable sampled from $\text{Beta}\left(b, N_{p,d}\right)$ with b as initial guess. However, we set the discount and concentration parameters manually to simplify the inference process and speed up processing time.

3.4 Topic models in retail analytics

Topic modelling, in particular LDA, has already been used to identify latent shopping motivations in retail data. For example, [11] applied LDA to grocery transactions from a major European supermarket to identify latent topics of product categories, intending to support an item recommendation system. [30] sketched the core of a recommender system to illustrate the managerial relevance of estimated topics, which were obtained from training LDA and the CTM on market baskets from a medium-sized German supermarket. [29] applied topic models to market baskets from a medium-sized online retailer in the Netherlands to identify latent motivations and to predict product purchasing in large assortments. [31, 97] compared LDA and CTM and other unsupervised probabilistic machine learning methods on point-of-sale transactions from a typical local grocery store in Austria, analysing 169 product categories.

The aforementioned works analysed market baskets as collections of product categories instead of the full product resolution, disregarding distinction between sizes, fragrances, flavours, package, etc. [11] analysed 473 synthetic categories; [30] analysed 60 product categories with the highest univariate purchase frequencies, and [29] analysed 394 categories. Analysing transactions using product categories reduces the dimensionality of the product assortment; thereby, decreasing computational time. However, topic analysis without visibility on individual products may dismiss product combinations with practical implications. [32] provided a direct application of a 25-topic LDA model on transactional data from a major British retailer to identify shopping goals.

There are a few topic models that were designed to analyse grocery retail data. For instance, [33] introduced ‘SHOPPER’, a sequential probabilistic model, that captures interaction among items and answers counterfactual queries about changes in prices. [34] combined the correlated topic model with the vector autoregression to account for product, customer and time dimensions existing in purchase history data. [98] introduced ‘Product2Vec’, a method based on the representation learning algorithm Word2Vec, to study product-level competition, when the number of products is large and produce more accurate demand forecasts and price elasticities estimations.

3.5 Summary

In this chapter, we have discussed two types of topic models: latent Dirichlet allocation and segmented topic model. We interpreted the generative of these models in terms of grocery retail data and described their MCMC inference methods. As we will show in subsequent chapters, LDA is used to identify customer behaviours among grocery retail transactions, and STM is used to identify customer behaviours that are relevant regarding store or time-aware aggregations of transactions. Applications of LDA and STM are presented in Chapters 4 and 5, respectively. In Chapter 6, we define SeqSTM, an extension of STM, and compare SeqSTM against STM and LDA.

Chapter 4

Summary and Evaluation of Topic Models

This Chapter is largely based on a paper due to be published in JRSSC titled "Modelling Grocery Retail Topic Distributions: Evaluation, Interpretability and Stability". arXiv:2005.10125

In this chapter, we apply LDA to identify customer behaviours of grocery retail data. We propose a clustering algorithm to summarise the posterior distribution of LDA and of the topics models we will use in subsequent sections. This clustering methodology provides measures of topic uncertainty, which allows users to select topics of high recurrence. We evaluate posterior summaries and posterior samples of LDA using a holistic evaluation framework that includes model generalisation and topic quality aspects such as topic coherence, topic distinctiveness and topic credibility. We also provide thresholds for interpreting topic quality aspects in the domain of our application.

4.1 Introduction

Summarising the posterior distribution of a topic model is not a trivial task. Topic models are, in essence, mixture models and thus subject to the label switching problem [99, 100, 101]. In other words, permutations of inferred topic distributions do not change the likelihood of the topic model. Because of this non-identifiability, there is no guaranteed correspondence between individual topics across samples; thereby, topic distributions cannot be averaged across samples for any analysis that relies on the content of specific topics [26].

Various methodologies aim to alleviate the label switching problem, assuming that topics are present (but with switched labels) across samples. [99, 100, 101] developed matching algorithms for assigning component labels per iteration such that an overall loss function is minimised. Other relabelling strategies considered label invariant loss functions [102, 103], identifiability constraints [104], and probabilistic relabelling [105, 106]. However, component relabelling techniques should not assume one-to-one matches, since matched topics may show large distributional dissimilarity among posterior samples.

In response, we propose a post-processing methodology that aggregates topic distributions obtained from multiple samples, which are obtained from running the topic model several times. This methodology groups topic distributions into an unconstrained number of clusters using a dissimilarity measure. Through hierarchical clustering, topics are grouped using the average link and cosine distance, which among other distributional measures correlates with human judgement on topic similarity [38]. A *clustered topic* is defined as the average topic distribution that exhibits the same theme, and its posterior uncertainty is given by its topic *recurrence*, i.e., the number of topics within the same cluster (the number of posterior samples exhibiting the same topic). Hierarchical clustering has been used previously to interactively align topics [37] and to aggregate topic models with small and large numbers of topics [107]. In comparison to these works, we aim to identify topics that illustrate different customer behaviours while measuring their uncertainty.

Depending on the domain of interest, users can set thresholds of minimum recurrence to select clustered topics of low uncertainty. Topics of low uncertainty appear consistently across posterior samples, reflecting ‘reliable’ topics. The selected clustered topics are then used to lead interpretations on customers’ needs instead of using a single posterior draw or a posterior mean (thus circumventing the label-switching problem).

The second contribution of this chapter is the definition of an evaluation framework for topic models. Typically, the evaluation of topic models is based on model fit metrics such as held-out-likelihood or perplexity [41, 42], which assess the model's generalisation capability by computing the model likelihood on unseen data. However, human evaluation may not agree with held-out metrics which lead to topic models with less semantically meaningful topics[39]. Inferred topics may not correspond to genuine and meaningful themes [40], affecting the user's confidence in the application of the topic model [108]. Thus, evaluation of topic models should include qualitative aspects such as topic *coherence* along with model generalisation, especially for applications with descriptive purposes.

Topic coherence [43] measures the interpretability of individual topics. It is typically quantified by co-occurrence metrics such as Pointwise Mutual Information (PMI) and Normalized Pointwise Mutual Information (NPMI) [109], which have been shown to correlate with human annotators in [43, 110]. Various methods have been proposed to improve topic coherence. For example, [94] used asymmetric priors over document distributions to capture highly frequent terms in few topics; [111] proposed the applications of regularisation methods; [108] applied the generalised the Pòlya urn model aiming to reduce the number of low-quality topics. In our application to retail data, we will show that the average NPMI of selected clustered topics (disregarding topics of high uncertainty) is larger than the average NPMI of single LDA posterior samples.

As observed by [59], inferred topics within one posterior sample may contain product combinations with so little variation that could be associated with the same semantic concept leading to a suboptimal outcome. In addition, topics may also exhibit significant variations across posterior samples [35, 36, 37], i.e., a topic associated with a particular semantic concept (in our case, a customer behaviour) may appear and disappear depending on its posterior uncertainty. Thus, we define two additional qualitative aspects: topic distinctiveness and topic credibility.

Topic distinctiveness measures the semantic dissimilarity among topics of the same posterior sample and topic credibility quantifies distributional similarity among posterior samples. Within topics of a single posterior sample, topic distinctiveness is defined as the minimum of the cosine distances between a topic and all the other topics. Across posterior samples, topic credibility is defined as the average maximum cosine similarity; where the maximum cosine similarity is w.r.t the topics of a different posterior sample. These two measures are based on the cosine distance, since it correlates with human judgement on topic similarity [38]. Thus, high-quality topics are not only coherent but also distinctive among them and identifiable in other posterior samples.

In summary, we establish a more holistic definition for model evaluation, which assesses topic models based on held-out likelihood and qualitative aspects such as topic coherence, distinctiveness and credibility.

In this chapter, we show an application of LDA to identify customer behaviours in a large collection of transactions from a major retailer in the UK. To guide interpretations of the qualitative metrics, we carried out a user study in which experts in grocery retail analytics assessed topics for their interpretability and similarity. We evaluate LDA, varying the number of topics, and show that not all the LDA topics are the most coherent, distinctive, and credible, concurring with [40, 108, 59, 35, 36, 37]. Moreover, we demonstrate that the selection of recurrent topics through the clustering methodology provides subsets of clustered topics with better model likelihood, greater credibility and improved interpretability.

Through the application of LDA and the proposed clustering methodology to retail data, we identify credible and coherent topics that exhibit a variety of shopping motivations such as diet orientations, ingredients for specific dishes, foods for specific events, pet ownership, household composition, festivities, preference for budget/premium products, seasonal demand, to name a few. Practical implications derive from the analysis of the identified topics, which also offer new means for social, cultural and dietary research.

4.2 Topic model evaluation

Topic model evaluation is typically based on model fit metrics such as held-out-likelihood and perplexity [41, 42], which assess the generalisation capability of the model by computing the model likelihood on unseen data. However, the LDA likelihood may lead to topic models with less semantically meaningful topics according to human annotators [39]. The evaluation of topic models should not, therefore, be exclusively based on likelihood metrics, but should also include topic quality metrics such as topic coherence [43], topic distinctiveness, and topic credibility.

In this section, we summarise metrics of model generalisation, topic coherence, and introduce metrics for topic distinctiveness and topic credibility. These four metrics will be used to evaluate topic models throughout this thesis.

4.2.1 Model generalisation

Model fit metrics such as perplexity or held-out-likelihood of unseen documents (transactions) estimate a model's capability for generalisation or predictive power. Perplexity is a measurement of how well the probability model predicts a sample of unseen (or seen) data. A lower perplexity indicates the topic model is better at predicting the sample. Mathematically,

$$\text{Perplexity} = -\frac{\log P(\mathbf{w}'|\Phi, \boldsymbol{\alpha})}{N'}, \tag{4.1}$$

where $\mathbf{w}'$ is a set of unseen products in a document, N' is the number of products in $\mathbf{w}'$, $\Phi = [\phi_1, \phi_2, \ldots, \phi_K]$ is a posterior estimate or draw of topics and $\boldsymbol{\alpha}$ is the posterior estimate or draw of the Dirichlet hyperparameters.

Computing the log-likelihood of a topic model on unseen data is an intractable task. Several estimation methods are described in [41, 42]. We use the left-to-right algorithm with 30 particles to approximate the log-likelihood on held-out documents [41, 95]. The left-to-right algorithm breaks the problem of approximating the log-likelihood of one document (transaction) in a series of parts, where each part is associated with the probability of observing one term (product) given the previously observed terms. The likelihood of each term is approximated using an approach inspired by sequential Monte Carlo methods, where topic assignments are resampled for the previously observed terms to simulate topical mixtures over observed terms. The likelihood is given by the summation over topics of the product between the probability of the topic in the document and the probability of the term under the topic distribution. This procedure is repeated for a number of iterations (particles) and the likelihood of the term is given by averaging the per-particle likelihood.

4.2.2 Topic coherence

A topic is said to be coherent when its most likely terms can be interpreted and associated with a single semantic concept [43]. For instance, ‘a bag of egg noodles’, ‘a package of prepared stir fry’, and ‘a sachet of Chinese stir fry sauce’ are items that can be easily associated with the topic of ‘Asian stir fry’. On the other hand, a non-coherent topic highlights products that do not seem to fulfil a particular customer need. For example, ‘a bag of egg noodles’, ‘a bunch of bananas’, and ‘a lemon cake’ are items that together do not convey a clear purpose.

Human judgement on topic coherence tends to correlate with metrics of product co-occurrence such as the Pointwise Mutual Information (PMI) and Normalized Pointwise Mutual Information (NPMI) [109] shown in [43, 110]. PMI measures the probability of seeing two products within the same topic in comparison to the probability of seeing them individually. NPMI standardizes PMI, providing a score in the range of $[-1,1]$. NPMI towards 1 corresponds to high co-occurrence.

$$\mathrm{PMI}(w_i, w_j) = \log\Big(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\Big); \quad i \neq j,\ 1 \leq i, j \leq 15. \tag{4.2}$$

$$\mathrm{NPMI}(w_i, w_j) = \frac{\mathrm{PMI}(w_i, w_j)}{-\log P(w_i, w_j)}; \quad i \neq j,\ 1 \leq i, j \leq 15. \tag{4.3}$$

In the literature, average NPMI and PMI are usually measured using the top 10 terms [110, 112, 113, 43]. However, we choose to use the 15 most probable products given that human annotators are comfortable assessing 10 or more items but less than 20 items per topic. Thus, we will interpret and compute NPMI using the top 15 products.

Instead of selecting terms by their probability, they can be selected through distributional transformations [114, 115, 116], which highlight less frequent but topic-wise unique products. However, transformations may select terms with low probabilities under the topic distribution.

The coherence measure of a single topic is given by the average of the NPMI scores. For simplicity, we will refer to this measure as NPMI. Here, we focus on NPMI since it has been shown to have a higher correlation with the human evaluation of topic coherence than PMI [110].

4.2.3 Topic distinctiveness

Topic distinctiveness refers to the semantic dissimilarity of one topic in comparison to the topics of the same sample. For instance, 'a bottle of sparkling water hint apple', 'a bottle of sparkling water hint grape', and 'a bottle of sparkling water hint orange' are items that are interpreted as the topic of 'flavoured sparkling water'. This topic and the 'Asian stir fry' topic are distinct from one another. If a topic in the posterior sample is characterised by 'a bottle of sparkling water hint lemon', 'a bottle of sparkling water hint mango' and 'a bottle of sparkling water hint lime', it is interpreted as non-distinctive from 'flavoured sparkling water' since both topics exhibit the same theme.

Several measures have been used to identify similar topics: KL-divergence [64, 117, 118], the average log odds ratio [113], and the cosine distance [72, 119, 37, 120]. [38, 120] showed that cosine distance outperforms other distributional similarity measures, such as KL-divergence, Jensen Shannon divergence, Euclidean distance, and Jaccard similarity, according to human judgment on topic similarity. Thus, we define the distinctiveness of a topic ϕ_i^t of posterior draw t as the minimum of the cosine distances between the topic and the other topics $\Phi^t \setminus \phi_i^t$ within the same posterior sample, denoted by :

$$\mathrm{CD}_{\min}\left(\phi_i^t, \Phi^t \setminus \phi_i^t\right) = \min\left[\mathrm{CD}(\phi_i^t, \phi_1^t), \ldots, \mathrm{CD}(\phi_i^t, \phi_{i-1}^t), \mathrm{CD}(\phi_i^t, \phi_{i+1}^t), \ldots, \mathrm{CD}(\phi_i^t, \phi_K^t)\right], \tag{4.4}$$

where

$$\mathrm{CD}\left(\phi_i, \phi_j\right) = 1 - \frac{\phi_i \cdot \phi_j}{\parallel \phi_i \parallel \parallel \phi_j \parallel}. \tag{4.5}$$

Cosine distance between topics measures a slightly different aspect of a topic compared to the model likelihood, and thus the model may warrant the existence of two similar topics in terms of cosine distance, showing a low minimum distance. The distinctiveness of a set of topics in a posterior sample is given by the average per-topic distinctiveness.

4.2.4 Topic credibility

When comparing different LDA posterior samples, topics may appear and disappear as a result of posterior uncertainty, which negatively affects practitioners' confidence in the method. While topic distinctiveness within the same posterior sample is good, the high cosine distance of topic ϕ_i^t with all topics Φ^s in posterior sample $s \neq t$ indicates uncertainty about ϕ_i^t. To measure topic credibility of topic ϕ_i^t in posterior sample t, we compute the maximum cosine similarity between ϕ_i^t and all topics within posterior sample Φ^s, for $s \neq t$, and average across all posterior samples $s \neq t$. If a topic is highly credible, then we expect a very similar topic to appear in every single posterior sample, hence the average cosine similarity will be high. Note here that we are using cosine similarity, rather than cosine distance, to capture topic credibility.

In other words,

$$\mathrm{CS}_{\max}\left(\phi_i^t, \Phi^s\right) = \max\left[\mathrm{CS}(\phi_i^t, \phi_1^s), \ldots, \mathrm{CS}(\phi_i^t, \phi_K^s)\right], \tag{4.6}$$

where

$$\mathrm{CS}\left(\phi_i, \phi_j\right) = \frac{\phi_i \cdot \phi_j}{\parallel \phi_i \parallel \parallel \phi_j \parallel}. \tag{4.7}$$

Averaging across all other posterior samples,

$$\overline{\mathrm{CS}_{\max}}(\phi_i^t, \Phi^{1:S}) = \frac{\sum_{s \neq t} \mathrm{CS}_{\max}\left(\phi_i^t, \Phi_:^s\right)}{S-1}. \tag{4.8}$$

A large average of the maximum similarities (i.e., minimum distances) across samples indicates that the topic appears with high similarity across posterior samples. The credibility of a set of topics is given by the average per-topic credibility.

4.3 Posterior summary of topic distributions

We introduce a methodology that aims to summarise the posterior distribution of a topic model by quantifying the recurrence of topic modes across posterior samples. Recurrent topics tend to appear several times across posterior draws, showing higher credibility. To group topics across samples that represent the same theme, we use a hierarchical clustering approach that retrieves clusters of topic distributions with high similarity. The resulting clusters are used to quantify the topic posterior recurrence of a clustered topic, which is ultimately used to identify and filter out topics of high uncertainty. In this chapter, this methodology is used to summarise LDA posterior samples; however, this methodology can be used to summarise other topics models as we will show in subsequent chapters.

4.3.1 Hierarchical clustering

Agglomerative hierarchical clustering (AHC) is a widely used statistical method that groups units according to their similarity, following a bottom-up merging strategy. The algorithm starts with as many clusters as input topics, and at each step, the AHC merges the pair of clusters with the smallest distance. AHC finishes when all the units are aggregated in a single cluster or when the distance among clusters is larger than a fixed threshold. AHC does not require the user to fix the number of clusters a priori; instead, the clustering dendrogram can be 'cut' at a user's desired level, potentially informed by domain knowledge.

We use the AHC algorithm to aggregate and fuse topics from multiple posterior samples. To quantify cluster similarity, we use CD and the average linkage method. We opt for CD since it has outperformed correlation on human evaluation of topic similarity [38] and human rating of posterior variability [120]. We opt for the *average* linkage method since, empirically, it has worked better than *single* and *complete* linkage methods, i.e., single linkage tended to create an extremely large cluster of low coherence, and complete linkage tended to create clusters of low distinctiveness. However, we slightly modify the algorithm to merge only topics that come from different posterior samples and whose cosine distance is lower than a user-specified threshold. In this manner, we avoid merging topics that belong to the same posterior sample or that differ to such a large extent that merging them is meaningless. See in Appendix 1 for the pseudo-code.

The AHC retrieves a collection of clusters $C_1, ..., C_N$, which are represented by a *clustered topic* $\overline{\phi_k}$ with a *cluster size* $|C_k|$, where $k = 1, ...N$. The clustered topic is the average distribution of the topics that share the same membership. The cluster size is the number of members, e.g., clustering 100 identical posterior samples of 50 topics would retrieve 50 clusters of 100 members each. The cluster size also represents the uncertainty related to the clustered topic. For instance, a cluster of size one indicates that its associated topic does not reappear in other posterior samples. On the other hand, a recurrent topic would be associated with a cluster with a large cluster size, indicating that the topic consistently reappears across multiple samples. Thus, we measure the recurrence of a topic by its cluster size:

$$\text{recurrence}(\overline{\phi}_i) = |C_i|. \tag{4.9}$$

Subsequently, subsets of clustered topics filtered by their recurrence are evaluated to identify a subset of clustered topics with high credibility. As we will show in the next section, cluster size as a measure of topic recurrence leads to subsets of better topic quality.

4.4 LDA application to grocery retail

We apply topic models in the domain of the grocery retail industry, where topics are distributions over a fixed assortment of products and transactions are described as mixtures of topics. We analyse grocery transactions from a major retailer in the UK. Transactions are sampled randomly, covering 100 nationwide superstores between September 2017 and August 2018. The training data set contains 36,000 transactions and a total of 392,840 products and the test data set contains 3,600 transactions and a total of 38,621 products. Transactions contain at least three products and 10 products on average. The product assortment contains 10,000 products which are the most monthly frequent, ensuring the selection of seasonal and non-seasonal products. We count unique products in transactions, disregarding the quantities of repetitive products. For instance, five loose bananas count as one single product (loose banana). We do not use an equivalent of 'stop words' list (highly frequent terms), as we consider that every product or combination of them describe different customer needs. We disregard transactions with fewer than three products assuming that smaller transactions do not have enough products to exhibit a customer need. Transactions are not linked to customers or to other transactions in our analysis. Despite this, we identify customer behaviours from transactional data. No personal customer data were used for this research.

4.4.1 User study on topic interpretability and similarity

To aid the interpretation of topics within the context of the application, meaningful NPMI and cosine similarity thresholds are required. To this end, we carried out a user study to collect human judgement on the interpretability of individual topics and the similarity between pairs of topics and, ultimately, set empirical thresholds driven by users' interpretations. Experts from a leading data science company specialising in retail analytics participated in the user study.

Users were asked to evaluate topics using a discrete scale from 1 to 5. For similarity between a pair of topics, a score of 1 refers to highly different topics, and a score of 5 refers to highly similar topics. For interpretability, a score of

1 refers to highly incoherent topics, and a score of 5 refers to highly coherent topics. Topics were obtained from $25, 50, 75, 100, 125, 150$-topic LDA with hyperparameters $\alpha = [0.1, 0.01]$ and $\beta = [0.01, 0.001]$. The range in the number of topics corresponds to an initial belief of having no less than 25 topics and no more than 150 topics. Topics were represented by the top 10 most probable products. 189 and 935 evaluations for topic distinctiveness and topic coherence were collected, respectively.

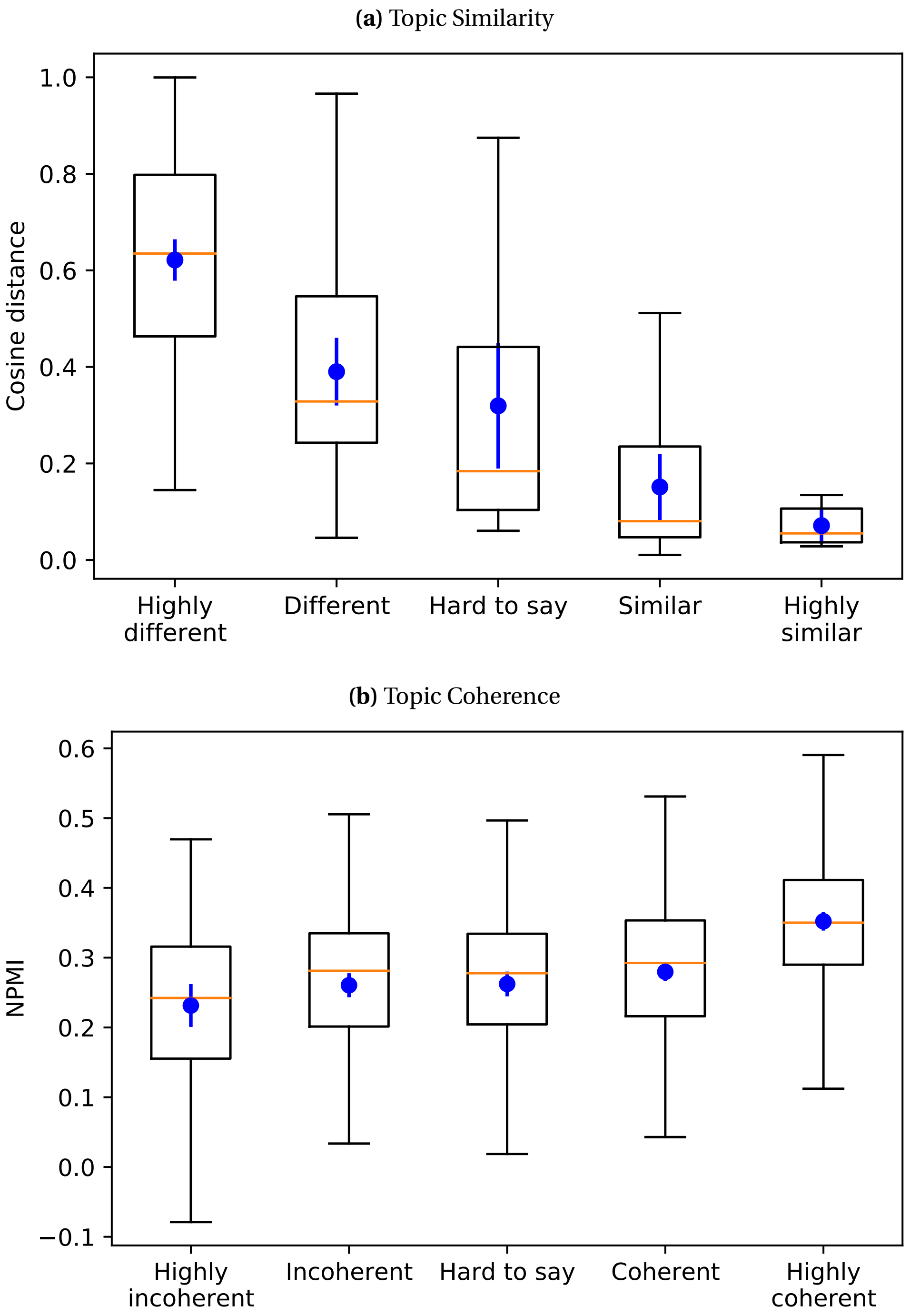
Figure 4.1b compares human judgment on topic coherence against NPMI. Despite the subtle positive correlation, there is no clear boundary of NPMI that can precisely identify coherent topics. However, we observe that 100% of topics with NPMI ≤ 0 were interpreted as highly incoherent, 65% of topics with NPMI ≥ 0.3 were interpreted as coherent, and 96% of topics with NPMI ≥ 0.5 were interpreted as highly coherent. We use these interpretations to guide the interpretation of topic coherence in the next sections.

Figure 4.1a compares human judgment on topic similarity against cosine distance. Unsurprisingly, the lower the cosine distance, the more similar the topic distributions. We observe that 70% of the pairs with CD ≤ 0.1 were interpreted as 'Similar' or 'Highly similar', and 95% of pairs with CD ≥ 0.5 were interpreted as 'Different' or 'Highly different'. While 38% of pairs were interpreted as 'Similar' or 'Highly similar' when $0.1 \leq$ CD ≤ 0.3, indicating some degree of topic similarity. Based on these results, we interpret topics with CD ≤ 0.1 as highly similar and with CD ≥ 0.5 as highly dissimilar. We use these thresholds to guide interpretations of topic distinctiveness and topic credibility.

4.4.2 LDA performance

We trained five LDA models with $K = 25, 50, 100, 200, 400$ topics, with symmetric Dirichlet hyperparameters $\alpha_k = 3/K$ and $\beta_v = 0.01$. Note that $\alpha_0 = 3$, which reflects the minimum transaction size. $\beta_v = 0.01$ is commonly used in the literature [108, 111].

Alternatively to a symmetric prior, an asymmetric prior over topical mixtures may improve topic interpretability by capturing highly frequent terms in a



(a) Topic Similarity

(b) Topic Coherence

Figure 4.1: Human evaluation on interpretability of individual topics and similarity between pairs of topics. Figure 4.1b shows coherence scores against topic NPMI. Figure 4.1a shows similarity scores against the cosine distance between compared topic distributions. Blue error bars show means and confidence intervals for the means. Interpreting results, a CD ≤ 0.1 indicates high similarity while CD ≥ 0.5 indicates high dissimilarity. It is also observed NPMI ≤ 0 responds to incoherent topics and NPMI ≥ 0.5 responds to highly coherent topics.

few topics [94]. Nevertheless, we do not use asymmetric priors since we found empirically that optimising the asymmetric hyperparameters lead to poor convergence of the Gibbs sampler in the context of our application.

For each model, four Markov chains are run for 50,000 iterations with a burn-in of 30,000 iterations. We evaluate convergence using the potential scale reduction factor $\widehat{R}$ [121]. We can assume that samples approximate the posterior distribution when $\widehat{R}$ is near 1, and values of $\widehat{R}$ below 1.1 are acceptable. Appendix B.1 shows the trace plot for the log-likelihood (measured at every 10 iterations) of LDA with 25, 50, 100, 200 and 400 topics. Chains for LDA with 25, 50, 100, 200 topics seem to be converged. The chains for LDA with 400 topics need to be further trained, however, preliminary evaluation of the topics from these chains already show lower performance than topics from chains with fewer topics.

LDA models are assessed on the four aforementioned quality aspects, using 20 posterior samples (five samples from each Markov chain). Samples are taken every 5,000 iterations which have non-significant autocorrelation as observed in Appendix B.2. Perplexity measures the generalisation of a group of topics, thus it is calculated for an entire collected sample. The other evaluation metrics are calculated at the topic level (rather than at the sample level) to illustrate individual topic performance.

Figure 4.2 shows the perplexity performance of LDA models. LDA samples of 50 and 100 topics tend to have the best generalisation capability. As observed in Figure 4.3a posterior draws with 25 and 50 topics show larger average NPMI, but there are no highly coherent topics (NMPI > 0.5). The posterior draws with 100 to 400 topics show some highly coherent topics, but also show many less coherent topics with low NPMI values. In agreement with [39], posterior samples with higher coherence do not necessarily have the highest likelihood, which is the case of 25-topic LDA samples. Figures 4.3b and 4.3c illustrate two topics with low/high coherence, respectively. The top topic displays product descriptions that do not show a specific meaning, purpose, or customer need. On the other hand, the bottom topic shows the soup topic, composed of branded soup items

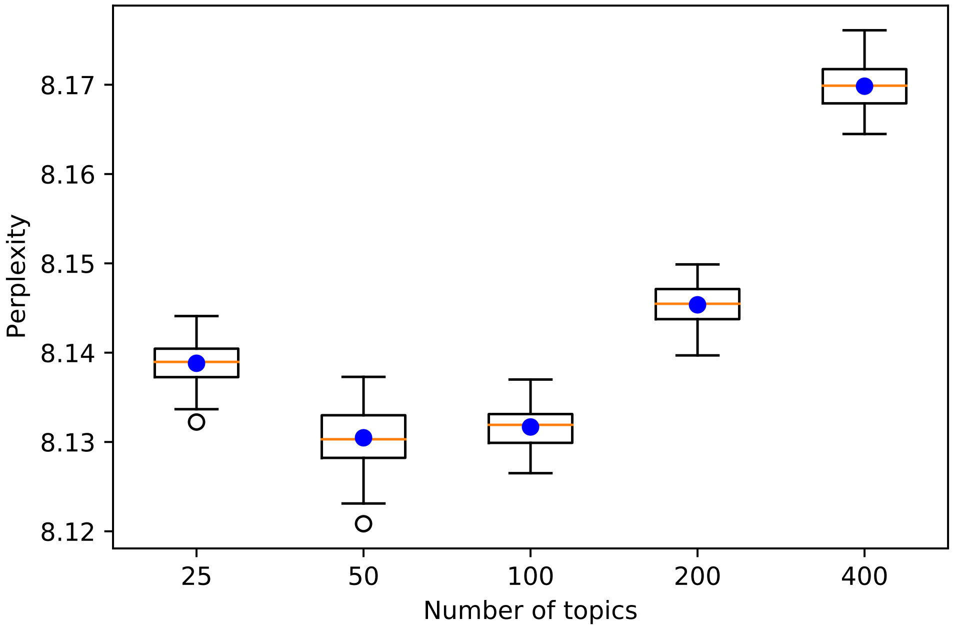
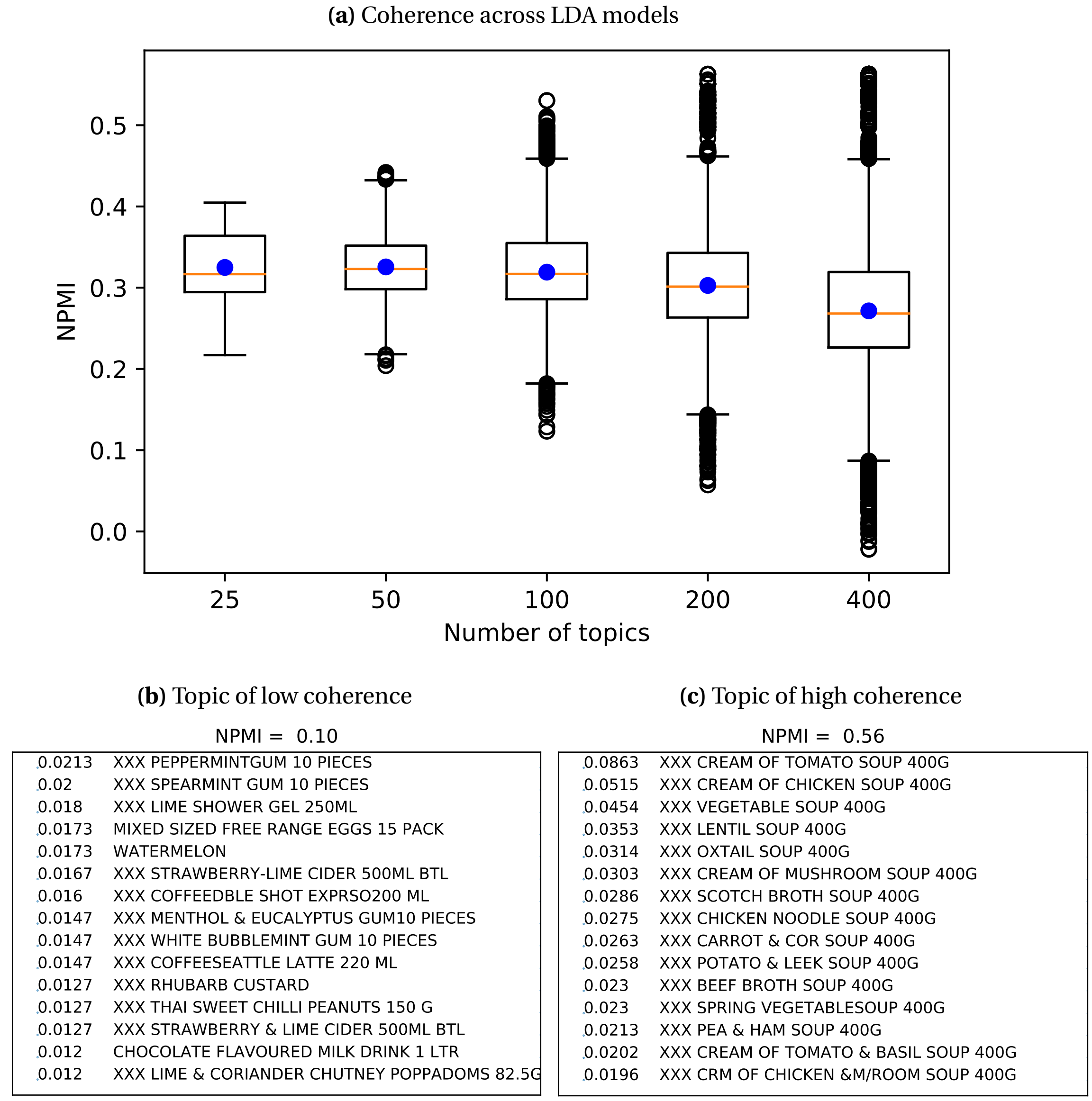


Figure 4.2: The perplexity of LDA models with 25, 50, 100, 200, 400 topics. Each boxplot represents the perplexity distribution over the 20 samples. Blue circles indicate the average perplexity; standard errors are smaller than the marker size.

that are frequently bought together due to promotional discounts.

In Figure 4.4a, we measure topic distinctiveness by computing the minimum cosine distance among topics of the same posterior draw. If two topics exhibit the same theme, and thereby similar distributions, then the cosine distance is close to 0. We observe that the majority of topics are highly distinct (CD ≥ 0.5) within their posterior draw. However, as expected, the larger the model, the more topics with some degree of similarity (CD ≤ 0.3) as seen in LDA models with 100 to 400 topics. Figures 4.4b and 4.4c show two topics with some degree of similarity, both illustrate collections of produce and red meat.

In Figure 4.5a, we measure topic credibility by averaging the maximum cosine similarity between a topic and the topics from the remaining posterior samples, so for each topic and each sample, there is one maximum cosine similarity from each remaining posterior sample. If one topic constantly appears across samples, then the average maximum cosine similarity tends to 1. Vice-versa, if the topic is not part of other samples, then the maximum cosine similarity of



(a) Coherence across LDA models

(b) Topic of low coherence

NPMI = 0.10

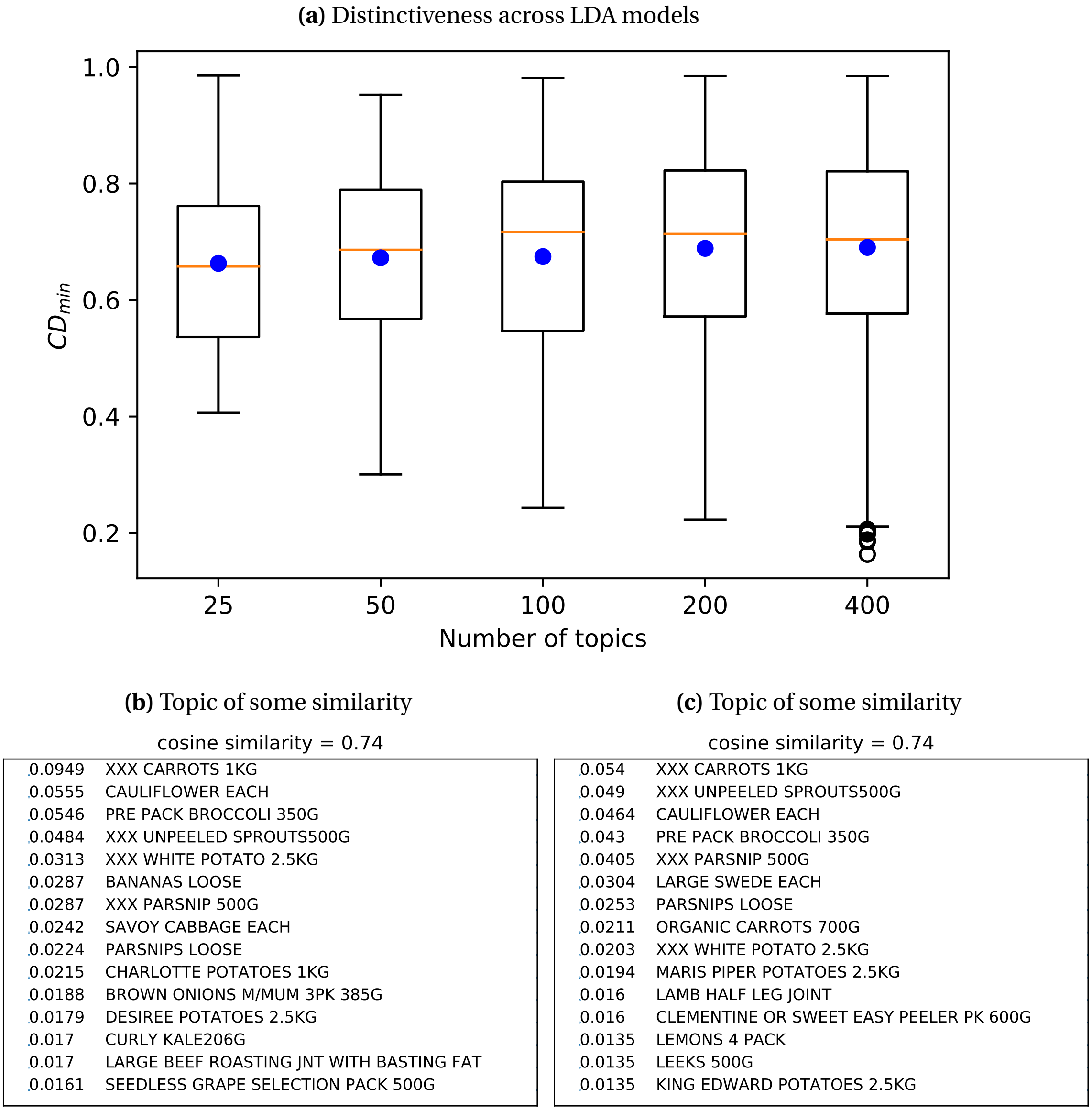
0.0213	XXX PEPPERMINTGUM 10 PIECES
0.02	XXX SPEARMINT GUM 10 PIECES
0.018	XXX LIME SHOWER GEL 250ML
0.0173	MIXED SIZED FREE RANGE EGGS 15 PACK
0.0173	WATERMELON
0.0167	XXX STRAWBERRY-LIME CIDER 500ML BTL
0.016	XXX COFFEEDBLE SHOT EXPRSO200 ML
0.0147	XXX MENTHOL & EUCALYPTUS GUM10 PIECES
0.0147	XXX WHITE BUBBLEMINT GUM 10 PIECES
0.0147	XXX COFFEESEATTLE LATTE 220 ML
0.0127	XXX RHUBARB CUSTARD
0.0127	XXX THAI SWEET CHILLI PEANUTS 150 G
0.0127	XXX STRAWBERRY & LIME CIDER 500ML BTL
0.012	CHOCOLATE FLAVOURED MILK DRINK 1 LTR
0.012	XXX LIME & CORIANDER CHUTNEY POPPADOMS 82.5G

(c) Topic of high coherence

NPMI = 0.56

0.0863	XXX CREAM OF TOMATO SOUP 400G
0.0515	XXX CREAM OF CHICKEN SOUP 400G
0.0454	XXX VEGETABLE SOUP 400G
0.0353	XXX LENTIL SOUP 400G
0.0314	XXX OXTAIL SOUP 400G
0.0303	XXX CREAM OF MUSHROOM SOUP 400G
0.0286	XXX SCOTCH BROTH SOUP 400G
0.0275	XXX CHICKEN NOODLE SOUP 400G
0.0263	XXX CARROT & COR SOUP 400G
0.0258	XXX POTATO & LEEK SOUP 400G
0.023	XXX BEEF BROTH SOUP 400G
0.023	XXX SPRING VEGETABLESOUP 400G
0.0213	XXX PEA & HAM SOUP 400G
0.0202	XXX CREAM OF TOMATO & BASIL SOUP 400G
0.0196	XXX CRM OF CHICKEN &M/ROOM SOUP 400G

Figure 4.3: Figure 4.3a: Topic-specific NPMI of 25/50/100/200/400-topic LDA model. Blue circles indicate the average NPMI; standard errors are smaller than the marker size. Figures 4.3b and 4.3c show (top) a topic with low coherence, (bottom) a topic with high coherence. Topics are illustrated with the probability and description of the top 15 products. Brands have been replaced by XXX.



(a) Distinctiveness across LDA models

(b) Topic of some similarity

cosine similarity = 0.74

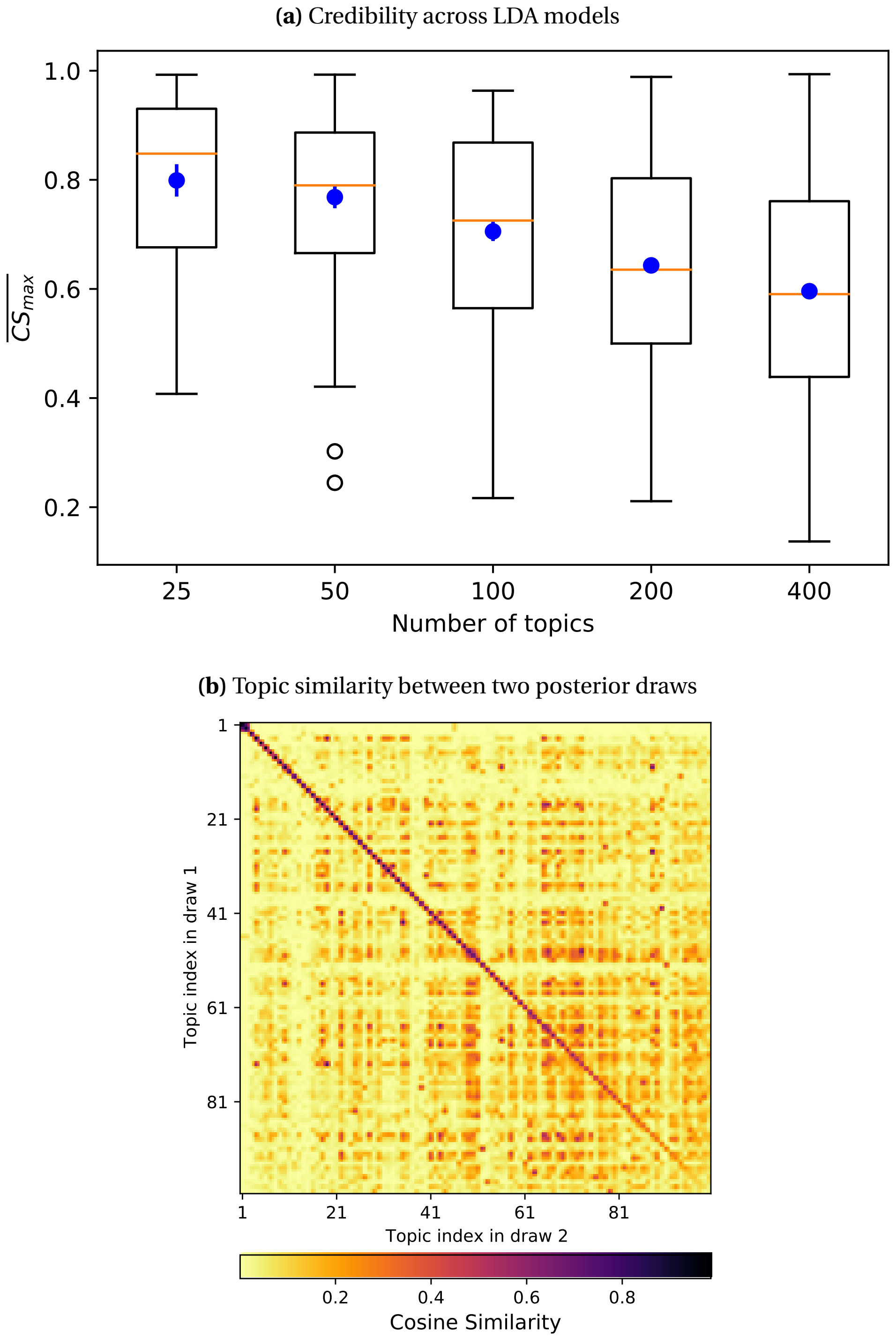
0.0949	XXX CARROTS 1KG
0.0555	CAULIFLOWER EACH
0.0546	PRE PACK BROCCOLI 350G
0.0484	XXX UNPEELED SPROUTS500G
0.0313	XXX WHITE POTATO 2.5KG
0.0287	BANANAS LOOSE
0.0287	XXX PARSNIP 500G
0.0242	SAVOY CABBAGE EACH
0.0224	PARSNIPS LOOSE
0.0215	CHARLOTTE POTATOES 1KG
0.0188	BROWN ONIONS M/MUM 3PK 385G
0.0179	DESIREE POTATOES 2.5KG
0.017	CURLY KALE206G
0.017	LARGE BEEF ROASTING JNT WITH BASTING FAT
0.0161	SEEDLESS GRAPE SELECTION PACK 500G

(c) Topic of some similarity

cosine similarity = 0.74

0.054	XXX CARROTS 1KG
0.049	XXX UNPEELED SPROUTS500G
0.0464	CAULIFLOWER EACH
0.043	PRE PACK BROCCOLI 350G
0.0405	XXX PARSNIP 500G
0.0304	LARGE SWEDE EACH
0.0253	PARSNIPS LOOSE
0.0211	ORGANIC CARROTS 700G
0.0203	XXX WHITE POTATO 2.5KG
0.0194	MARIS PIPER POTATOES 2.5KG
0.016	LAMB HALF LEG JOINT
0.016	CLEMENTINE OR SWEET EASY PEELER PK 600G
0.0135	LEMONS 4 PACK
0.0135	LEEKS 500G
0.0135	KING EDWARD POTATOES 2.5KG

Figure 4.4: Figure 4.4a: Topic-specific minimum cosine distance (among topics of the same posterior draw). Blue circles indicate the average minimum cosine distance; standard errors are smaller than the marker size. Figures 4.4b and 4.4c show two topics from a single Gibbs sample that show some similarity. Topics are illustrated with the probability and description of the top 15 products. Brands have been replaced by XXX.


(a) Credibility across LDA models

(b) Topic similarity between two posterior draws

Figure 4.5: Figure 4.5a: Topic-specific average maximum cosine similarity. For each topic, the maximum cosine similarity is calculated over the topics of a different posterior sample. Then, the average is taken over all maximum values. When a topic is highly credible, it will frequently appear across posterior samples, thus the average maximum cosine similarity tends to 1. Conversely, if a topic is highly uncertain and it does not appear in other posterior samples, then the maximum cosine similarity for each sample would tend to zero. Blue circles indicate the mean; standard errors are smaller than the marker size. Figure 4.5b: shows the cosine distance between topics of two posterior samples. Topics have been ordered using a greedy alignment algorithm that tries to find the best one-to-one topic correspondences.

each sample tends to 0, so does its average maximum cosine similarity. We observe 4%/ 3%/16%/ 25%/ 36% of topics with $\overline{\mathrm{CS}_{\mathrm{max}}} \leq 0.5$, indicating that they did not reappear in other posterior samples with high similarity. Figure 4.5b shows the cosine similarity matrix between two posterior LDA samples of 100 topics. Topics have been ordered using a greedy alignment algorithm that tries to find the best one-to-one topic correspondences as in [36]. This plot indicates that around one-fifth of the topics do not appear with high similarity ($\mathrm{CS} \leq 0.5$) in the other posterior draw. This implies that applying label-switching algorithms to resolve labelling for each posterior sample would inevitably 'match-up' topics that are semantically dissimilar. Instead of averaging over distinct modes, our methodology (described in the next section) would report separate clusters, each with its own credibility, reflecting the frequency with which each mode appears.

4.4.3 Clustering and selection of recurrent topics

In this section, we apply our methodology to summarise posterior LDA topic distributions and to quantify topic recurrence. We will show that topic recurrence can aid the selection of topics with better coherence, credibility and model generalisation.

We summarise LDA with 50, 100 and 200 topics. For each model, a bag of topics is formed from 20 samples that come from four separate Gibbs samplers. From each chain, samples are recorded after a burn-in period of 30,000 iterations and every 5,000 iterations to reduce autocorrelation as observed in Figure B.2.

We evaluate subsets of clustered topics obtained at different distance thresholds (cosine distance from 0 to 0.95 and every 0.05). We do not compute the evaluation metrics at each clustering step since computing perplexity is computationally expensive. Credibility is measured by comparing one clustering experiment against a second clustering experiment whose samples are recorded from four different Gibbs samplers. We do not further explore LDA samples with 25 and 400 topics, the former does not show a better variety of topics and the latter shows worse perplexities.

Figure 4.6 shows the evaluation of subsets of clustered topics obtained from clustering 50-topic LDA samples at different levels of topic recurrence, when the minimum cluster size is 1, 5, 10, and 20, representing 5%, 25%, 50%, and 100% of the samples.

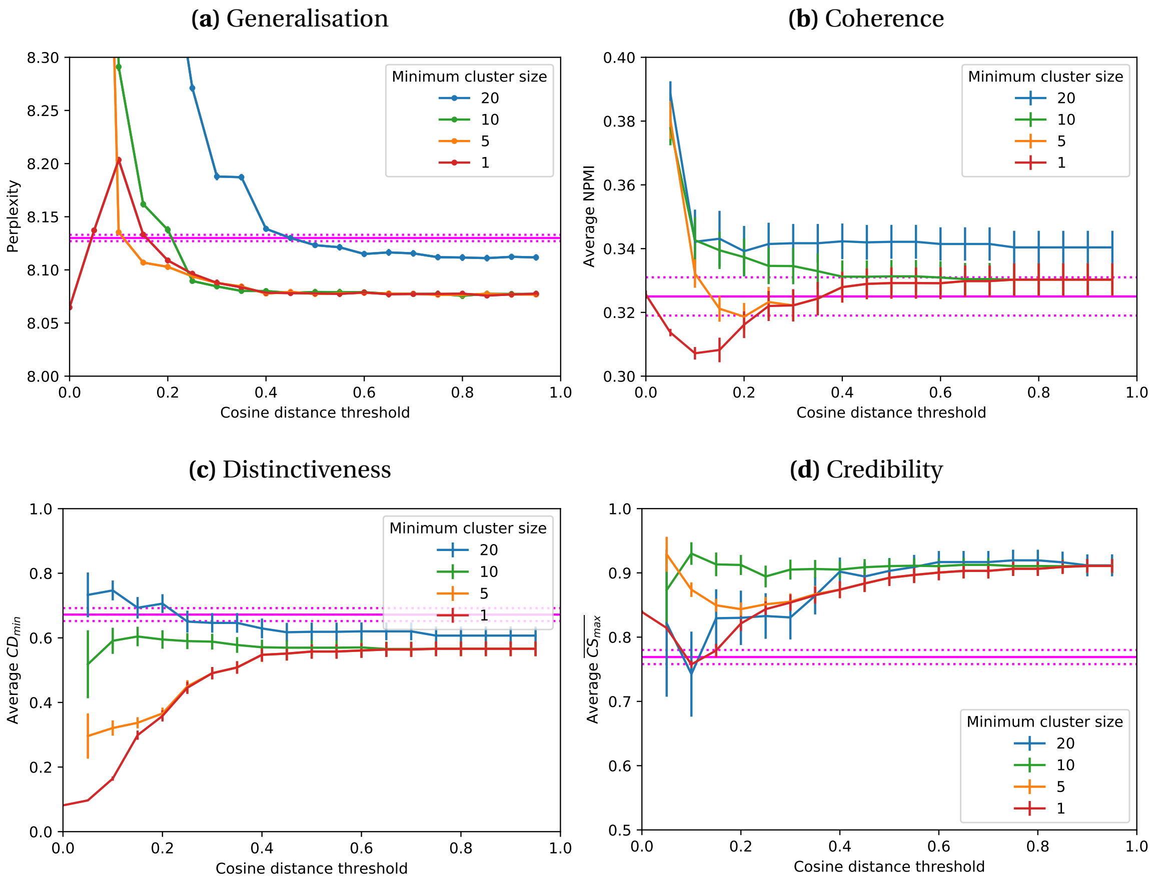


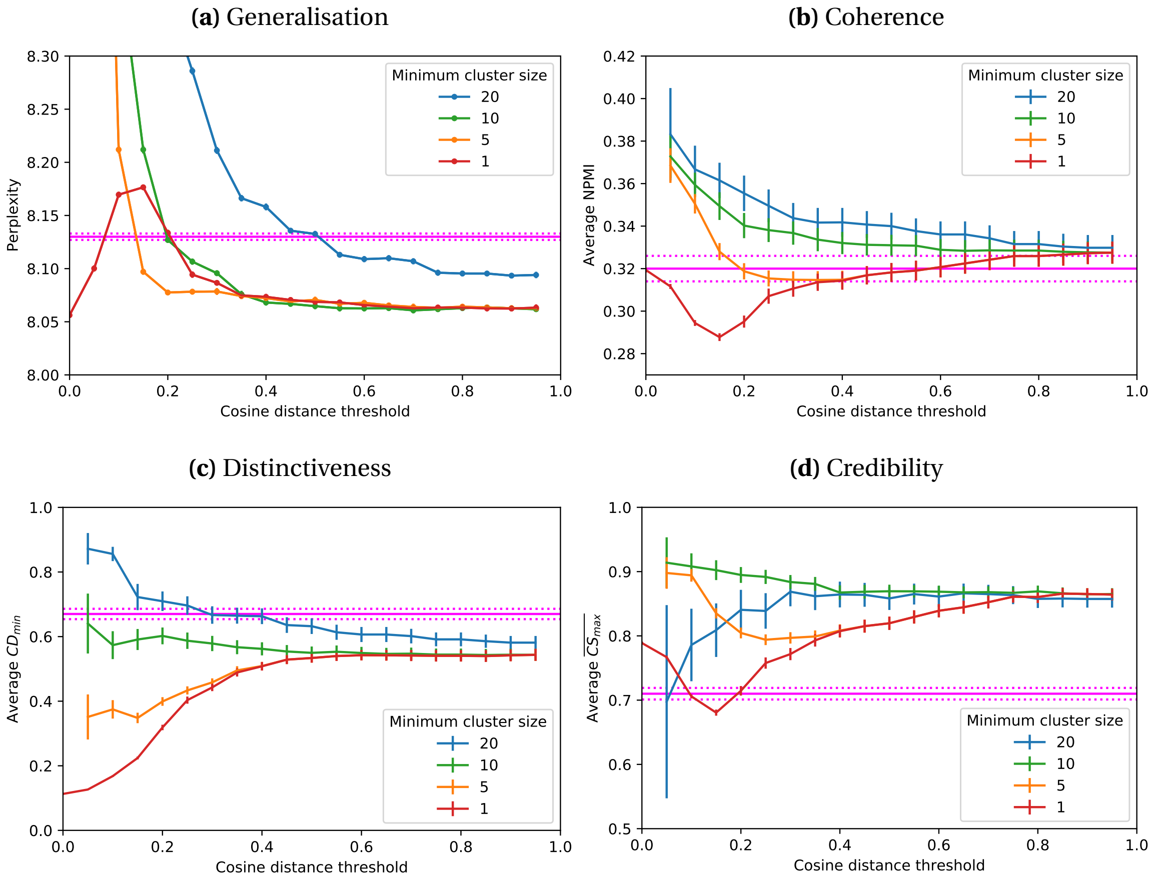
Figure 4.6: Subset evaluation using cosine distance (varying from 0 to 0.95 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). Clustered topics were obtained from clustering 20 samples of LDA with 50 topics. Vertical lines represent one standard error. Magenta lines show the average measures (± one standard error) of single LDA samples.

As observed in the perplexity plot in Figure 4.6a, the subset with a minimum cluster size 1 and cosine distance 0 shows the lowest perplexity; this is the original bag of 1,000 topics before merging. This subset has the lowest performance in distinctiveness; using this subset is inefficient as it contains too many repetitive topics. Subsets with minimum cluster size 1 and cosine distance 0.05 – 0.1 show increased perplexity because the most credible topics are reduced to a small number of clusters in comparison to topics that have not been clustered. Since a symmetric prior is used to compute perplexity, the uncertain topics outweigh the credible topics. More interestingly, various subsets show significantly better perplexity than the average perplexity of single LDA samples. For instance, the subset of cluster topics with a minimum cluster size of 10 and at cosine distance larger than 0.35.

Topic coherence in Figure 4.6b and topic distinctiveness in Figure 4.6c show that highly recurrent topics (with minimum cluster size 10) tend to be more coherent and distinctive. We also observe that measures of coherence and distinctiveness decrease when including topics of lower recurrence or when increasing the cosine distance (letting more clusters be merged, so the new cluster grows in size). Interestingly, the credibility plot in Figure 4.6d shows that the most credible subsets are formed with clusters of size 10 or more. Subsets of a minimum cluster size of 20 or cosine distance ≤ 0.1 are formed by a reduced number of clustered topics as shown in Figure 4.9. These topics may not recur with the same certainty in other samples, and, therefore, subsets with a small number of clusters tend to show high variability. Similar patterns are found when clustering LDA samples with 100 and 200 topics as shown in Figures 4.7 and 4.8.

Figure 4.9 shows the number of clustered topics obtained from clustering 20 LDA posterior samples with: 50 topics in Figure 4.9a, 100 topics in Figure 4.9b and 200 topics in Figure 4.9c, varying cosine distance thresholds and minimum cluster size. For visualisation purposes, subsets with a large number of clustered topics are not shown, i.e., subsets with more than 400 clustered topics in Figure 4.9c. Note that highly recurrent topics are always fewer than the number of topics in LDA samples. No more than 40/80/120 clustered topics appear in each of the 20 LDA samples with 50/100/200 topics as observed in Figures 4.9a, 4.9b, 4.9c, respectively. This confirms the low credibility and high uncertainty of some inferred topics.

Based on this analysis, we select a subset generated by minimum cluster size 10 and 0.35 CD threshold. A minimum cluster size of 20 may lead to greater coherence but lower perplexity and, vice-versa, a minimum cluster size of 1 or


(a) Generalisation

(b) Coherence

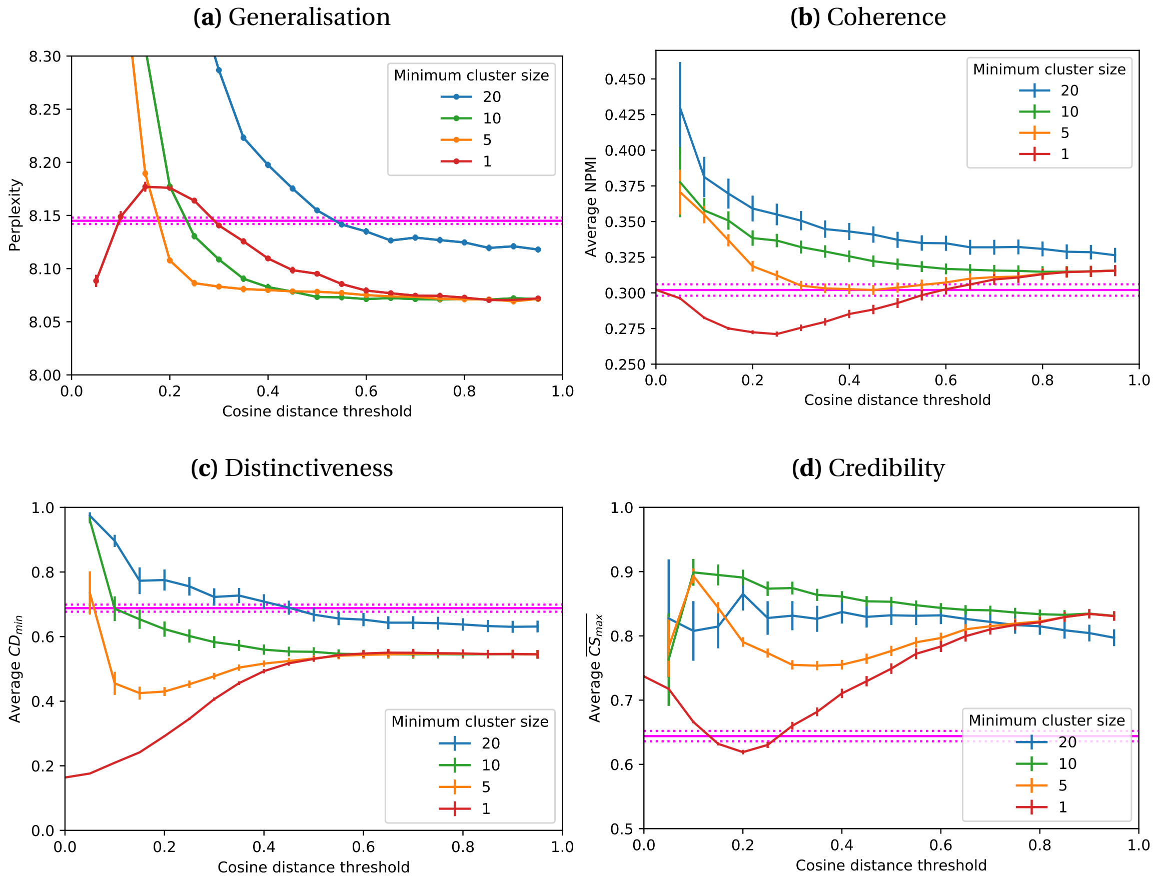
(c) Distinctiveness

(d) Credibility

Figure 4.7: Subset evaluation using cosine distance (varying from 0 to 0.95 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). Clustered topics were obtained from clustering 20 samples of LDA with 100 topics. Vertical lines represent one standard error. Magenta lines show the average measures (± one standard error) of single LDA samples.

5 leads to better perplexity but worse coherence. After the 0.35 CD threshold, perplexity is no longer significantly improved by increasing the CD threshold. Both thresholds are also used to select a subset of clustered topics obtained from 100-topic LDA samples, and 0.45 CD for clustered topics obtained from 200-topic LDA samples.

We repeat the 3 experiments of clustering LDA samples with 50, 100 and 200 topics, but this time, the clustering algorithm can merge topics within the same posterior sample. This allows us to determine if a topic model succeeds in identifying distinct customer behaviours. A model that is too large will identify many fewer clustered topics.


(a) Generalisation

(b) Coherence

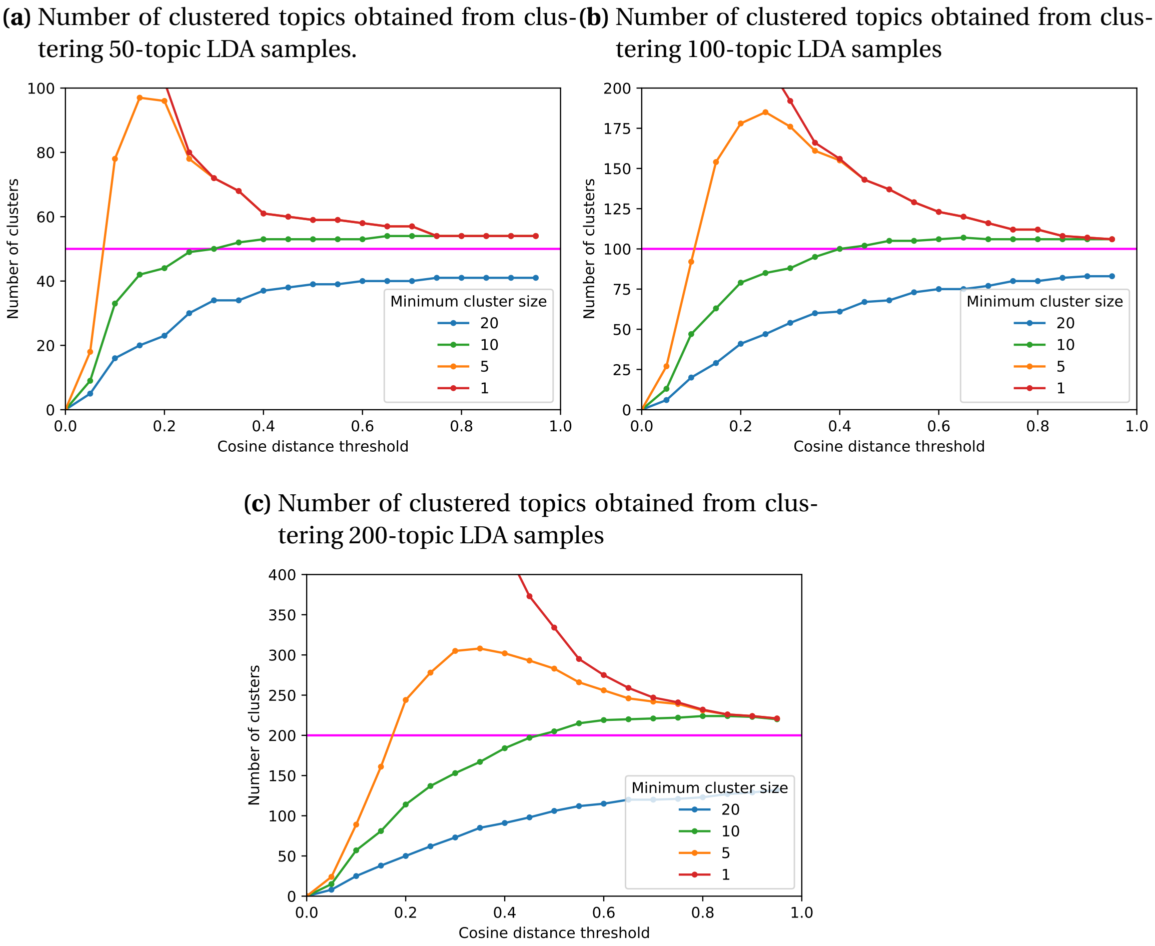
(c) Distinctiveness

(d) Credibility

Figure 4.8: Subset evaluation using cosine distance (varying from 0 to 0.95 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). Clustered topics were obtained from clustering 20 samples of LDA with 200 topics. Vertical lines represent one standard error. Magenta lines show the average measures (± one standard error) of single LDA samples.

In Table 4.1, we compare the performance of selected subsets of clustered topics (HC-LDA), and subsets of clustered topics formed by allowing merging topics from the same sample (HC-LDA-WS), against the average performance of LDA posterior samples. Subsets of clustered topics show significantly lower measures of generalisation, larger topic coherence and larger topic credibility than LDA topics from single samples. LDA samples show larger distinctiveness than those of subsets of clustered topics; this might be the case with LDA samples that contain highly distinctive but non-recurrent topics. Allowing merging topics from the same sample retrieves fewer topics but does not improve performance.

Different numbers of topics may produce similar performances. For exam-



(a) Number of clustered topics obtained from clustering 50-topic LDA samples.

(b) Number of clustered topics obtained from clustering 100-topic LDA samples

(c) Number of clustered topics obtained from clustering 200-topic LDA samples

Figure 4.9: The number of clustered topics obtained at varying cosine distance (from 0 to 0.95 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). The magenta line shows the number of topics in the LDA samples. For visualisation purposes, large number of clustered topics (with more than twice the number of topics in the LDA sample) are not shown, i.e, more than 100 clustered topics when clustering LDA samples of 50 topics.

ple, Table 4.1 shows that subsets of clustered topics achieve similar average measures of perplexity, coherence and credibility; LDA models with 50 and 100 topics show the same levels of perplexity, coherence and distinctiveness. However, LDA samples with a large number of topics (and thereby their derived clustered topics) cover a wider variety of topics, highlighting important customer behaviours. For example, the Scottish topic illustrated in Figure 4.11h is only found in LDA samples with 200 topics. Clustered topics may be included in a subset derived from larger LDA samples. For instance, Figure 4.10 shows that the clustered top-

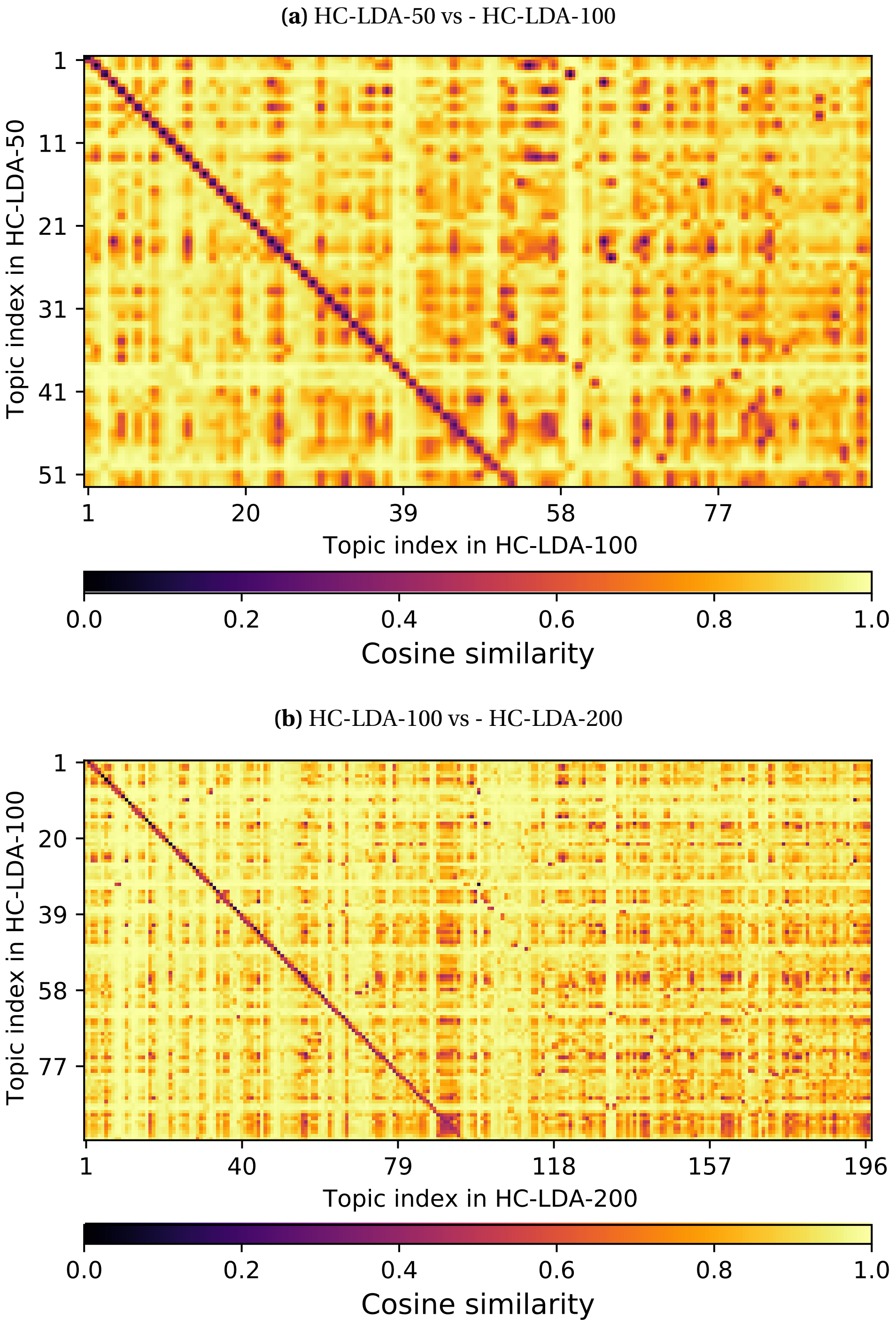
Table 4.1: Generalisation, coherence, distinctiveness and stability metrics of LDA samples and subsets of clustered topics (HC-LDA and HC-LDA-WS) obtained from clustering LDA samples with 50, 100 and 200 topics.

Model	Topics	Generalisation	Coherence	Distinctiveness	Credibility
		Perplexity	NPMI	$\mathrm{CD}_{\min}$	$\overline{\mathrm{CS}}_{\max}$
		Mean ± SE	Mean ± SE	Mean ± SE	Mean ± SE
LDA-50	50	8.130 ± 0.003	0.325 ± 0.006	**0.672** ± 0.020	0.769 ± 0.011
HC-LDA-50	52	**8.079** ± 0.006	**0.333** ± 0.006	0.580 ± 0.023	**0.916** ± 0.014
HC-LDA-WS-50	50	**8.083** ± 0.005	**0.333** ± 0.006	0.601 ± 0.021	**0.907** ± 0.014
LDA-100	100	8.131 ± 0.003	0.319 ± 0.006	**0.674** ± 0.016	0.716 ± 0.009
HC-LDA-100	96	**8.076** ± 0.006	**0.333** ± 0.005	0.565 ± 0.021	**0.890** ± 0.010
HC-LDA-WS-100	86	**8.086** ± 0.005	**0.331** ± 0.005	0.621 ± 0.018	**0.882** ± 0.012
LDA-200	200	8.145 ± 0.003	0.302 ± 0.004	**0.688** ± 0.011	0.644 ± 0.008
HC-LDA-200	198	**8.078** ± 0.005	0.32 ± 0.004	0.555 ± 0.014	**0.864** ± 0.007
HC-LDA-WS-200	145	8.132 ± 0.003	**0.335** ± 0.005	0.664 ± 0.011	**0.848** ± 0.011

ics in HC-LDA-50 (obtained from clustering 50-topic LDA samples) are also identified among the clustered topics in HC-LDA-100 (derived from 100-topic LDA samples). The latter is also identified among the clustered topics in HC-LDA-200 (derived from 200-topic LDA samples). Thus, the analysis of clustered topics obtained from LDA topics with a large number of topics may be warranted if the results reveal topics of interest, and the application of our clustering methodology can alleviate poor generalisation for the over-parametrised model.

4.5 British customer behaviour in grocery retail transactions

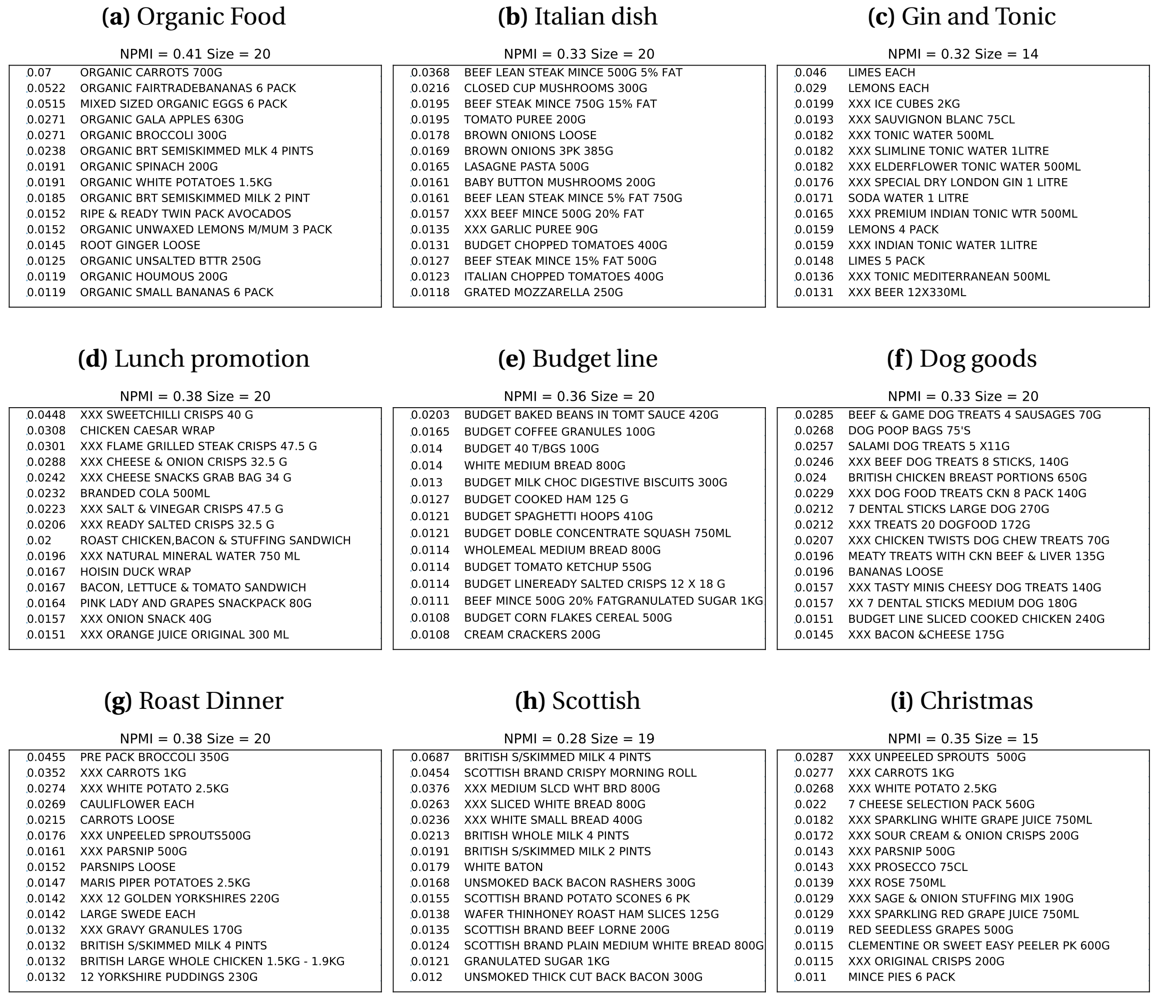
Interpreting topics by analysing the descriptions of the most likely products reveals customer preference for 'organic foods' as shown in Figure 4.11a, for ingredients for specific dishes such as the 'Italian dish' illustrated in Figure 4.11b,



(a) HC-LDA-50 vs - HC-LDA-100

(b) HC-LDA-100 vs - HC-LDA-200

Figure 4.10: Clustered topics correspondence between clustering of LDA samples with 50, 100 and 200 topics.



(a) Organic Food

NPMI = 0.41 Size = 20

0.07	ORGANIC CARROTS 700G
0.0522	ORGANIC FAIRTRADEBANANAS 6 PACK
0.0515	MIXED SIZED ORGANIC EGGS 6 PACK
0.0271	ORGANIC GALA APPLES 630G
0.0271	ORGANIC BROCCOLI 300G
0.0238	ORGANIC BRT SEMISKIMMED MLK 4 PINTS
0.0191	ORGANIC SPINACH 200G
0.0191	ORGANIC WHITE POTATOES 1.5KG
0.0185	ORGANIC BRT SEMISKIMMED MILK 2 PINT
0.0152	RIPE & READY TWIN PACK AVOCADOS
0.0152	ORGANIC UNWAXED LEMONS M/MUM 3 PACK
0.0145	ROOT GINGER LOOSE
0.0125	ORGANIC UNSALTED BTTR 250G
0.0119	ORGANIC HOUMOUS 200G
0.0119	ORGANIC SMALL BANANAS 6 PACK

(b) Italian dish

NPMI = 0.33 Size = 20

0.0368	BEEF LEAN STEAK MINCE 500G 5% FAT
0.0216	CLOSED CUP MUSHROOMS 300G
0.0195	BEEF STEAK MINCE 750G 15% FAT
0.0195	TOMATO PUREE 200G
0.0178	BROWN ONIONS LOOSE
0.0169	BROWN ONIONS 3PK 385G
0.0165	LASAGNE PASTA 500G
0.0161	BABY BUTTON MUSHROOMS 200G
0.0161	BEEF LEAN STEAK MINCE 5% FAT 750G
0.0157	XXX BEEF MINCE 500G 20% FAT
0.0135	XXX GARLIC PUREE 90G
0.0131	BUDGET CHOPPED TOMATOES 400G
0.0127	BEEF STEAK MINCE 15% FAT 500G
0.0123	ITALIAN CHOPPED TOMATOES 400G
0.0118	GRATED MOZZARELLA 250G

(c) Gin and Tonic

NPMI = 0.32 Size = 14

0.046	LIMES EACH
0.029	LEMONS EACH
0.0199	XXX ICE CUBES 2KG
0.0193	XXX SAUVIGNON BLANC 75CL
0.0182	XXX TONIC WATER 500ML
0.0182	XXX SLIMLINE TONIC WATER 1LITRE
0.0182	XXX ELDERFLOWER TONIC WATER 500ML
0.0176	XXX SPECIAL DRY LONDON GIN 1 LITRE
0.0171	SODA WATER 1 LITRE
0.0165	XXX PREMIUM INDIAN TONIC WTR 500ML
0.0159	LEMONS 4 PACK
0.0159	XXX INDIAN TONIC WATER 1LITRE
0.0148	LIMES 5 PACK
0.0136	XXX TONIC MEDITERRANEAN 500ML
0.0131	XXX BEER 12X330ML

(d) Lunch promotion

NPMI = 0.38 Size = 20

0.0448	XXX SWEETCHILLI CRISPS 40 G
0.0308	CHICKEN CAESAR WRAP
0.0301	XXX FLAME GRILLED STEAK CRISPS 47.5 G
0.0288	XXX CHEESE & ONION CRISPS 32.5 G
0.0242	XXX CHEESE SNACKS GRAB BAG 34 G
0.0232	BRANDED COLA 500ML
0.0223	XXX SALT & VINEGAR CRISPS 47.5 G
0.0206	XXX READY SALTED CRISPS 32.5 G
0.02	ROAST CHICKEN,BACON & STUFFING SANDWICH
0.0196	XXX NATURAL MINERAL WATER 750 ML
0.0167	HOISIN DUCK WRAP
0.0167	BACON, LETTUCE & TOMATO SANDWICH
0.0164	PINK LADY AND GRAPES SNACKPACK 80G
0.0157	XXX ONION SNACK 40G
0.0151	XXX ORANGE JUICE ORIGINAL 300 ML

(e) Budget line

NPMI = 0.36 Size = 20

0.0203	BUDGET BAKED BEANS IN TOMT SAUCE 420G
0.0165	BUDGET COFFEE GRANULES 100G
0.014	BUDGET 40 T/BGS 100G
0.014	WHITE MEDIUM BREAD 800G
0.013	BUDGET MILK CHOC DIGESTIVE BISCUITS 300G
0.0127	BUDGET COOKED HAM 125 G
0.0121	BUDGET SPAGHETTI HOOPS 410G
0.0121	BUDGET DOBLE CONCENTRATE SQUASH 750ML
0.0114	WHOLEMEAL MEDIUM BREAD 800G
0.0114	BUDGET TOMATO KETCHUP 550G
0.0114	BUDGET LINEREADY SALTED CRISPS 12 X 18 G
0.0111	BEEF MINCE 500G 20% FATGRANULATED SUGAR 1KG
0.0108	BUDGET CORN FLAKES CEREAL 500G
0.0108	CREAM CRACKERS 200G

(f) Dog goods

NPMI = 0.33 Size = 20

0.0285	BEEF & GAME DOG TREATS 4 SAUSAGES 70G
0.0268	DOG POOP BAGS 75'S
0.0257	SALAMI DOG TREATS 5 X11G
0.0246	XXX BEEF DOG TREATS 8 STICKS, 140G
0.024	BRITISH CHICKEN BREAST PORTIONS 650G
0.0229	XXX DOG FOOD TREATS CKN 8 PACK 140G
0.0212	7 DENTAL STICKS LARGE DOG 270G
0.0212	XXX TREATS 20 DOGFOOD 172G
0.0207	XXX CHICKEN TWISTS DOG CHEW TREATS 70G
0.0196	MEATY TREATS WITH CKN BEEF & LIVER 135G
0.0196	BANANAS LOOSE
0.0157	XXX TASTY MINIS CHEESY DOG TREATS 140G
0.0157	XX 7 DENTAL STICKS MEDIUM DOG 180G
0.0151	BUDGET LINE SLICED COOKED CHICKEN 240G
0.0145	XXX BACON &CHEESE 175G

(g) Roast Dinner

NPMI = 0.38 Size = 20

0.0455	PRE PACK BROCCOLI 350G
0.0352	XXX CARROTS 1KG
0.0274	XXX WHITE POTATO 2.5KG
0.0269	CAULIFLOWER EACH
0.0215	CARROTS LOOSE
0.0176	XXX UNPEELED SPROUTS500G
0.0161	XXX PARSNIP 500G
0.0152	PARSNIPS LOOSE
0.0147	MARIS PIPER POTATOES 2.5KG
0.0142	XXX 12 GOLDEN YORKSHIRES 220G
0.0142	LARGE SWEDE EACH
0.0132	XXX GRAVY GRANULES 170G
0.0132	BRITISH S/SKIMMED MILK 4 PINTS
0.0132	BRITISH LARGE WHOLE CHICKEN 1.5KG - 1.9KG
0.0132	12 YORKSHIRE PUDDINGS 230G

(h) Scottish

NPMI = 0.28 Size = 19

0.0687	BRITISH S/SKIMMED MILK 4 PINTS
0.0454	SCOTTISH BRAND CRISPY MORNING ROLL
0.0376	XXX MEDIUM SLCD WHT BRD 800G
0.0263	XXX SLICED WHITE BREAD 800G
0.0236	XXX WHITE SMALL BREAD 400G
0.0213	BRITISH WHOLE MILK 4 PINTS
0.0191	BRITISH S/SKIMMED MILK 2 PINTS
0.0179	WHITE BATON
0.0168	UNSMOKED BACK BACON RASHERS 300G
0.0155	SCOTTISH BRAND POTATO SCONES 6 PK
0.0138	WAFER THINHONEY ROAST HAM SLICES 125G
0.0135	SCOTTISH BRAND BEEF LORNE 200G
0.0124	SCOTTISH BRAND PLAIN MEDIUM WHITE BREAD 800G
0.0121	GRANULATED SUGAR 1KG
0.012	UNSMOKED THICK CUT BACK BACON 300G

(i) Christmas

NPMI = 0.35 Size = 15

0.0287	XXX UNPEELED SPROUTS 500G
0.0277	XXX CARROTS 1KG
0.0268	XXX WHITE POTATO 2.5KG
0.022	7 CHEESE SELECTION PACK 560G
0.0182	XXX SPARKLING WHITE GRAPE JUICE 750ML
0.0172	XXX SOUR CREAM & ONION CRISPS 200G
0.0143	XXX PARSNIP 500G
0.0143	XXX PROSECCO 75CL
0.0139	XXX ROSE 750ML
0.0129	XXX SAGE & ONION STUFFING MIX 190G
0.0129	XXX SPARKLING RED GRAPE JUICE 750ML
0.0119	RED SEEDLESS GRAPES 500G
0.0115	CLEMENTINE OR SWEET EASY PEELER PK 600G
0.0115	XXX ORIGINAL CRISPS 200G
0.011	MINCE PIES 6 PACK

Figure 4.11: Topics in grocery retail transactions in the UK. Each topic is characterised by the 15 products with the largest probabilities. Probabilities and products are sorted in descending order. Brand names have been replaced by XXX for anonymity purposes. Size shows the number of topics distributions associated with each cluster. Topics reflect a variety of shopping motivations, i.e., diet orientations, international dishes, specific events, ready-to-eat meals, preference for budget/premium product lines, pet ownership/household composition, special dishes, geography-related topics and temporal topics.

and items for a specific event such as a party given by the ‘Gin and Tonic’ topic illustrated in Figure 4.11c. Along with these topics, other identified topics show a preference for vegetarian foods, free-from lactose/gluten foods, ingredients for cooking Asian, Mexican, or Indian recipes, items for baking, picnics, barbecues and flower gifts.

Topics show customer preference for convenience foods such as ready-to-eat meals and lunch promotions (composed of a sandwich, a bottle of soda or water, and a package of prepared fruit or crisps) as shown in Figure 4.11d, preference for supermarket’s budget line or premium line, e.g., Figure 4.11e gathers products from a ‘budget line’ which offers products of a lower price than branded substitutes. Topics also suggest pet ownership, for instance, Figure 4.11f lists ‘dog goods’, including food, meat, and cleaning items. Other topics include baby-related foods and large size items indicating household composition.

Topics reveal customer motivations that are driven by geography or seasonality. For instance, Figure 4.11g depicts the ‘roast dinner’ which is a traditional British main meal that is typically served on Sunday. Topics also reveal specific shopping themes that are driven by product availability, i.e., seasonal products or locally-supplied products. For example, Figure 4.11h reveals Scottish-branded products in the ‘Scottish’ topic. Similarly, a ‘Northern Irish’ topic includes locally packed and supplied foods. Figure 4.11i shows the ‘Christmas’ topic which is characterised by mince pies, sparkling grape juice, vegetable, and snacks. Easter and Halloween are also depicted by topics that contain the iconic products: chocolate egg and pumpkin, respectively.

Our approach allows us to provide measures of uncertainty for each inferred topic. For example, the topics ‘organic food’, and ‘Italian dish’ appeared in every single posterior draw. Therefore, corresponding commercial decisions can be made with relative confidence in these shopping themes. On the other hand, less frequent topics can be identified, for instance, the topics ‘Scottish’ and ‘Christmas’ appeared 19 and 15 times, respectively, within the 20 LDA posterior draws. The lower frequency of these topics might be explained by the small representation in our data due to their regional/seasonal nature. More importantly, naive averaging of posterior draws would have damaged these topics by merging them with non-related topics.

4.6 Practical implications

Commercially speaking, the identification of product combinations that fulfil specific needs aids shelf management, planning aisle layouts and improving distribution. For instance, the most likely products of a topic could be placed on the same shelf, so customers easily find and choose products that otherwise might be forgotten. Understanding product combinations aids retailers to design marketing campaigns, i.e., creating spatial combos; and improving product distribution by allocating product combinations that are likely to be highly purchased. Product recommendations could be designed by offering ranked products within topics.

Describing transactions and customer purchases through topical mixtures aids retailers to design customised promotions, i.e., giving discounts for highly ranked products of a highly preferred topic; thereby, encouraging future transactions. Topical mixtures can be used in further customer analysis such as customer segmentation and customer profiling. Ultimately, understanding customer needs and stimulating customers with relevant offers improve customer experience and build brand loyalty.

Understanding grocery consumption through topic models not only aids marketing practices but also opens up new means for social, cultural and dietary research. For instance, topics such as ‘vegetarian’ and ‘organic’ demonstrate food movements and their significance in British food consumption. The ‘roast’ topic is an icon of British cuisine, but the significant presence of topics such as ‘stir fry’ and ‘fajita’ show how international dishes have been adopted by people in Britain. Topics such as ‘fizzy drinks,’ ‘beers’, ‘promotion meal’, and ‘convenience prepared foods’ can be related to high consumption of sugar, alcohol, salt and fat; and topics such as ‘prepared fruit’ and ‘seasonal produce’ can be related to healthy eating habits. Typically, dietary studies are limited to survey data such as food frequency questionnaires and open-ended dietary assessment [122, 123, 124, 125]. In contrast, topic models offer new methodologies to process transactions on a big scale and to track food consumption patterns at a low cost.

4.7 Summary

In this chapter, we proposed to evaluate topic models in four aspects: generalisation, topic coherence, topic distinctiveness, and topic credibility. We proposed a methodology that post-processes posterior topic distributions to identify customer behaviours and quantify their uncertainty. Recurrence, defined as the cluster size, provides a measure of uncertainty and allows users to select topics according to their desired level of (un)certainty. Using a survey with experts in retail analytics, we suggest thresholds of NPMI and CD that aid the evaluation of interpretability, distinctiveness and credibility. Empirically, we showed the advantages of the proposed methodology, which can capture topics of enhanced coherence, greater credibility (low uncertainty) and better generalisation than single posterior samples.

We identified credible and coherent topics that exhibit a variety of shopping motivations such as diet orientations, ingredients for specific dishes, foods for specific events, pet ownership, household composition, festivities, preference for budget/premium products, seasonal demand, to name a few. Analysis of grocery transactions through topic modelling has commercial implications. For instance, identifying groups of products that are frequently bought together for the fulfilment of specific needs helps to plan store layouts and design marketing campaigns. Topic models may also support sociological, cultural and dietary studies through the analysis of topic probabilities, which are associated with topics that exhibit (un)healthy food habits, eating trends and cultural differences.

Chapter 5

Identifying Regional Behaviours and Modelling Spatial Prevalence

This Chapter is largely based on a paper due to be submitted in Annals of Applied Statistics titled "Finding Regional Topics in British Grocery Retail Transactions".

In this chapter, we apply the segmented topic model to accommodate store hierarchy over transactions. In this manner, product co-occurrence is relevant within the store context, giving visibility of store-specific topics. We complement the analysis of regional behaviours by modelling topic prevalence over the UK. Using linear Gaussian process regression, we determine the significance of a topic over the constituent countries of the UK and English regions while accounting for spatial autocorrelation.

5.1 Introduction

The standard LDA model identifies topics of highly co-occurrent products across transactions that are assumed to be non-spatial, i.e., there is no spatial distinction between transactions and topics derived from a unique product assortment. However, stores from big retailers are located over large territories, even though several countries. Thus, stores offer a large variety of products that are supplied locally and nationally and not all the products are equally available. Customer behaviours may respond to local supply and local preferences for regional foods.

Customer preference for regional foods has been mainly discussed in anthropological and sociological works. *Regional foods* is defined as the food of a particular area of the country, often representing a regional speciality [126]. Regional foods are perceived as 'regional products' or 'regional recipes', which are associated with speciality, handicraft products or home-cooked dishes [127]. In comparison to these studies that employed market research methods such as focus groups and questionnaires, we aim to identify regional customer behaviours directly from transactional data. We define regional topics as groups of products that are frequently purchased together and show regional preference.

Spatial analysis of grocery retail data in the UK has shown applications of agent-based models and gravity models to investigate store catchment and store performance. For instance, [128] used an agent-based model to extract key customer behaviours about shopping frequency, shopping mission, store choice and spending. [129] applied the Huff gravity model, a spatial interaction modelling (SIM) technique, to create catchment areas and investigate the spatial variation on competition, sales area, trade intensity, among other factors. [130] also applied a SIM approach to forecast store patronage and store revenues using grocery retail data from Cornwall in the South West region. [131] explored spatiotemporal fluctuations of store sales and catchment areas in two English regions. [132] examined workplace geographies and census statistics to investigate store trading characteristics in inner London. None of the existing literature fully explores the spatial distribution of topics and thereby groups of products that are frequently purchased together (as opposed to individual products).

Customer behaviours can be modelled through the application of topic modelling (TM). As mentioned before, TM describes transactions as probabilistic mixtures of topics, and topics are distributions over a fixed product assortment that express customer behaviours; thereby, different topics exhibit different combinations of products with high probability. In the literature, latent Dirichlet allocation (LDA) [61] has been applied to retail data [11, 30, 29, 133, 97, 134] to identify customer behaviours. However, the standard LDA model cannot take into account store hierarchy over transactions, i.e., analysing topics within store context. Like LDA, several topic models do not exploit location, dismissing spatial patterns such as topics being more (or less) likely in certain areas or that nearby stores tend to show topics with similar intensities. Thus, analysing retail data through LDA might overlook topics that reflect regional customer behaviours.

In response, we employ the segmented topic model (STM) [44], a hierarchical extension of LDA, which not only provides topic distributions and transaction-specific topical mixtures, but also store-specific topical compositions. Transactions taking place at the same store are expected to exhibit more similar topical mixtures than transactions from other stores. STM harnesses store hierarchy over transactions, thereby product co-occurrence is relative to store context. Otherwise, regionally purchased products would be drowned out by the sheer volume of nationally supplied products, hampering the identification of regional topics.

The segmented topic model has not been applied in the analysis of market baskets to the best of our knowledge. STM has been mainly used in text applications. For instance, [135] used STM to match experts with questions in Community Question Answering websites, in which questions answered by a user are concatenated together to build a user profile. [136] applied STM to analyse multi-aspect sentiment in customer reviews from a variety of internet platforms, in which different aspects form part of a service review.

We summarise the posterior distribution of STM by identifying thematic modes following the clustering methodology in Section 4.3. As mentioned before, this methodology fuses posterior samples of several MCMC chains to identify recurrent topics and their associated uncertainties. Ultimately, topics that are grouped into clusters are represented by their average distribution, named *clustered* topic. We analyse the clustered topics obtained from 20 STM posterior samples and demonstrate that interpreting product descriptions and mapping store-specific topic probabilities lead to the identification of regional topics. Mapping topic probabilities also reveals cross-regional prevalence. We show that STM can identify regional clustered topics that cannot be captured under the LDA model.

Finally, we employ linear Gaussian Process regression (LGPR) [137, 138, 139] on regional topic probabilities to model topical prevalence over the UK. LGPR accounts for store meta-data and the geographical proximity between stores, neither of which are accounted for by STM. Specifically, we employ a spatial Gaussian process within a linear model on regional covariates, where spatial dependence is represented by the square exponential covariance function. LGPR allows us to identify and characterise variation in topic probabilities which are explained by spatial autocorrelation as well as geographical covariates.

We implement these methods on a nationwide collection of grocery transactions from a major supermarket chain in the UK and identify topics that characterise Scotland, Northern Ireland, Wales, and England from September 2017 to August 2018. After analysing each topic's most likely products and mapping their spatial distribution, we identify 6 topics that show significant regional differences, e.g., Welsh, English-North and Centre, Organic to name a few. LGPR complements the analysis by modelling store-specific topic probabilities as a response of constituent countries - Scotland, Northern Ireland, Wales - and English regions while accounting for geographical autocorrelation. We show that LGPR naturally achieves good out-of-sample predictive behaviour by borrowing information from neighbouring stores while affording an interpretable model with quantifiable uncertainty.

The analysis of grocery transactions within geographical space can provide insights into shopping patterns at a much higher resolution, which may help to customise store assortments and layout, improve distribution with locally purchased product combinations, launch marketing campaigns and predict thematic composition for new stores. The geographical resolution of shopping patterns may aid in our understanding of social and cultural habits driven by local demand and local supply.

5.2 STM application to grocery retail

We analyse grocery transactions from a major retailer in the UK. Transactions are sampled randomly, covering 100 nationwide superstores between September 2017 and August 2018. The training data set contains 36,000 transactions and a total of 392,840 products and the test data set contains 36 hundred transactions and a total of 38,621 products. Transactions contain around 10 products on average. The product assortment contains 10,000 products which are the most monthly frequent, ensuring the selection of seasonal and non-seasonal products. We count unique products in transactions, disregarding the quantities of repetitive products. For instance, five loose bananas count as one product (loose banana). We do not use an equivalent of 'stop words' list (highly frequent terms), as we consider that every product or combination of them tell different customer needs. Transactions with fewer than three products are disregarded assuming that smaller transactions do not have enough products to exhibit a regional topic. Transactions are not linked to customers or to other transactions in our analysis. Despite this, we identify customer behaviours from transactional data. No personal customer data were used for this research.

We apply the segmented topic model (STM) [44] to identify grocery topics and estimate store-specific distributions as well as transaction-specific topical mixtures. We explore the STM with 100 topics, assuming that 100 topics retrieve a large enough model to capture customer behaviours in the data of our application. STM is set with symmetric priors with Dirichlet hyperparameters $\alpha_k = 1,000/K$ and $\beta_v = 0.01$. The Dirichlet precision $\alpha_0 = 1,000$ is chosen empirically by assigning a significant value with respect to the number of active tables per store. $\beta_v = 0.01$ is commonly used in the literature [108, 111]. We empirically set the values of PDP hyperparameters $b = 3.0$ and $a = 0.5$ as a balance between flexibility and shrinkage and aid model exploration.

STM runs for four Markov chains with 100,000 iterations and a burn-in of 80,000 iterations. MCMC trace plots are shown in Appendix C.1 where the convergence is satisfactory. Posterior samples are recorded every 5,000 iterations to ensure little autocorrelation as shown in Appendix C.2.

As before, we do not use asymmetric priors since we found, empirically, that optimising the asymmetric hyperparameters lead to poor convergence of the Gibbs sampler in the context of our application.

5.2.1 Posterior summary of STM topic distributions

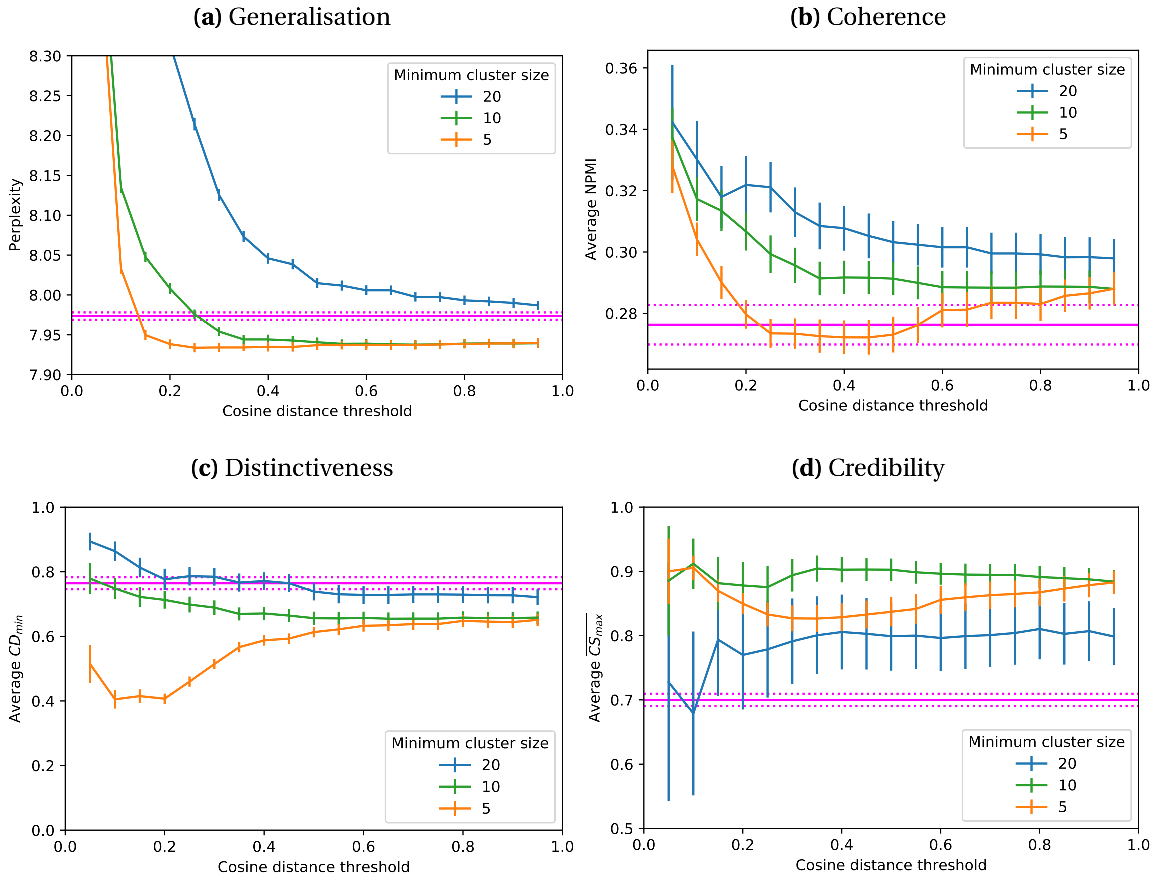
Summarising the posterior distribution of a topic model is challenging because the posterior distribution is often highly multi-modal, resulting in Gibbs samples that capture different semantic modes [37]. Thus, component-wise posterior averaging (after resolving component-labelling) inevitably merges different semantic concepts. Additionally, posterior variance results in topics of high uncertainty which can be less semantically meaningful and may not represent genuine themes [39, 40].

In response, we follow the methodology described in Section 4.3 to construct a summary of topical modes that captures credible topics across posterior draws and quantifies individual topic uncertainty. We take five posterior samples from each of the four aforementioned Markov chains, forming a bag of 2,000 topics. We form subsets of clustered topics varying the cosine distance from 0.05 to 0.95 with steps of 0.05 and minimum clusters size of five, 10 and 20.

5.2.2 Evaluation and selection of topic models

We evaluate subsets of clustered topics that have been formed by varying cosine distance threshold and acceptable recurrence (setting a minimum cluster size). Each subset is evaluated on four aspects: generalisation or predictive power of a subset of topics, coherence of individual topics, the distinctiveness of a topic w.r.t. the other topics in the same posterior sample, and credibility of a topic w.r.t. the topics from other posterior samples.

Topic coherence, distinctiveness and credibility are measured as described in Sections 4.2.2, 4.2.3 and 4.2.4. Model generalisation, however, is measured by the perplexity of unseen transactions given topics, store-specific topical mixtures



(a) Generalisation

(b) Coherence

(c) Distinctiveness

(d) Credibility

Figure 5.1: Evaluation of subsets of STM clustered topics. Subsets are formed with combinations of minimum cluster size and cosine distance thresholds. Horizontal and dotted lines show the average measures and ±2 standard error of the STM posterior samples.

and PDP parameters:

$$\text{Perplexity} = -\frac{\log P(\mathbf{w}'_d | \Phi, \theta_d, a, b)}{N'}, \tag{5.1}$$

where $\mathbf{w}'_d$ is a set of products in a held-out transaction at store d, N' is the number of products in $\mathbf{w}'_d$, $\Phi = [\phi_1, \phi_2, \ldots, \phi_K]$ the set of inferred topics, θ_d is the store-specific topical mixtures associated to store d, a and b are the PDP parameters.

As shown in Figure 5.1, we observe that subsets of clusters formed with a least 10 members (which represent 50% of the samples) and a cosine distance threshold ≥ 0.35 show greater coherence, credibility and generalisation, concurring with Section 4.4.3. Based on these results, we summarise the posterior distribution of STM with 104 topical modes using a cosine distance threshold of 0.35 and a minimum cluster size of 10. These 104 clustered topic distributions are used to obtain posterior samples of the store-specific topical mixtures. This time, we model transactions from 500 nationwide superstores between September 2017 and August 2018, which allow us to have a broader picture of topic probabilities around the UK. Topical mixtures are estimated by averaging 30 posterior samples recorded after a burn-in period of 1,000 iterations and every 500 iterations. MCMC trace plots are shown in Appendix C.3, where convergence is satisfactory. Samples show little autocorrelation as shown in Appendix C.4.

5.3 Regional British behaviours in grocery retail

In this section, we describe topics with regional patterns. First, we analyse the product descriptions of the most likely products. Second, we map the topic probability corresponding to 500 stores across the UK. Finally, we compare clustered topics obtained from STM and LDA and determine the advantage of STM over LDA in identifying regional patterns.

5.3.1 Interpreting regional behaviours

We interpret six out of the 104 clustered topics as they capture a regional pattern. The remaining topics show ubiquitous distributions over the UK. We interpret topics by analysing the product descriptions of the 15 products with the largest probabilities in each topic. Topics are manually named after the regional pattern or customer preference reflected on the product descriptions. Note that the illustrated topics appeared consistently across the 20 posterior samples (size = 20), indicating low posterior uncertainty.

Product descriptions in Figures 5.2b, 5.2a and 5.2c suggest foods provided locally or by local brands associated with Scotland, Northern Ireland and Wales. For instance, the Scottish topic includes ‘Scottish-branded skinless sausages’ and ‘Scottish-branded potato scones’, the Northern Irish topic shows the ‘North Ireland semi-skimmed milk’, ‘white potatoes packed in Northern Ireland’, and the Welsh topic contains ‘Welsh jacket potatoes’ and ‘Welsh-branded bread’. Hence, we name the Scottish topic, Northern Irish topic and Welsh topic after the nationality that their product descriptions suggest.

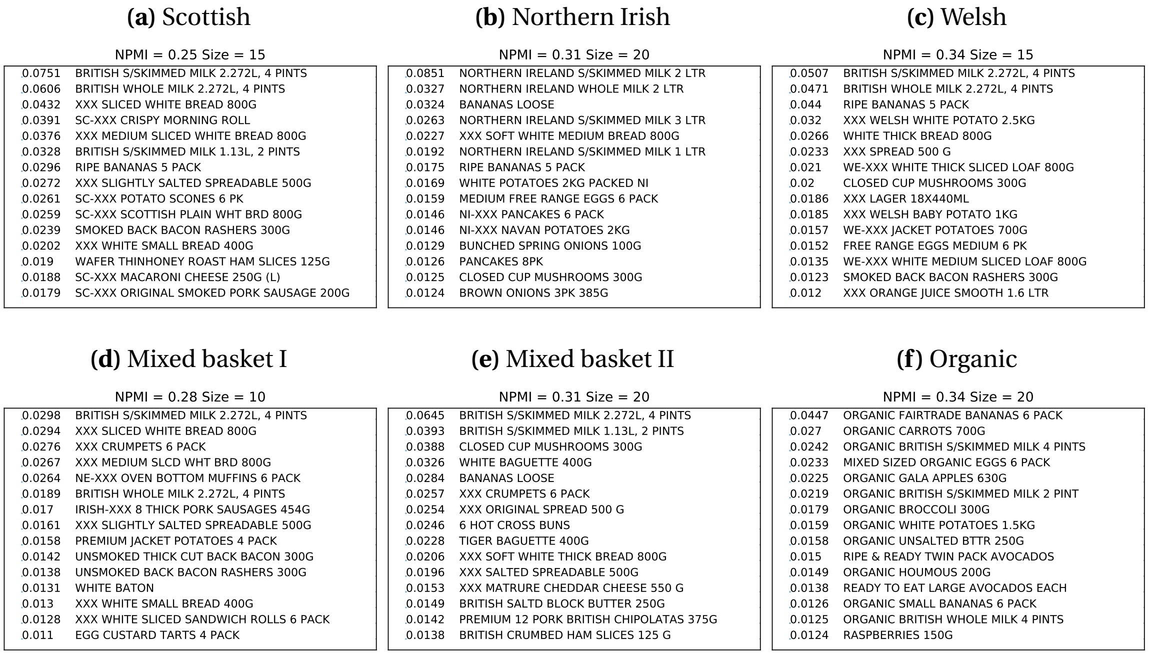
Figures 5.2d and 5.2e show a variety of products such as types of milk, types of bread, fruits and vegetables, etc. Close inspection of product descriptions such as ‘oven bottom muffin’, ‘fruit teacake’, and ‘potato and meat pie’ in Figure 5.2d, and ‘pork pies’ and ‘scotch eggs’ in Figure 5.2e may reveal a regional topic when regional expertise is available. Since these product descriptions do not provide interpretations that can be directly associated with specific regions, we name these topics ‘Mixed basket I’ and ‘Mixed basket II’. Figures 5.2f shows ‘organic’ quality foods, indicating a specific customer preference, however, the topic does not suggest any specific regional pattern.

Analysing topics by interpreting product descriptions may reveal the existence of regional topics, i.e., the Northern Irish, Scottish, Welsh topics. Topic interpretations may dismiss regional topics that exhibit products that are not directly linked with specific areas or local brands, such as the topics in Figures 5.2d, 5.2e and 5.2f. Thus, interpreting topic descriptions is not sufficient to identify geographically driven shopping motivations, reinforcing the need for exploring store-specific topical mixtures.

5.3.2 Mapping regional behaviours

Aiming to find topics with regional patterns, we map the store-specific topic probabilities using store locations to spatially describe the six previously discussed topics. First, store postcodes are linked with location coordinates through querying stores’ postcodes in the lookup table from the Office for National Statistics [140]. Second, and for each topic, store-specific topic probabilities are placed at their associated store location. Figure 5.3 shows the topic probabilities of the six clustered topics mapped across the UK.

Figures 5.3a, 5.3b, 5.3c clearly confirm that the Scottish, Northern Irish and Welsh topics are driven regionally. Each of them shows high topic probabilities at stores in their associated constituent country. Besides, Figure 5.3c shows the prevalence of the Welsh topic over neighbouring regions. Figure 5.3d shows high



(a) Scottish

NPMI = 0.25 Size = 15

0.0751	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0606	BRITISH WHOLE MILK 2.272L, 4 PINTS
0.0432	XXX SLICED WHITE BREAD 800G
0.0391	SC-XXX CRISPY MORNING ROLL
0.0376	XXX MEDIUM SLICED WHITE BREAD 800G
0.0328	BRITISH S/SKIMMED MILK 1.13L, 2 PINTS
0.0296	RIPE BANANAS 5 PACK
0.0272	XXX SLIGHTLY SALTED SPREADABLE 500G
0.0261	SC-XXX POTATO SCONES 6 PK
0.0259	SC-XXX SCOTTISH PLAIN WHT BRD 800G
0.0239	SMOKED BACK BACON RASHERS 300G
0.0202	XXX WHITE SMALL BREAD 400G
0.019	WAFER THINHONEY ROAST HAM SLICES 125G
0.0188	SC-XXX MACARONI CHEESE 250G (L)
0.0179	SC-XXX ORIGINAL SMOKED PORK SAUSAGE 200G

(b) Northern Irish

NPMI = 0.31 Size = 20

0.0851	NORTHERN IRELAND S/SKIMMED MILK 2 LTR
0.0327	NORTHERN IRELAND WHOLE MILK 2 LTR
0.0324	BANANAS LOOSE
0.0263	NORTHERN IRELAND S/SKIMMED MILK 3 LTR
0.0227	XXX SOFT WHITE MEDIUM BREAD 800G
0.0192	NORTHERN IRELAND S/SKIMMED MILK 1 LTR
0.0175	RIPE BANANAS 5 PACK
0.0169	WHITE POTATOES 2KG PACKED NI
0.0159	MEDIUM FREE RANGE EGGS 6 PACK
0.0146	NI-XXX PANCAKES 6 PACK
0.0146	NI-XXX NAVAN POTATOES 2KG
0.0129	BUNCHED SPRING ONIONS 100G
0.0126	PANCAKES 8PK
0.0125	CLOSED CUP MUSHROOMS 300G
0.0124	BROWN ONIONS 3PK 385G

(c) Welsh

NPMI = 0.34 Size = 15

0.0507	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0471	BRITISH WHOLE MILK 2.272L, 4 PINTS
0.044	RIPE BANANAS 5 PACK
0.032	XXX WELSH WHITE POTATO 2.5KG
0.0266	WHITE THICK BREAD 800G
0.0233	XXX SPREAD 500 G
0.021	WE-XXX WHITE THICK SLICED LOAF 800G
0.02	CLOSED CUP MUSHROOMS 300G
0.0186	XXX LAGER 18X440ML
0.0185	XXX WELSH BABY POTATO 1KG
0.0157	WE-XXX JACKET POTATOES 700G
0.0152	FREE RANGE EGGS MEDIUM 6 PK
0.0135	WE-XXX WHITE MEDIUM SLICED LOAF 800G
0.0123	SMOKED BACK BACON RASHERS 300G
0.012	XXX ORANGE JUICE SMOOTH 1.6 LTR

(d) Mixed basket I

NPMI = 0.28 Size = 10

0.0298	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0294	XXX SLICED WHITE BREAD 800G
0.0276	XXX CRUMPETS 6 PACK
0.0267	XXX MEDIUM SLCD WHT BRD 800G
0.0264	NE-XXX OVEN BOTTOM MUFFINS 6 PACK
0.0189	BRITISH WHOLE MILK 2.272L, 4 PINTS
0.017	IRISH-XXX 8 THICK PORK SAUSAGES 454G
0.0161	XXX SLIGHTLY SALTED SPREADABLE 500G
0.0158	PREMIUM JACKET POTATOES 4 PACK
0.0142	UNSMOKED THICK CUT BACK BACON 300G
0.0138	UNSMOKED BACK BACON RASHERS 300G
0.0131	WHITE BATON
0.013	XXX WHITE SMALL BREAD 400G
0.0128	XXX WHITE SLICED SANDWICH ROLLS 6 PACK
0.011	EGG CUSTARD TARTS 4 PACK

(e) Mixed basket II

NPMI = 0.31 Size = 20

0.0645	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0393	BRITISH S/SKIMMED MILK 1.13L, 2 PINTS
0.0388	CLOSED CUP MUSHROOMS 300G
0.0326	WHITE BAGUETTE 400G
0.0284	BANANAS LOOSE
0.0257	XXX CRUMPETS 6 PACK
0.0254	XXX ORIGINAL SPREAD 500 G
0.0246	6 HOT CROSS BUNS
0.0228	TIGER BAGUETTE 400G
0.0206	XXX SOFT WHITE THICK BREAD 800G
0.0196	XXX SALTED SPREADABLE 500G
0.0153	XXX MATRURE CHEDDAR CHEESE 550 G
0.0149	BRITISH SALTD BLOCK BUTTER 250G
0.0142	PREMIUM 12 PORK BRITISH CHIPOLATAS 375G
0.0138	BRITISH CRUMBED HAM SLICES 125 G

(f) Organic

NPMI = 0.34 Size = 20

0.0447	ORGANIC FAIRTRADE BANANAS 6 PACK
0.027	ORGANIC CARROTS 700G
0.0242	ORGANIC BRITISH S/SKIMMED MILK 4 PINTS
0.0233	MIXED SIZED ORGANIC EGGS 6 PACK
0.0225	ORGANIC GALA APPLES 630G
0.0219	ORGANIC BRITISH S/SKIMMED MILK 2 PINT
0.0179	ORGANIC BROCCOLI 300G
0.0159	ORGANIC WHITE POTATOES 1.5KG
0.0158	ORGANIC UNSALTED BTTR 250G
0.015	RIPE & READY TWIN PACK AVOCADOS
0.0149	ORGANIC HOUMOUS 200G
0.0138	READY TO EAT LARGE AVOCADOS EACH
0.0126	ORGANIC SMALL BANANAS 6 PACK
0.0125	ORGANIC BRITISH WHOLE MILK 4 PINTS
0.0124	RASPBERRIES 150G

Figure 5.2: Most probable products in grocery regional topics. Each topic is interpreted using the 15 products with the largest probabilities. Probabilities and products are sorted in descending order. General brand names have been replaced by XXX. Local brands in North Ireland, Scotland, Wales and North of England have been replaced by NI-XXX, SC-XXX, WE-XXX, NE-XXX. NPMI and size are measures of topic coherence and recurrence.

topic probabilities concentrated in the North West and surrounding regions. Figure 5.3e shows high topic probabilities in the central and southern English regions. We rename both topics as 'North and Centre', and 'South and Midlands' due to their cross-regional predominance. Figure 5.3f, which maps the Organic topic, shows high probabilities concentrated in London.

In comparison to the Scottish, Northern Irish, and Welsh topics the interpretations of the North and Centre, South and Midlands, and Organic topics based on their most likely items do not easily suggest a geographical pattern. Thus, mapping the store-specific topic probabilities aids the analysis and interpretation of topics.

5.3.3 What does STM have that LDA does not?

STM shows two advantages over LDA. First, STM provides topical summaries for stores, by including the store hierarchy above transactions. Second, and less ob-

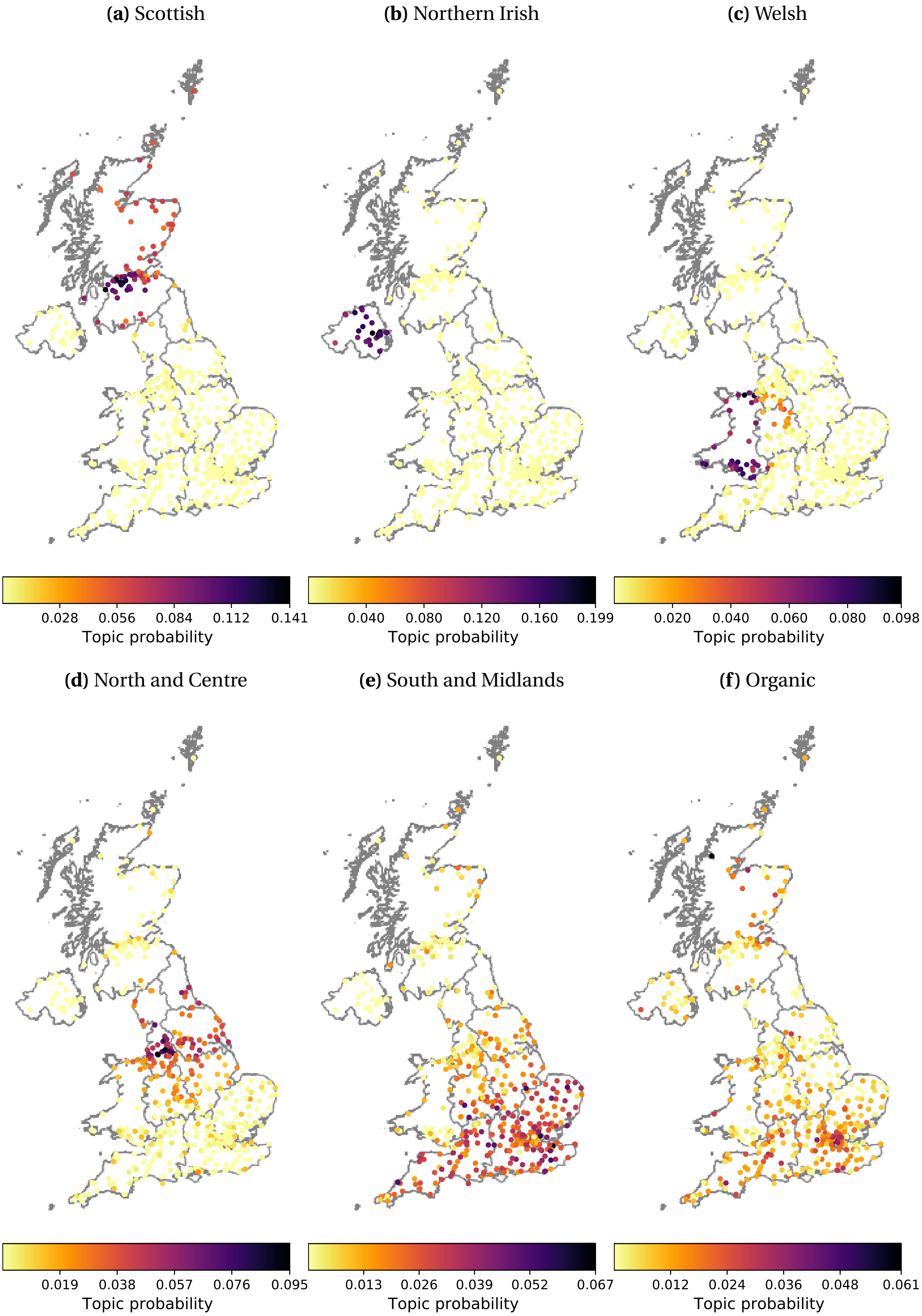


Figure 5.3: Topic probabilities $\theta_{i,k}$ of store i and clustered topic k. Purple and yellow points reflect the largest and smallest topic probabilities, respectively.

viously, STM discovers topics that are relevant within their store context. In comparison, LDA finds products that are frequently bought together across all transactions. Thus, a product combination that is only frequent in a few stores may not be shown against LDA topics. The ability to capture store-specific topics is key to our subsequent spatial modelling analysis.

We compare the 104 STM clustered topics (HC-STM-100) against the posterior summaries of the LDA model with 100 and 200 topics. The posterior summaries of LDA were obtained using the same training data and following the clustering methodology in [134]. The posterior summary of LDA with 100 topics (HC-LDA-100) gathered 96 clustered topics and the posterior summary of LDA with 200 topics (HC-LDA-200) gathered 198 clustered topics as shown in Table 4.1.

Figure 5.4a shows the cosine similarity between (HC-STM-100) 104 clustered topics and (HC-LDA-100) 96 clustered topics. Clustered topics are ordered to visualise their high similarity in the diagonal. As observed, the majority of clustered topics are identified in both models, STM and LDA, with high cosine similarity > 0.7. Analysing the distributions of the maximum similarity of each topic regarding the topics of the other model, Figure 5.5a shows that 70% of the (HC-STM-100) clustered topics are found among HC-LDA-100 clustered topics; and 85% of the HC-LDA-100 clustered topics are found among the HC-STM-100 clustered topics. For instance, the Northern Irish topic is found in both models with high cosine similarity (0.97). As depicted in Figure 5.6a, Northern Ireland related products rank in the top 15 products in both topics. The Organic topic was also found among HC-LDA-100 clustered topics with high cosine similarity (0.95).

We also compared the 104 STM clustered topics against the 198 LDA clustered topics obtained from summarising LDA posterior samples of 200 topics. This comparison allows the identification of regional topics which were not inferred in LDA samples with 100 topics. For instance, the Scottish topic described in Figure 5.2b, is not found in the HC-LDA-100 subset, but it is found in the HC-LDA-200 subset with a cosine similarity of 0.83. As observed in Figure 5.4b, the

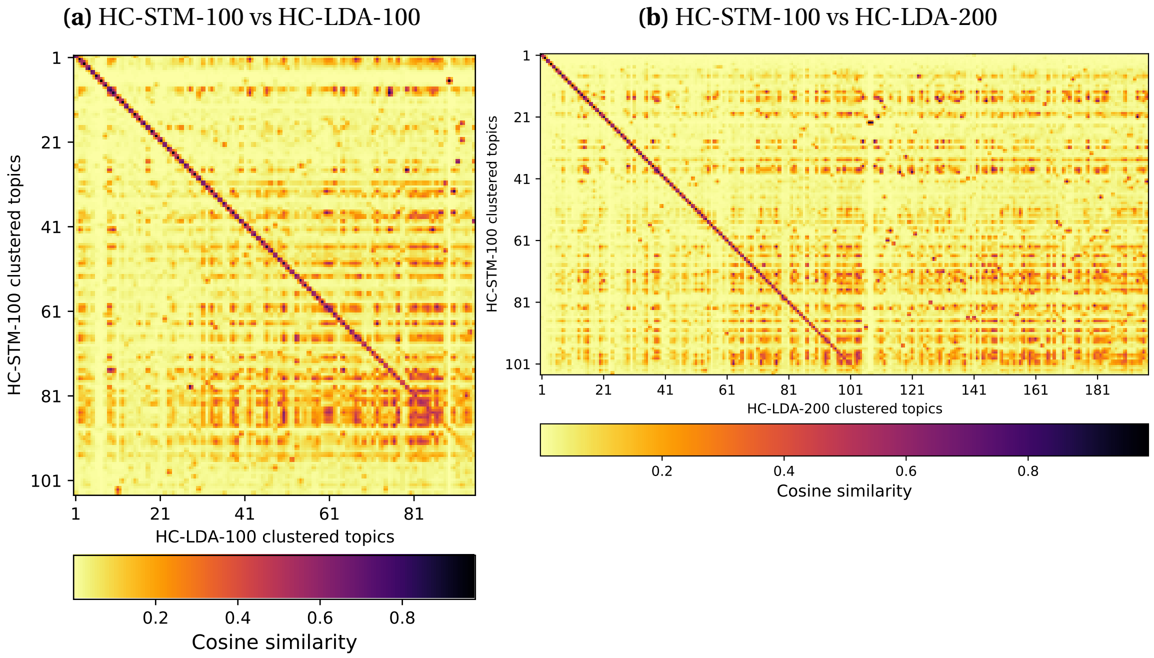
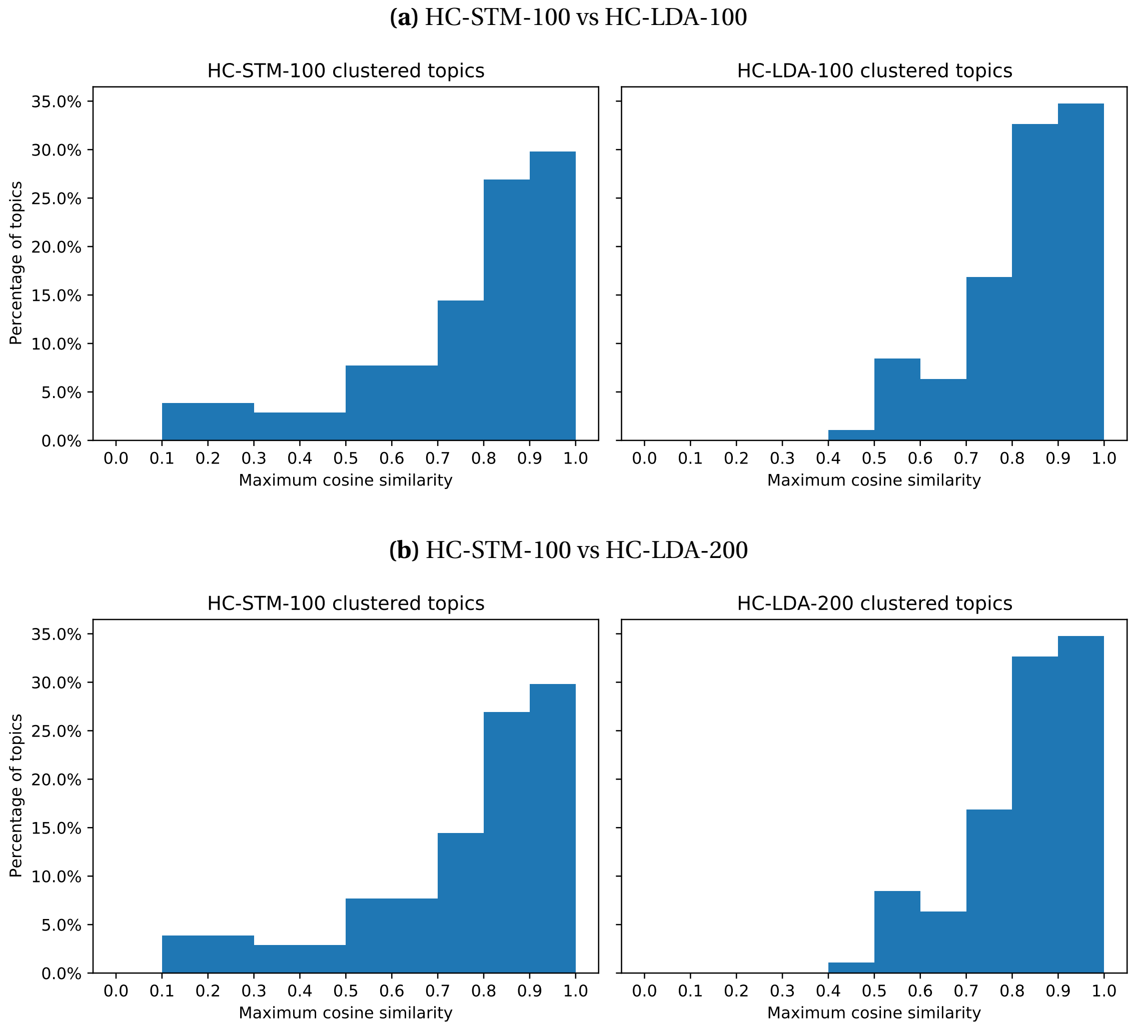


Figure 5.4: Cosine Similarity between clustered topics obtained from posterior summaries of STM with 100 topics and LDA with 100 and 200 topics. Topics have been aligned following a greedy algorithm that at each step searches and pairs topics (that have not been paired) with the highest cosine similarity.

majority of the 104 clustered topics are found among the (HC-LDA-200) 198 LDA clustered topics with high cosine similarity (> 0.7). However, Figure 5.5b shows that there are still some STM clustered topics that do not match with any of the LDA clustered topics with high similarity. For instance, the Welsh topic described in Figure 5.2c, is not found in either of the two subsets of LDA clustered topics. The Welsh topic and the closest clustered topic in HC-LDA-200 (with 0.67 cosine similarity) are listed in Figure 5.6b; as observed, few products are shared by the topics but Welsh products are not described in both topics. The North and Centre topic and the South and Midlands topic were not found among the HC-LDA-200 clustered topics either.

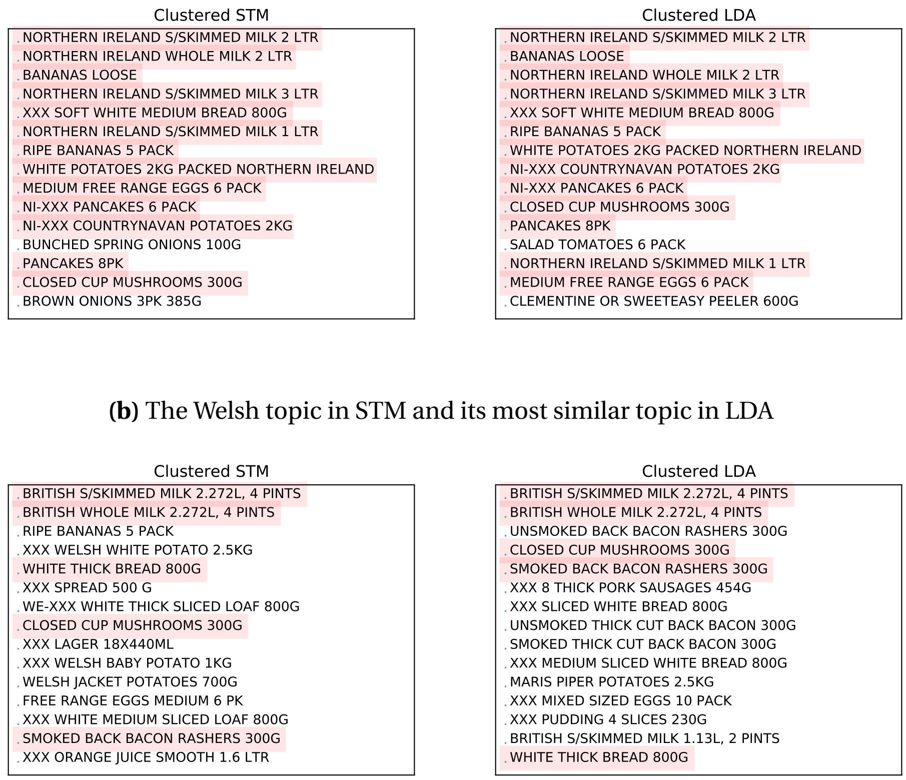
Of the six analysed topics, three topics were identified by STM and three regional topics were identified by both STM and LDA models. One of these topics was captured by a large LDA model. While large LDA models may capture more topics with spatial patterns, larger models are computationally expensive



(a) HC-STM-100 vs HC-LDA-100

(b) HC-STM-100 vs HC-LDA-200

Figure 5.5: Distributions of the maximum cosine distance obtained from each cosine similarity matrix in Figure 5.4. Figure 5.5a plots maximum cosine distances between clustered STM topics (HC-STM-100) against the posterior summary of LDA with 100 topics (HC-LDA-100) (left); and from HC-LDA-100 to HC-STM-100 (right). Figure 5.5b plots maximum cosine distances between HC-STM-100 against the posterior summary of LDA with 200 topics (HC-LDA-200) (left); and from HC-LDA-200 to HC-STM-100 (right).



(a) The Northern Irish topic in STM and LDA

(b) The Welsh topic in STM and its most similar topic in LDA

Clustered STM
NORTHERN IRELAND S/SKIMMED MILK 2 LTR
NORTHERN IRELAND WHOLE MILK 2 LTR
BANANAS LOOSE
NORTHERN IRELAND S/SKIMMED MILK 3 LTR
XXX SOFT WHITE MEDIUM BREAD 800G
NORTHERN IRELAND S/SKIMMED MILK 1 LTR
RIPE BANANAS 5 PACK
WHITE POTATOES 2KG PACKED NORTHERN IRELAND
MEDIUM FREE RANGE EGGS 6 PACK
NI-XXX PANCAKES 6 PACK
NI-XXX COUNTRYNAVAN POTATOES 2KG
BUNCHED SPRING ONIONS 100G
PANCAKES 8PK
CLOSED CUP MUSHROOMS 300G
BROWN ONIONS 3PK 385G
Clustered LDA
NORTHERN IRELAND S/SKIMMED MILK 2 LTR
BANANAS LOOSE
NORTHERN IRELAND WHOLE MILK 2 LTR
NORTHERN IRELAND S/SKIMMED MILK 3 LTR
XXX SOFT WHITE MEDIUM BREAD 800G
RIPE BANANAS 5 PACK
WHITE POTATOES 2KG PACKED NORTHERN IRELAND
NI-XXX COUNTRYNAVAN POTATOES 2KG
NI-XXX PANCAKES 6 PACK
CLOSED CUP MUSHROOMS 300G
PANCAKES 8PK
SALAD TOMATOES 6 PACK
NORTHERN IRELAND S/SKIMMED MILK 1 LTR
MEDIUM FREE RANGE EGGS 6 PACK
CLEMENTINE OR SWEETEASY PEELER 600G
Clustered STM
BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
BRITISH WHOLE MILK 2.272L, 4 PINTS
RIPE BANANAS 5 PACK
XXX WELSH WHITE POTATO 2.5KG
WHITE THICK BREAD 800G
XXX SPREAD 500 G
WE-XXX WHITE THICK SLICED LOAF 800G
CLOSED CUP MUSHROOMS 300G
XXX LAGER 18X440ML
XXX WELSH BABY POTATO 1KG
WELSH JACKET POTATOES 700G
FREE RANGE EGGS MEDIUM 6 PK
XXX WHITE MEDIUM SLICED LOAF 800G
SMOKED BACK BACON RASHERS 300G
XXX ORANGE JUICE SMOOTH 1.6 LTR
Clustered LDA
BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
BRITISH WHOLE MILK 2.272L, 4 PINTS
UNSMOKED BACK BACON RASHERS 300G
CLOSED CUP MUSHROOMS 300G
SMOKED BACK BACON RASHERS 300G
XXX 8 THICK PORK SAUSAGES 454G
XXX SLICED WHITE BREAD 800G
UNSMOKED THICK CUT BACK BACON 300G
SMOKED THICK CUT BACK BACON 300G
XXX MEDIUM SLICED WHITE BREAD 800G
MARIS PIPER POTATOES 2.5KG
XXX MIXED SIZED EGGS 10 PACK
XXX PUDDING 4 SLICES 230G
BRITISH S/SKIMMED MILK 1.13L, 2 PINTS
WHITE THICK BREAD 800G

Figure 5.6: Comparison of topics identified in STM and LDA posterior samples. Highlighted products appear in both topics. While the Northern Irish topic is clearly identified by both models, the Welsh topic is only found by the STM model.

and also tend to retrieve less distinctive topics as shown in [134]. Thus, STM has advantages over LDA when identifying regional topics.

5.4 Spatial topic prevalence

Mapping topics by store location aids the analysis of topics and shopping behaviours, but it does not quantify regional topic prevalence. In this section, we implement linear Gaussian process regression to identify and characterise regional topics.

5.4.1 Linear Gaussian process regression

According to Tobler's first law of geography [141], everything is related to everything else, but near things are more related than distant things. Thus, we expect that nearby stores show similar shopping patterns and that some specific patterns may be limited to particular geographical areas. STM assumes that stores are independent of each other and does not take into account store location or proximity; although such a model would be mathematically possible, it would be computationally intractable at the level of resolution of interest. Instead, we use the summarised posterior distributions of topics obtained from STM and take a spatial modelling approach to capture their geographical structure and regional behaviour.

We aim to model topic probabilities across stores in the UK by constructing a linear model with fixed effects associated with the constituent countries of the UK (Wales, Scotland, Northern Ireland) and the nine English regions, and imposing spatial dependency through a Gaussian process that captures residual spatial association. In this manner, we can quantify the significance of a topic to a region or constituent country. This administrative division was chosen assuming that each country and region would broadly show differences in customer behaviour. Analysis over other subdivisions is possible, but out of the scope of this chapter.

5.4.1.1 Model

A linear regression with a spatial process is defined as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{5.2}$$

where $\mathbf{Y}$ is the dependent variable, $\mathbf{X}$ is the matrix of p covariates associated with locations $\mathbf{s}_1, ..., \mathbf{s}_n$, $\boldsymbol{\beta}$ is a p-dimensional fixed effect, $\boldsymbol{\eta}$ is a spatial process which captures spatial residual, and $\boldsymbol{\varepsilon}$ is an independent process which models pure error, also known as the *nugget* effect [138].

Our dependent variable $\mathbf{Y}$ is the logit transformation of store-specific topic probabilities $[\widehat{\theta}_{\mathbf{s}_1,k}, \widehat{\theta}_{\mathbf{s}_2,k}, ..., \widehat{\theta}_{\mathbf{s}_n,k}]$, given by:

$$\mathbf{Y} = \text{logit}([\widehat{\theta}_{\mathbf{s}_1,k}, \widehat{\theta}_{\mathbf{s}_2,k}, ..., \widehat{\theta}_{\mathbf{s}_n,k}]), \tag{5.3}$$

where each $\widehat{\theta}_{\mathbf{s}_i,k}$ is the average probability over 30 posterior samples of the k^{th} topic at store location $\mathbf{s}_i$. Samples are obtained from Section 5.2.1. Topic probabilities for different topics are modelled independently.

The logit transformation not only avoids predicting nonsensical values (i.e., topic probabilities > 1 or < 0) but also aids the visualisation of topic probabilities that cannot be appreciated in the original scale. For instance, Figure 5.7 (left panel) highlights stores in the South West that do not seem to show a significant probability of the Welsh topic in Figure 5.3c.

The covariates are dummy variables responding to the constituent countries: 'North Ireland', 'Scotland', 'Wales'; and the English regions: 'North East', 'North West', 'Yorkshire and the Humber', 'East Midlands', 'West Midlands', 'South West', 'South East', and 'East Anglia', where 'London' is the reference category. For the purpose of this Chapter, we only use English regions and constituent countries of the UK, but other store-specific covariates could be added.

Distributionally, the errors $\varepsilon(\mathbf{s}_1),...,\varepsilon(\mathbf{s}_n)$ are assumed $i.i.d \sim N(0,\sigma^2)$ and the spatial process $\eta(\mathbf{s}_1),...,\eta(\mathbf{s}_n) \sim GP(0,C_{\boldsymbol{\eta}})$ is a zero-mean Gaussian process with positive definitive covariance matrix $C_{\boldsymbol{\eta}}$. Here, we use the positive definitive square exponential covariance function,

$$C_{\boldsymbol{\eta}}(\mathbf{s}_i,\mathbf{s}_j|\alpha,\rho) = \alpha^2 \exp\Big(-\frac{\mathrm{dist}(\mathbf{s}_i,\mathbf{s}_j)^2}{2\rho^2}\Big), \tag{5.4}$$

where parameters α and ρ control the amplitude and length-scale of the spatial dependence, respectively. $\mathrm{dist}(\mathbf{s}_i,\mathbf{s}_j)$ is a measure of distance between locations.

Spatial distance between stores is calculated by, firstly finding the latitude-longitude coordinates associated with the store's postcode, secondly computing the distance between pair of coordinates using the Haversine formula [142]. The Haversine formula provides accurate approximations of distance for locations over large areas. Postcode coordinates are queried from the postcode lookup table from the Office for National Statistics. Spatial distance is measured in kilometres.

The distribution of $\mathbf{Y}$ can be written as:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \Sigma), \tag{5.5}$$

where $\Sigma = C_{\boldsymbol{\eta}}(\cdot|\alpha, \rho) + \sigma^2 I$.

We complete the hierarchical model assuming weakly informative priors: $\sigma^2 \sim \text{half}N(0,1)$; $\beta \sim N(0,10)$; $\alpha \sim N(0,2)$; and $\rho \sim IG(2,50)$.

5.4.1.2 Methods

Linear Gaussian process regression specified in equation 5.2 is fitted using Stan [143]. Stan is a state-of-the-art platform for statistical modelling and high-performance statistical computation. Stan facilitates Bayesian inference by gradient-based sampling techniques such as Hamiltonian Monte Carlo methods [144] and variational inference [145]. In our study, the inference is computed by the default Stan algorithm No-U-Turn Sampler (NUTS) [146]. NUTS is an extension of the Hamiltonian Monte Carlo (HMC) algorithm that effectively explores the parameter space by avoiding retaking previously sampling paths in a U-turn style.

We obtain posterior samples from linear Gaussian process regression with 2 chains, 2,000 total iterations, 1,000 burn-in iterations, and a thin of five iterations. Results show satisfactory convergence with scale factor reduction $\widehat{R} = 0.998$.

5.4.1.3 Predictions

Predicted topic probabilities $\mathbf{Y}^\star = [Y^\star(\mathbf{s}_1), ..., Y^\star(\mathbf{s}_n)]$ at new locations $\mathbf{s}_1^\star, ..., \mathbf{s}_n^\star$ are distributed as:

$$\mathbf{Y}^\star|\mathbf{Y}, \boldsymbol{\beta}, \Theta, \mathbf{X}^\star, \mathbf{X} \sim N(\mathbf{X}^\star\boldsymbol{\beta} + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}), \tag{5.6}$$

where $\mathbf{X}^\star$ is the matrix of p covariates at the new locations. Here, Σ_{11} is the covariance matrix of $\mathbf{s}_1, ..., \mathbf{s}_n$ locations, $\Sigma_{12} = \Sigma_{21}$ the covariance matrix between $\mathbf{s}_1, ..., \mathbf{s}_n$ and $\mathbf{s}_1^\star, ..., \mathbf{s}_n^\star$, and Σ_{22}, covariance matrix of $\mathbf{s}_1^\star, ..., \mathbf{s}_n^\star$.

Note that expected topic probabilities $E(\mathbf{Y}^\star)$ are computed by two quantities. The first quantity is obtained by multiplying the covariate matrix by the fixed effects as in multiple linear regression. The second quantity pulls the expected value of the topic probability at a new store towards the topic probabilities of the nearby stores if spatial dependence is significant.

5.4.2 Prevalence of regional behaviours in the United Kingdom

Table 5.1 shows posterior summaries of the linear Gaussian process regression. The intercept can be interpreted as how likely (in logit scale) a topic is at a store in London and vice versa. Positive average coefficients indicate that the topic is more likely in those regions than in London. Average coefficients that are highlighted in red correspond to non-zero 95% credible intervals with $0 >$ upper bound, and bold average coefficients correspond to non-zero 95% credible intervals with $0 <$ lower bound.

Unsurprisingly, the Scottish, Northern Irish and Welsh topics show positive average coefficients with non-zero credibility intervals for the respective constituent countries. This indicates that their topic probability increases significantly for stores in Scotland, North Ireland and Wales, respectively.

Table 5.1: Regression parameters for regional topics Red/**bold** mean estimates for coefficients with non-zero credibility intervals that decrease/increase the topic probability, respectively.

	Northern Irish		Scottish		Welsh		English - North and Centre		English - South and Midlands		Organic	
Parameter	Avg.	SE	Avg.	SE	Avg.	SE	Avg.	SE	Avg.	SE	Avg.	SE
Intercept	-10.4	0.02	-9.52	0.03	-8.9	0.04	-6.34	0.04	-4.42	0.02	-4.62	0.05
Northern Ireland	**8.67**	0.03	-0.72	0.04	-1.44	0.07	-4.11	0.05	-5.77	0.03	-1.25	0.06
Scotland	0.19	0.02	**6.84**	0.04	-1.12	0.05	-1.93	0.04	-1.82	0.03	-1.34	0.06
Wales	-0.4	0.03	-0.57	0.03	**5.63**	0.07	0.39	0.04	-2.27	0.03	-1.27	0.06
North West	0.15	0.03	**1.54**	0.04	0.1	0.08	**3.3**	0.05	-0.99	0.03	-1.91	0.06
North East	-0.86	0.04	**3.27**	0.05	-0.33	0.08	**3.05**	0.06	-1.25	0.04	-2.5	0.07
Yorkshire	0.04	0.03	1.08	0.04	-0.43	0.05	**2.98**	0.06	-0.43	0.03	-1.68	0.05
West Midlands	-0.15	0.02	-0.24	0.03	**1.89**	0.07	**1.95**	0.05	0.26	0.03	-1.01	0.05
East Midlands	-0.47	0.03	0.68	0.04	0.67	0.05	1.45	0.05	0.31	0.06	-1.47	0.05
East Anglia	-0.27	0.02	-0.28	0.03	-0.38	0.05	-0.31	0.04	**0.99**	0.02	-1.03	0.05
South East	-0.21	0.2	0.56	0.03	-0.25	0.04	-1.07	0.04	0.66	0.02	-0.51	0.05
South West	-0.26	0.2	-0.1	0.03	**1.26**	0.05	-0.64	0.04	0.71	0.03	-0.02	0.05
Length-scale ρ	63.85	5.95	92.07	19.95	55.31	1.32	51.32	15.53	50.23	3.84	34.67	3.13
Amplitude α	0.13	0.01	0.3	0.03	1.04	0.01	0.74	0.02	0.23	0.01	0.86	0.02
σ	0.78	0.01	1.38	0.01	1.43	0.01	1.37	0.01	1.15	0.01	1.58	0.01

Interestingly, Wales's neighbouring regions, West Midlands and South West show positive average coefficients with non-zero credibility intervals. As shown in Figure 5.7 (left panel), store-specific probabilities of the Welsh topic (in logit scale) are large for stores in Wales and some stores near Wales. The covariate coefficients of the Welsh topic that correspond to West Midlands and South West produce larger estimates than further regions as observed in Figure 5.7 (central panel). These estimates are the same for all stores within the same region, which does not fit the observed spatial pattern (far-from-Wales stores show lower logit probabilities than close-to-Wales stores). Then, the Gaussian process (GP) captures spatial residuals. As illustrated in Figure 5.7 (right panel), red dots indicate where the topic is more popular; this popularity decreases as stores locate further from the south of the North West region. On the other hand, blue dots indicate where the topic is less popular; and negative spatial estimates reduces the fixed effects estimate. Thus, the GP distinguishes the stores in the neighbouring regions that are close to Wales from the stores (in the same regions) that are at further distances.

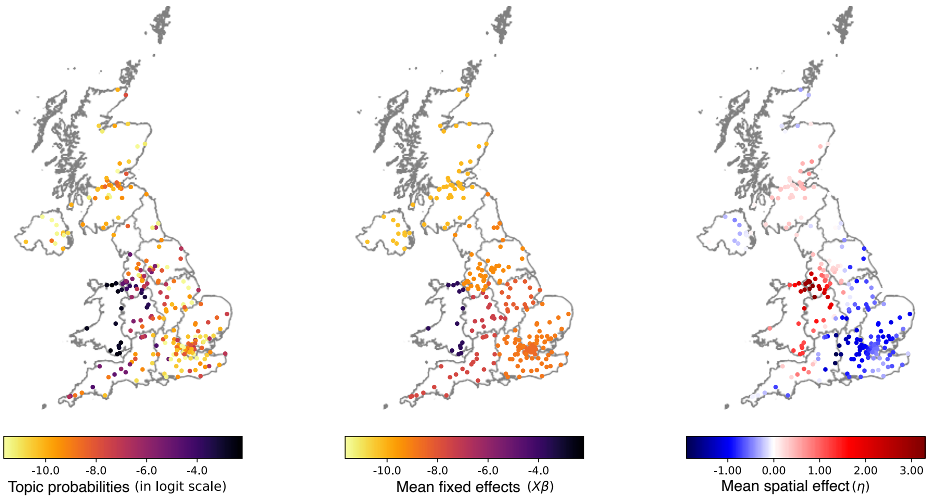

-10.0 -8.0 -6.0 -4.0
Topic probabilities (in logit scale)
-10.0 -8.0 -6.0 -4.0
Mean fixed effects $(X\beta)$
-1.00 0.00 1.00 2.00 3.00
Mean spatial effect (η)

Figure 5.7: Welsh topic: (left panel) observed topic probabilities in logit scale; (central panel) probability estimates (in logit scale) using only fixed effects; (right panel) spatial residuals captured by the Gaussian process.

Covariates of the GP regression for the Scottish topic also show significant positive coefficients for Scotland's neighbouring regions: North East and North West. As shown in Figure 5.8 (left panel), the probabilities in logit scale are larger at stores in Scotland and at few neighbouring stores; covariates for neighbouring regions produce slightly larger estimates than the covariates corresponding to other regions as observed in Figure 5.8 (central panel). A spatial pattern is clear in Figure 5.8 (right panel), where the gradation of colours goes from north to south. Note that spatial estimates for the furthest stores (in the South East, South West and East Anglia) would reduce the estimates obtained only with fixed effects (vice-versa, spatial estimates for stores in the centre of Scotland would augment the estimates obtained only with fixed effects); however, the scale of the spatial estimates is very small, indicating a reduced GP contribution.

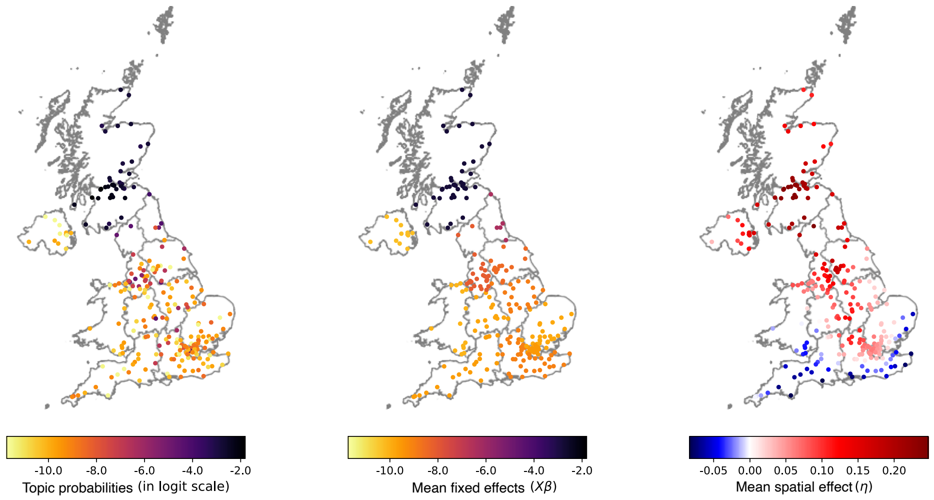

-10.0
-8.0
-6.0
-4.0
-2.0
Topic probabilities (in logit scale)
-10.0
-8.0
-6.0
-4.0
-2.0
Mean fixed effects $(X\beta)$
-0.05
0.00
0.05
0.10
0.15
0.20
Mean spatial effect (η)

Figure 5.8: Scottish topic: (left panel) observed topic probabilities in logit scale; (central panel) probability estimates (in logit scale) using only fixed effects; (right panel) spatial residuals captured by the Gaussian process.

In contrast, Figure 5.9 (left panel) illustrates large logit probabilities of the Northern Irish topic only at stores in Northern Ireland, showing no spatial variation across the rest of the UK. The fixed effects covariates provide good estimates and the spatial estimates are seldom and not significant as shown in Figure 5.9

(central and right panel).

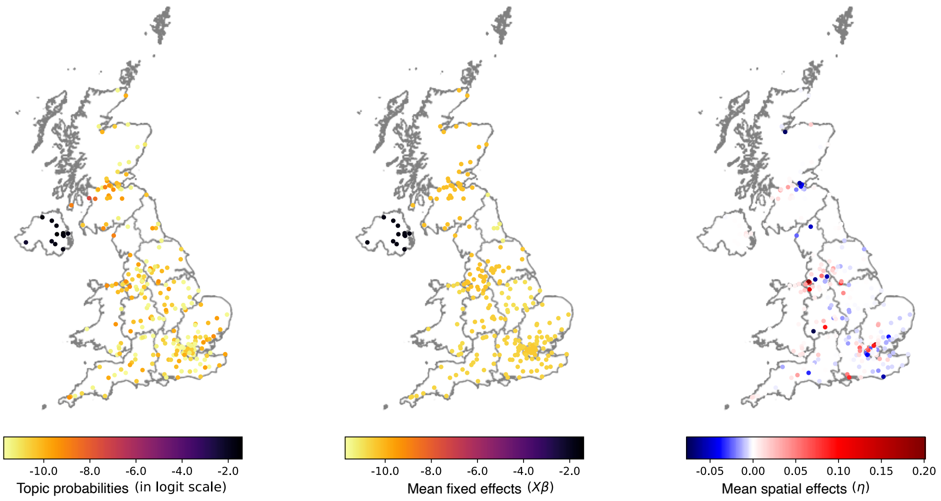

-10.0
-8.0
-6.0
-4.0
-2.0
Topic probabilities (in logit scale)
-10.0
-8.0
-6.0
-4.0
-2.0
Mean fixed effects $(X\beta)$
-0.05
0.00
0.05
0.10
0.15
0.20
Mean spatial effects (η)

Figure 5.9: Northern Irish topic: (left panel) observed topic probabilities in logit scale; (central panel) probability estimates (in logit scale) using only fixed effects; (right panel) spatial residuals captured by the Gaussian process.

The coefficients for the English-North and Centre topic clearly show that the topic is more likely in the North West, North East, Yorkshire and West Midlands and is less likely in Northern Ireland and Scotland as observed in Figure 5.10 (left panel). On the other hand, the coefficients for the English-South and Midlands show that on average the topic is more likely in the southern and central English regions as shown in Figure 5.11 (left panel); however, only the coefficient of East England has a non-zero 95% credibility interval. These two topics show opposite spatial patterns, the North and Centre topic shows spatial estimates that increase the probability of the topic in the North West and neighbouring regions and that decrease the probability of the topic in London and Scotland as shown in Figure 5.10 (right panel). On the contrary, the topic of South and Midlands presents spatial estimates that decrease the probability of the topic in the North West and neighbouring regions as observed in Figure 5.11 (right panel). However, the scale of the spatial estimates for the South and Midlands topic is very small, indicating a reduced GP contribution.

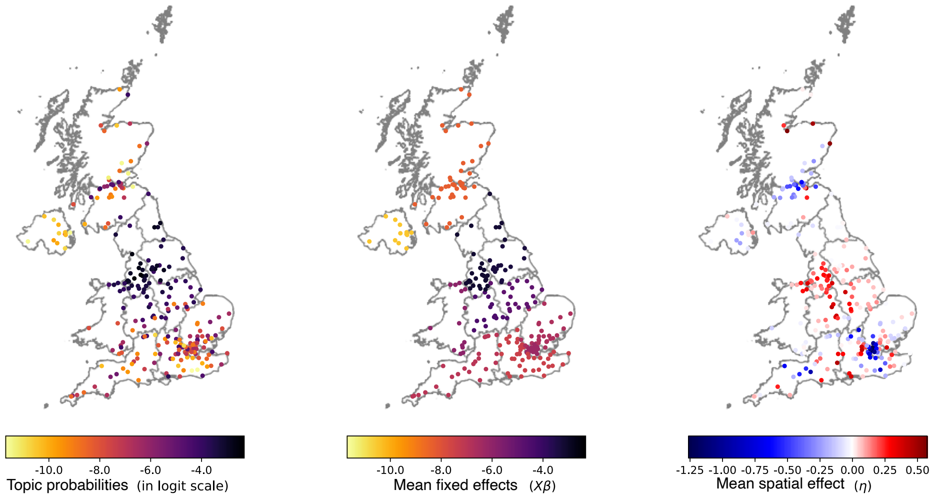


Figure 5.10: North and Centre topic: (left panel) observed topic probabilities in logit scale; (central panel) probability estimates (in logit scale) using only fixed effects; (right panel) spatial residuals captured by the Gaussian process.

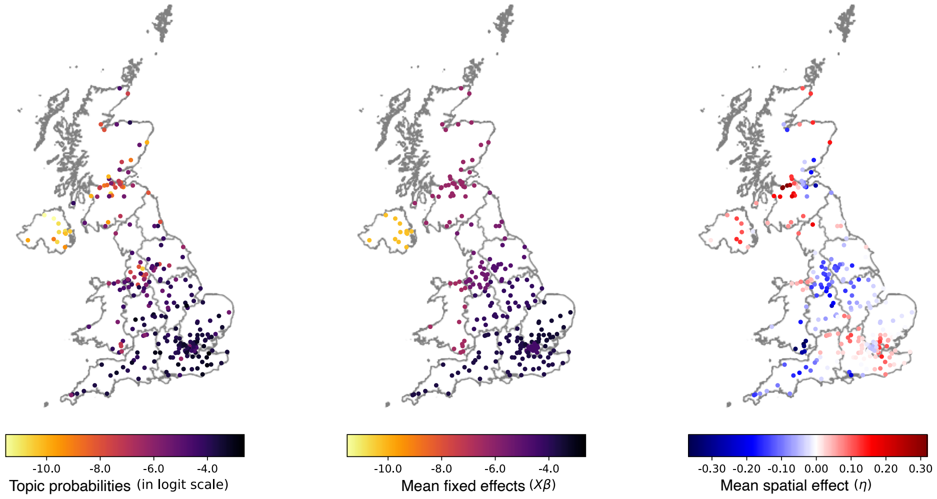


Figure 5.11: South and Midlands topic: (left panel) observed topic probabilities in logit scale; (central panel) probability estimates (in logit scale) using only fixed effects; (right panel) spatial residuals captured by the Gaussian process.

The Organic topic shows a different pattern, its average coefficients are negative; this indicates that the probability of the Organic topic is on average lower than the average topic probability in London. In other words, the Organic topic is more likely in London than in any other region or constituent country; however, the coefficients show 95% credible intervals containing zero, suggesting that the regional effect may not be significant. Observing Figure 5.12 (left panel), we can see that the Organic topic is popular but also unpopular across all the regions, except for London. The GP adjust the fixed effects estimates (central panel in Figure 5.12) to slightly increase the topic probability (in logit scale) at stores in London and to decrease the topic probability (in logit scale) at specific areas such as the centre and south of Scotland, the east of North Ireland, East Midlands to name a few, as shown in Figure 5.12 (right panel).

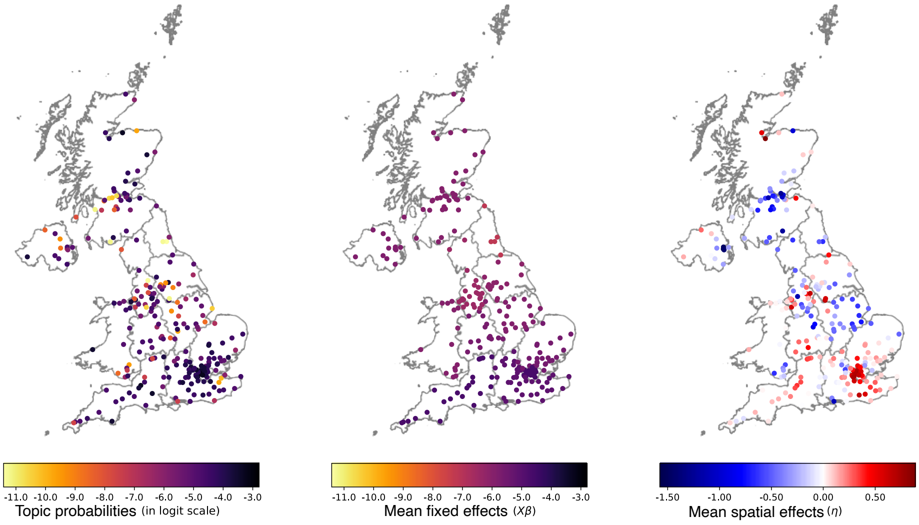
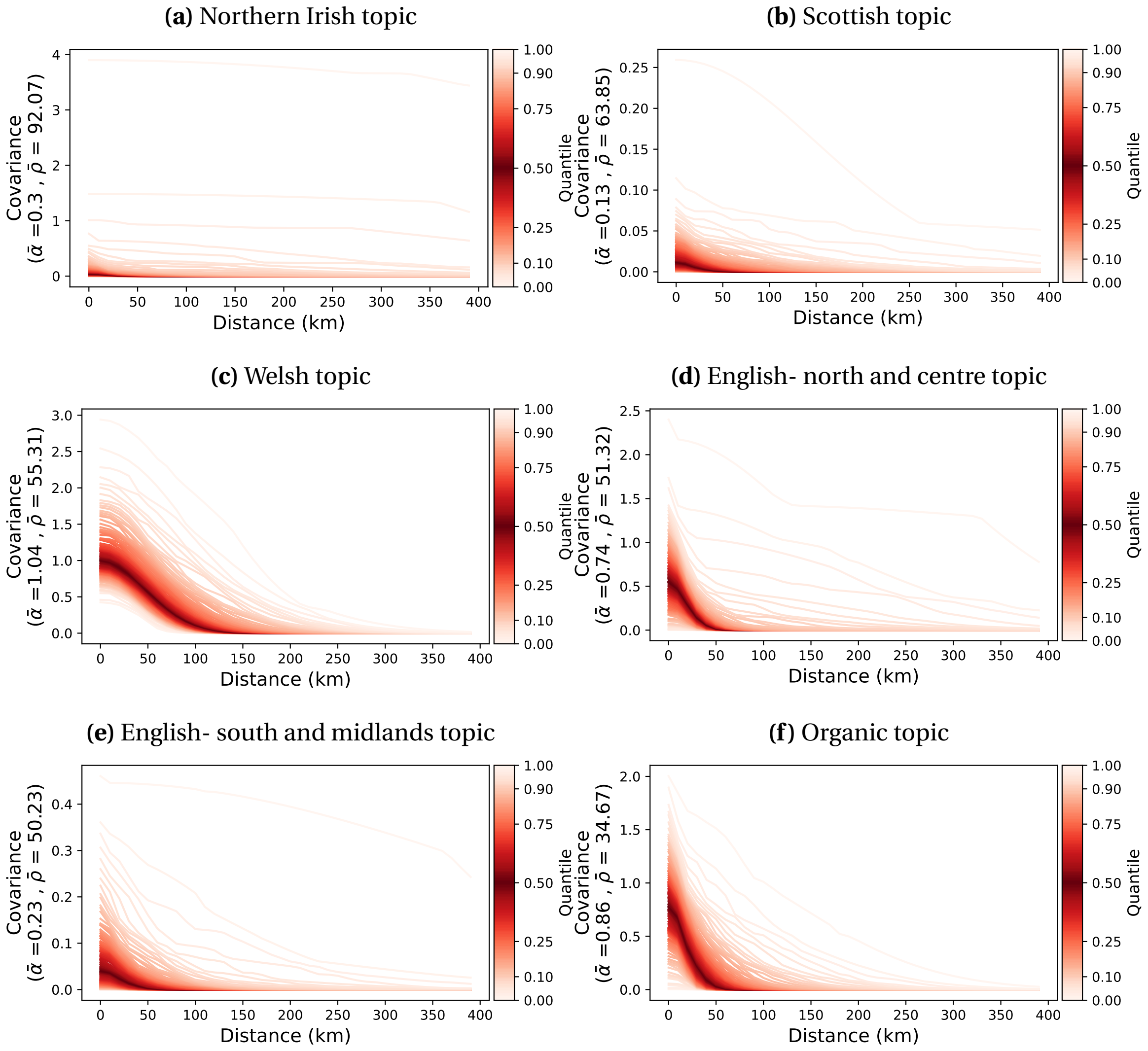


Figure 5.12: Organic Topic: (left panel) observed topic probabilities in logit scale; (central panel) probability estimates (in logit scale) using only fixed effects; (right panel) spatial residuals captured by the Gaussian process.

Analysing the posterior distribution of the covariance parameters: length-scale ρ and amplitude α in Figure 5.13, we observed that the Welsh topic shows a strong covariance function within stores that are not further than 100 km. The

North and Centre topic and the Organic topic also show strong covariance within stores that are not further than 50 km, and the South and Midlands topic shows a weak covariance within the same distance. On the other hand, the Northern Irish topic and the Scottish topic seem to show no significant covariance functions.


(a) Northern Irish topic

(b) Scottish topic

(c) Welsh topic

(d) English- north and centre topic

(e) English- south and midlands topic

(f) Organic topic

Figure 5.13: Posterior distribution of the covariate function of regional topics. Lines are computed with posterior samples of α and ρ.

5.4.3 Linear Gaussian process regression vs Linear regression

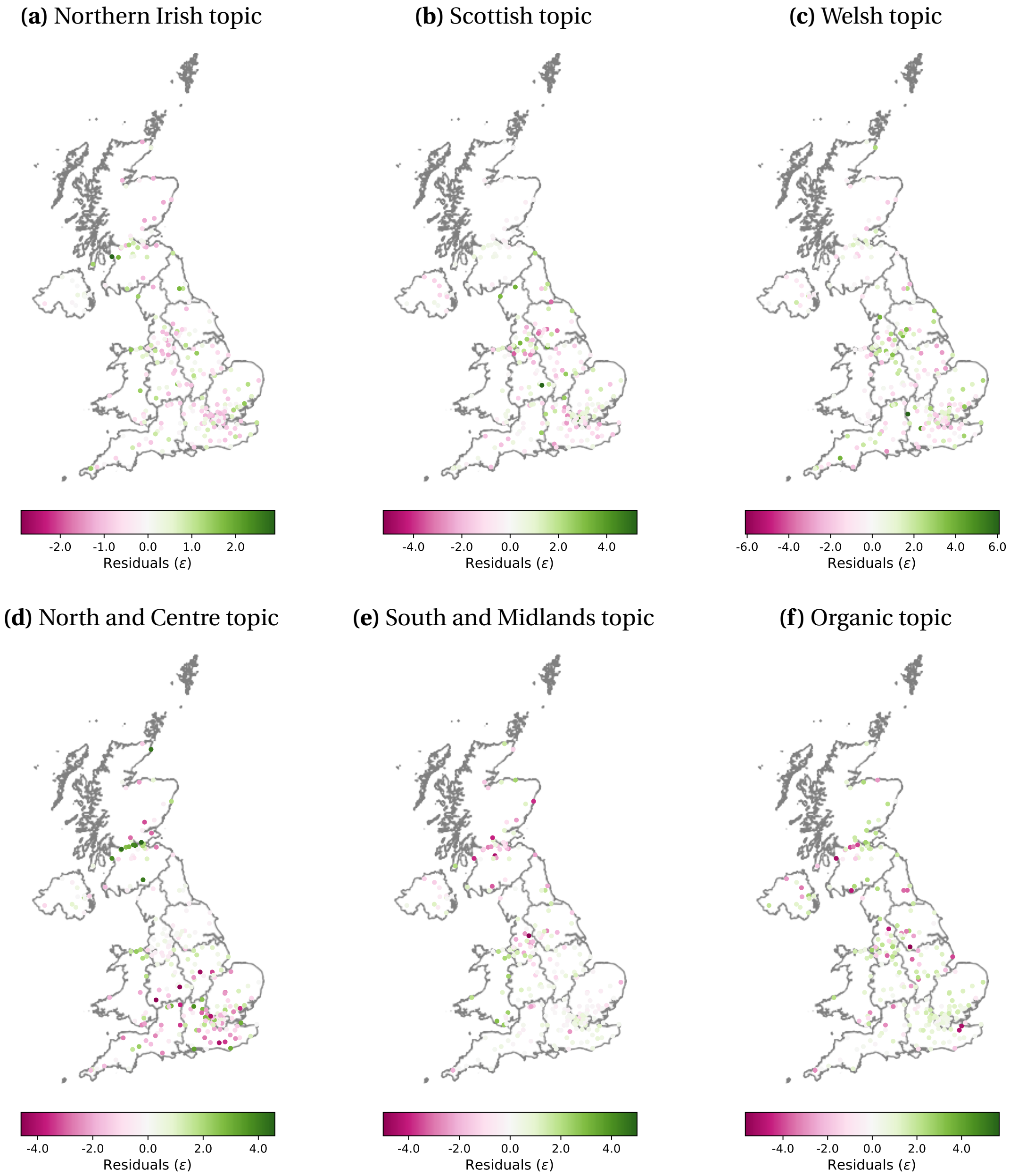
Here, we compare *mean squared error* and the log of the probability density on held-out data obtained from model topic prevalence using linear Gaussian process regression (LGPR) and the linear regression (LR). We will show that the former model retrieves more accurate estimates and better predictive likelihood by modelling residual spatial effect.

Table 5.2 shows that LGPR improves the prediction of topic probabilities of the Welsh, English-Northern and Centre, South and Midlands and Organic topics. The difference between the mean squared error of these topics is statistically significant at the 0.05 level, indicating that the GP provides significant model improvement. Similarly, the log predictive likelihood of the four aforementioned topics is significantly better at the 0.05 level. On the contrary, the LGPR doesn't show significantly improved predictions of the Scottish and Northern Irish topics. The difference of their mean squared errors is not statistically significant at the 0.05 level; however, the LGPR shows significantly better predictive log-likelihood at the 0.05 level.

Table 5.2: Comparison of the linear Gaussian process regression (LGPR) vs linear regression (LR). lppd: log posterior predictive density on test data. p-values are computed for the pointwise difference of the two methods at each observation in the test set.

	Northern Irish	Scottish	Welsh	English-North and Centre	English-South and Midlands	Organic
LR: MSE (SE)	0.64 (0.001)	2.16 (0.004)	3.18 (0.006)	3.36 (0.007)	1.65 (0.004)	3.39 (0.007)
LGPR: MSE (SE)	0.63 (0.001)	2.15 (0.004)	2.64 (0.005)	3.24 (0.004)	1.62 (0.003)	3.18 (0.006)
p-value	0.5877	0.1664	0.0000	0.0000	0.0000	0.0000
LR lppd (SE)	-298.3 (0.30)	-450.2 (0.25)	-499.1 (0.23)	-513.5 (0.31)	-418.8 (0.38)	-506 (0.26)
LGPR lppd (SE)	-296.5 (0.28)	-449.3 (0.26)	-476.9 (0.26)	-504.9 (0.43)	-412.6 (0.40)	-493.9 (0.27)
p-value	0.0000	0.0169	0.0000	0.0000	0.0000	0.0000

Examining residuals in Figure 5.14, we still observe spatial patterns that are not captured by the Gaussian process. For instance, concentrations of underestimated probabilities around North West in Figure 5.14a, around the centre of Scotland in Figure 5.14b, South West and East Anglia in Figure 5.14c; and overestimated probabilities around South East in Figure 5.14b. Further work could explore the Gaussian process with non-stationary covariance to capture local spatial patterns.



(a) Northern Irish topic

(b) Scottish topic

(c) Welsh topic

(d) North and Centre topic

(e) South and Midlands topic

(f) Organic topic

-2.0 -1.0 0.0 1.0 2.0
Residuals (ε)
-4.0 -2.0 0.0 2.0 4.0
Residuals (ε)
-6.0 -4.0 -2.0 0.0 2.0 4.0 6.0
Residuals (ε)
-4.0 -2.0 0.0 2.0 4.0
Residuals (ε)
-4.0 -2.0 0.0 2.0 4.0
Residuals (ε)
-4.0 -2.0 0.0 2.0 4.0
Residuals (ε)

Figure 5.14: Residuals of modelling the regional topics with linear Gaussian process regression. Pink/green dots denote over/under estimated topic probabilities; the less colourful the dots are, the smaller the residuals are.

5.5 Practical implications

Commercially speaking, analysing grocery transactions using segmented topic model aids retailers to understand the spatial difference in customer behaviours, avoiding topics that do not characterise specific stores. For instance, identifying product combinations that are relevant in specific areas aids the designing of marketing campaigns to target local demand, i.e., designing a promotion that utilises the ranked products in the 'North and Centre' topic at stores in northern and central English regions. By comparing store-specific topical mixtures, retailers could benchmark their stores and identify opportunities to offer products that could fulfil local demand, i.e., customising product assortments with highly ranked products of a locally relevant topic. The Gaussian process regression not only quantifies the importance of a topic over regions, but it might also help retailers to plan the product assortment of a new store, i.e., predicting topic probabilities at a new location given geographical features and distance to other stores.

Analysing grocery retail transactions through a segmented topic model provides new venues for sociological research. For instance, the topic probability of the 'organic' topic is larger in London than in other regions, this could be of interest to social scientists that aim to understand how economical and social factors influence food consumption. Sociologists could also find evidence of cultural identity in the identified regional topics as they show specific products such as 'potato scones' in Scotland or 'bottom oven muffin' in the North West region. While these products are widely recognised by the public, they are bought in combination with other less obvious products such as types of bread or types of sausages that could also express local identity. Social research could also analyse eating patterns at smaller statistical areas such as Middle layer Super Output Areas (MSOAs) and Lower layer Super Output Areas (LSOAs), investigate for demographical factors influence food consumption.

5.6 Summary

In this chapter, we showed an application of STM to grocery retail transactions and identified various customer needs, particularly, those that reflect regional demand. STM harnesses store structure, describes transactions and stores as topical mixtures and can identify regional topics that otherwise would be overlooked by the widely used topic model, the LDA. Summarising the posterior distribution of STM by aggregating multiple posterior samples and selecting topics of low uncertainty achieves better model generalisation, larger coherence and better credibility than topics from single posterior samples. Topic analysis, through linear Gaussian process regression, quantifies regional effects and captures spatial dependence through the squared exponential covariance function.

We identified six topics that reflect local supply and/or local demand; three topics demonstrate customer behaviours associated with the constituent countries of the UK, two topics show customer behaviours from the northern and southern English regions, and one topic is highly associated with London. Analysing store-specific topical mixtures and topics' product composition could help retailers to customise product assortments and design local promotions. Linear Gaussian process regression could aid analysts to plan product assortments for new stores. The application of STM to the analysis of grocery retail data provides new venues for sociological research.

Chapter 6

Finding Temporal Behaviours: the Sequential Segmented Topic Model

In the standard LDA model, transactions are assumed to be exchangeable. This ignores any temporal order between transactions. The lack of a temporal aspect in the analysis of retail data translates into not acknowledging that customer behaviours respond to temporal patterns due to seasonal product availability and to seasonal demand. We introduce a new topic model, the sequential segmented topic model, that accommodates temporal hierarchy over transactions while accounting for temporal sequence between time slices. In this manner, we identify customer behaviours with temporal patterns that are associated to festive, seasonal and periodic themes.

6.1 Introduction

Topic models, in particular LDA, have been applied to retail data and have proven their capacity to identify topics that reflect customers' shopping needs and to summarise transactions as mixtures of topics [11, 30, 29, 31, 97, 32, 134]. LDA assumes that transactions are exchangeable, and thereby, ignores the temporal aspects of grocery consumption, disregarding transactional metadata such as timestamps, i.e., transaction purchasing time. Shopping motivations respond to temporal patterns, e.g., customer behaviours in December may be different from the shopping patterns during summer. Thus, a more realistic representation of shopping motivations in grocery retail data needs to accommodate temporal metadata.

Various topic models exploit timestamps. The dynamic topic model (DTM) [68] extended LDA to let the topics and the prior distribution of topic distributions evolve across discretised units of time (month/week/year/etc.). The continuous dynamic topic model (cDTM) [69] extended the DTM in a continuous representation of time; so the only discretisation is the resolution at which timestamps are measured. In the retail context, DTM and cDTM offer methods to detect how products gain/lose importance within topics. For example, a 'fruit' topic shows summer fruits with high probabilities during the summer months and low probabilities during the winter months in which winter fruits become more probable. DTM and cDTM aim to capture the product dynamics in topic distributions by using a state-space model in which natural parameters of the multinomial distribution evolve with Gaussian noise. The Gaussian distribution is not conjugate to the multinomial distribution and an efficient collapsed Gibbs sampler cannot be derived; instead, these models use variational inference methods. Empirically, we have found that these models are challenging to fit; and that variational inference methods retrieve less interpretable topics than MCMC methods such as collapsed Gibbs sampling.

Topic models that exploit temporal metadata using MCMC methods are the dynamic mixture models (DMM) [46] and sequential latent Dirichlet allocation (SeqLDA) [67]. In both models, topics are static and topical mixtures are time-changing. DMM assumes that the semantic composition of a document depends on the semantic composition of the previous document. SeqLDA is a hierarchical topic model that interprets documents as collections of ordered segments, where segment-specific topical mixtures are chained with a first-order Markov assumption and documents are conditionally independent (exchangeable). Both DMM and SeqLDA can be applied to retail data to model sequential transactions when customer purchase history is available. In this case, transactions are modelled as sequential (non-exchangeable) purchases linked to the same customer. Thereby, past transactions may influence future transactions. However, these models are not suitable for our purposes since customer data are not available.

Since customer data are not available, time-ordered transactions cannot be linked to the same customer. Without customer dependency, transactions are exchangeable and a temporal sequence over transactions no longer applies. However, temporal metadata can still be exploited by grouping transactions using timestamps, creating a hierarchy between transactions and time slices (month/week/year/etc.). The Segmented Topic Model (STM) [44] was originally introduced to exploit document structure, i.e., documents are collections of paragraphs (segments) and paragraphs are interpreted as bags-of-words. In the retail context, STM can be applied to describe transactions and time slices as topical mixtures. However, STM does not exploit temporal sequence; and thereby, time slices are exchangeable.

Here, we propose the Sequential Segmented Topic Model (SeqSTM) that lays a temporal hierarchy over transactions (segments); transactions are exchangeable within a specific time slice (month/ week/ year/ etc.) and time slices follow a temporal sequence. In SeqSTM, transactions are characterised by topical mixtures that derive from their associated time-specific topical mixture, and time-specific topical mixtures are chained with a first-order Markov assumption through their prior distribution. For instance, topical mixtures that describe transactions purchased in December derive from December's topical mixture, which is influenced by November's topical mixture. Thus, we aim to identify grocery topics that respond to time-variant customer behaviours by assuming temporal sequence and temporal hierarchy over transactions.

Our work is inspired by the aforementioned dynamic mixture model (DMM) [46] and segmented topic model (STM) [44]. In DMM, topical mixtures are chained through a time-depending modified prior but no structure accommodates transactions under time slices. In STM, such a structure exists, but there is no temporal sequence between time slices. In response, SeqSTM extends STM to accommodate time dependence among time-specific topical mixtures using a time-depending modified prior. SeqSTM differs from SeqLDA, which assumes time dependency between transaction-specific topical mixtures.

We apply the SeqSTM, STM and LDA to data from a major grocery retailer in the UK and summarise the posterior distribution of the SeqSTM, STM and LDA by identifying thematic modes following the methodology in [134]. Thematic modes correspond to topics (and associated uncertainties) that consistently appear across multiple posterior samples, preventing the selection of nonsensical topics [108, 39, 40], and avoiding highly uncertain topics [37] while capturing posterior topic variability. We demonstrate that SeqSTM can identify temporal topics that STM or LDA fuse or overlook. We discuss the temporal topics that characterise grocery British consumption and illustrate customer behaviours that are driven by seasonal product availability and seasonal demand.

6.2 Sequential segmented topic model

We develop the sequential segmented topic model (SeqSTM), which is inspired by the STM and DMM. DMM's generative process is similar to LDA's but the transaction-specific topical mixtures are not exchangeable. In retail terms, the DMM postulates that the first transaction-specific topical mixtures θ_1 has a Dirichlet prior, and the subsequent transaction-specific topical mixtures θ_d depend on their previous transaction-specific topical mixtures θ_{d-1}. In SeqSTM, transactions are exchangeable within their respective time slice as in STM. However, time-specific topical mixtures respond to a temporal sequence and all months except the first one have a Dirichlet prior with a Markovian dependence on the previous time slice.

SeqSTM follows a generative process in which topic distributions, $[\phi_1, ...\phi_K]$, are drawn from a Dirichlet distribution governed by hyperparameters $\boldsymbol{\beta}$. The first time-specific topical mixture, θ_1, is drawn from Dirichlet distribution governed by hyperparameters $\boldsymbol{\alpha}$, and the subsequent time-specific topical mixture, θ_d, are drawn from Dirichlet distribution governed by hyperparameters $\alpha_0\theta_{d-1}$. Transaction-specific topical mixtures $\nu_{p,d}$ are drawn from a two parameter Poisson-Dirichlet Process (PDP) distributed with parameters a, b and θ_d. Then, for each item in a transaction, a topic assignment $z_{n,p,d}$ is sampled from $v_{p,d}$ and the item is sampled from the topic distribution $\phi_{v_{p,d}}$. Mathematically,

$$\begin{aligned} \phi_k &\sim \text{Dirichlet}(\boldsymbol{\beta}) \\ \theta_d &\sim \text{Dirichlet}(\boldsymbol{\psi}) \\ v_{p,d} &\sim \text{PDP}(a, b, \theta_d) \\ z_{n,p,d} &\sim \text{Multinomial}(\delta_{p,d}) \\ w_{n,p,d} &\sim \text{Multinomial}(\phi_{z_{n,p,d}}). \end{aligned} \tag{6.1}$$

where $\boldsymbol{\psi} = \boldsymbol{\alpha}$ for $d = 1$ and $\boldsymbol{\psi} = \alpha_0\theta_{d-1}$ for $d > 1$. $\alpha_0 = \sum_{k=1}^{K} \alpha_k$.

The graphical model of the SeqSTM is depicted in Figure 6.1 where only the products are observed. Note that only the first time-specific topical mixture derives from the Dirichlet prior and the following time-specific topical mixtures derive from their previous time-specific topical mixture. Transaction-specific topical mixtures derive from their corresponding time-specific topical mixture. Thus, transactions and time slices share the space of latent topics. SeqSTM also assumes that product order is disregarded ('bag-of-products'). Transactions are only exchangeable within their time slice.

The Dirichlet distribution is parametrised by a base measure and precision parameter [50]. Thus, θ_d is Dirichlet distributed with precision parameter, α_0, and base measure θ_{d-1}; and θ_1 is Dirichlet distributed with precision parameter, α_0, and a uniform base measure. Then, we re-express Equation 3.15 to include time-dependent priors.

$$\begin{aligned} &p(\mathbf{z}, \mathbf{w}, \mathbf{t} | \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) = \\ &\prod_d \frac{\text{Beta}_K(\boldsymbol{\psi} + \sum_p \mathbf{t}_{p,d})}{\text{Beta}_K(\boldsymbol{\psi})} \prod_{p,d} \frac{(b|a)_{\sum_k t_{p,d,k}}}{(b)_{N_{p,d}}} \prod_{p,d,k} S^{N_{k|p,d}}_{t_{p,d,k},a} \prod_k \frac{\text{Beta}_V(\boldsymbol{\beta} + \mathbf{N}_k)}{\text{Beta}_V(\boldsymbol{\beta})}, \end{aligned} \tag{6.2}$$

where $\boldsymbol{\psi} = \boldsymbol{\alpha}$ for $d = 1$ or $\boldsymbol{\psi} = \alpha_0\theta_{d-1}$ for $d > 1$ and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. $t_{p,d,k}$ is the table count for transaction p, time period d and topic k. $\text{Beta}_K(\boldsymbol{\alpha})$ is K dimensional

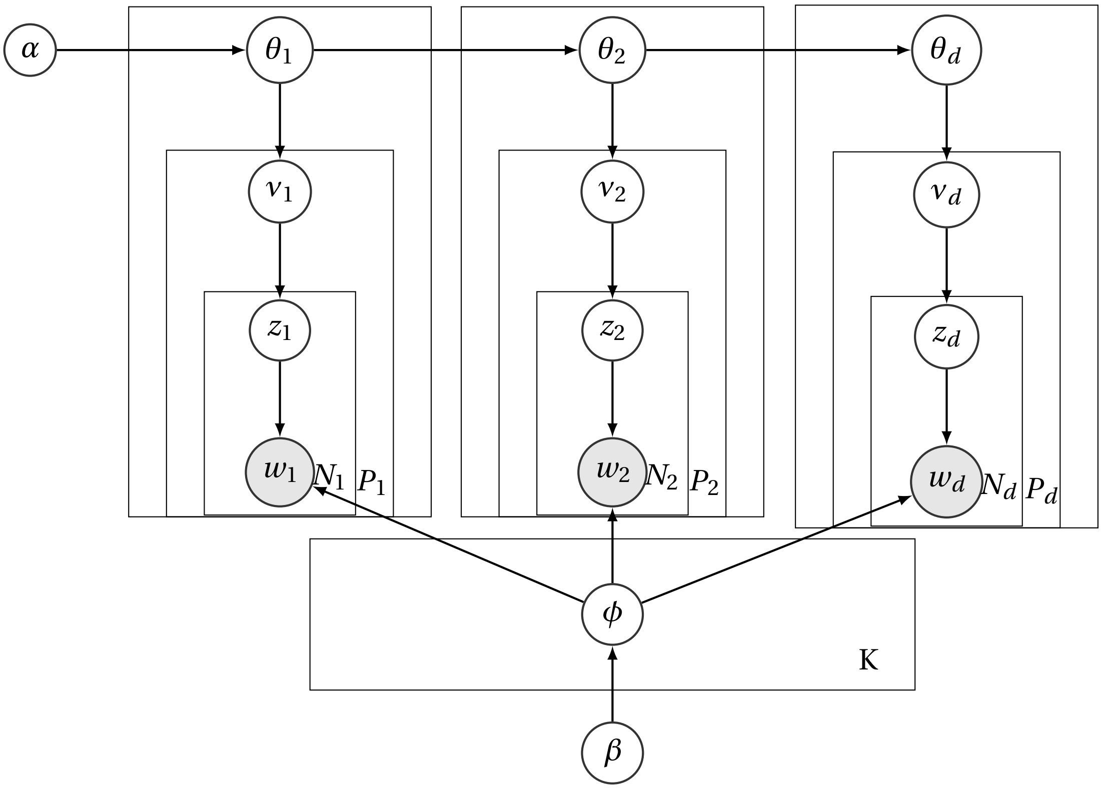


Figure 6.1: SeqSTM graphical model for d periods. Nodes denote random variables and edges denote dependencies. Unshaded node denote hidden random variables and shaded nodes denote observed random variables. Plates denote replication. The hidden variables are z topic assignments, θ period-specific topical mixtures, v transaction-specific topical mixtures, ϕ topic distributions, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ Dirichlet hyperparameters. K number of topics, P number of transactions, and N number of products.

beta function that normalises the Dirichlet distribution defined in Equation 2.2; $\mathbf{t}_{p,d}$ is a table count vector (i.e. $t_{p,d,1},...,t_{p,d,K}$); $(x|y)_N$ denotes the Pochhammer symbol defined in Equation 2.16; $N_{p,d}$ size of transaction p in store d; $S^N_{M,a}$ is a generalised Stirling number defined in Equation 2.17; $N_{k|p,d}$ number of topic assignments of topic k in transaction p in period d. $\text{Beta}_V(\boldsymbol{\beta})$ is V dimensional beta function that normalises the Dirichlet distribution; $\mathbf{N}_k$ is a vector of $N_{v|k}$, which is the number of products of type v assigned to topic k.

After a burn-in period, states of the Markov chain are recorded with an appropriate lag to ensure low autocorrelation between samples. For a single sample s, topics ϕ, time-specific topical mixtures θ and the transaction-specific topical mixtures v are estimated by their conditional posterior means given by:

$$\widehat{\theta}^s_{1,k} = E(\theta^s_{1,k}|\mathbf{t}^s, \boldsymbol{\alpha}) = \frac{\alpha_k + \sum_p t^s_{p,1,k}}{\alpha_0 + \sum_{p,k} t^s_{p,1,k}}, \tag{6.3}$$

$$\widehat{\theta}^s_{d,k} = E(\theta^s_{d,k}|\mathbf{t}^s, \boldsymbol{\alpha}) = \frac{\psi_k + \sum_p t^s_{p,d,k}}{\psi_0 + \sum_{p,k} t^s_{p,d,k}}, \tag{6.4}$$

$$\widehat{v}_{p,d,k} = E(v^s_{p,d,k}|\mathbf{z}^s, \mathbf{t}^s, a, b) = \frac{N^s_{p,d,k} - a \times t^s_{p,d,k}}{b + N^s_{p,d}} + \theta_{d,k} \frac{\sum_k t^s_{p,d,k} \times a + b}{b + N^s_{p,d}}, \tag{6.5}$$

$$\widehat{\phi}_{k,v} = E(\phi^s_{k,v}|\mathbf{z}^s, \boldsymbol{\beta}) = \frac{\beta_v + N^s_{k,v}}{\beta_0 + N^s_k}, \tag{6.6}$$

where $\alpha_0 = \sum_k^K \alpha_k$, $\psi_d = \alpha_0 \theta_{d-1}$, $\psi_0 = \sum_k^K \alpha_0 \theta_{d-1,k} = \alpha_0$, and $\beta_0 = \sum_v^V \beta_v$.

6.2.1 Block Gibbs sampler for SeqSTM

The block Gibbs sampler for SeqSTM follows the block Gibbs algorithm described in 3.3.2. However, we need to modify Equation 3.20 to modify the Dirichlet prior using:

$$p(z_n = k, u_n = 1|\mathbf{z} - \{z_n\}, \mathbf{u} - \{u_n\}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) \propto$$
$$\frac{\alpha_0 \theta_{d-1,k} + t'_{d,k}}{\alpha_0 + t'_d} \frac{b + a t'_{p,d}}{b + N'_{p,d}} \frac{S^{N'_{p,d,k}+1}_{t'_{p,d,k}+1}}{S^{N'_{p,d,k}}_{t'_{p,d,k}}} \frac{t'_{p,d,k} + 1}{n'_{p,d,k} + 1} \frac{\beta_v + N'_{k,w_{p,d,n}}}{\beta_0 + N'_k}, \tag{6.7}$$

where θ is given by Equations 6.3 and 6.4.

6.3 Topic model applications for temporal analytics

In this section, we apply STM and SeqSTM to the same data set described in Section 5.2, aiming to identify grocery topics that reflect time-variant customer behaviours. We discretise time over months, assuming that temporal customer behaviour extends over calendar months. Month-specific topical mixtures are then represented as topical mixtures that share the same finite number of topics. Posterior summaries and single posterior samples of STM and SeqSTM are compared against posterior summaries and single posterior samples of the LDA from Chapter 4. In comparison to SeqSTM and STM, LDA does not directly provide time-specific topical mixtures. However, time-specific topical mixtures can be estimated by grouping and averaging topical mixtures according to their timestamps as in [26].

SeqSTM and STM are set with symmetric priors with Dirichlet hyperparameters $\alpha_0 = 10,000$ and $\beta_v = 0.01$. The Dirichlet precision α_0 is chosen empirically by assigning a significant value with respect to the number of active tables per time slice. $\beta_v = 0.01$ is commonly used in the literature [108, 111]. We empirically set the values of PDP hyperparameters $b = 5.0$ and $a = 0.5$ to aid convergence of Markov chains. LDA settings are the same as in Section 4.3. We explore topic models with 100 topics, assuming that 100 topics are a large enough model to capture customer behaviours in the data of our application.

We run four Markov chains of SeqSTM and STM with 100 topics; each Markov chain runs for 150,000 iterations with a burn-in of 100,000 iterations; samples were recorded every 10,000 iterations. The convergence of the aforementioned models is satisfactory as shown in Appendix D.1. Samples show little autocorrelation as shown in Appendix D.2. Markov chain settings of LDA are the same as in Chapter 4.3.

6.3.1 Posterior summary of SeqSTM and STM topic distributions

As mentioned in previous chapters, it is challenging to summarise the posterior distribution of a topic model on real-data applications. The posterior distribution is often highly multimodal, so Gibbs sampling methods usually cannot fully explore the entire posterior distribution. As such, we use the clustering methodology detailed in Section 4.3 to summarise the posterior distribution of SeqSTM and STM.

For each model, we form a bag of topics using 20 posterior samples obtained from the four Markov chains mentioned above (five samples per chain), forming a bag of 2,000 topics. Several subsets of clusters are formed of varying cosine distance from 0.05 to 0.95 with steps of 0.05 and minimum clusters size of five, 10 and 20.

6.3.2 Evaluation and selection of topic models

Each subset is evaluated on four aspects: generalisation or predictive power of a subset of topics, coherence of individual topics, the distinctiveness of a topic w.r.t. the other topics in the same posterior sample, and credibility of a topic w.r.t. the topics from other posterior samples. Topic coherence, distinctiveness and credibility are measured as described in Sections 4.2.2, 4.2.3 and 4.2.4. Model generalisation is measured by the perplexity of unseen transactions given topics, time-specific topical mixtures and PDP parameters:

$$\text{Perplexity} = -\frac{\log P(\mathbf{w}'_d|\Phi, \theta_d, a, b)}{N'}, \tag{6.8}$$

where $\mathbf{w}'_d$ is a set of products in a held-out transaction at store d, N' is the number of products in $\mathbf{w}'_d$, $\Phi = [\phi_1, \phi_2, \ldots, \phi_K]$ the set of inferred topics, θ_d is the time-specific topical mixtures associated to store d, a and b are the PDP parameters.

We observe in Figures 6.2, 6.3 similar patterns as in 4.7 from Section 4.4.3. Subsets of clusters formed with a minimum cluster size of 10 show greater coherence and credibility, and the subsets formed with a cosine distance threshold larger than 0.3 show better generalisation (in comparison to the average generalisation of the STM posterior samples). Subsets with a minimum cluster size of 10 show less distinctive clustered topics, which might result from filtering out distinctive but uncertain topics. Cosine distance threshold larger than 0.35 cosine distance does not significantly improve perplexity.

As shown in Figures 6.2 and 6.3, we observe that subsets of clusters formed with a least 10 members (which represent 50% of the samples) and a cosine distance threshold ≥ 0.35 show the greatest coherence, credibility and generalisation, concurring with [134]. Based on these results, we summarise the posterior distribution of SeqSTM with 98 clustered topics, STM with 97 clustered topics and LDA with 96 clustered topics using a cosine distance threshold of 0.35 and a minimum cluster size of 10. The performance of the topic models and subsets of clustered topics is presented in Table 6.1.

As observed in Table 6.1, the 3 subsets of clustered topics (HC-LDA-100, HC-

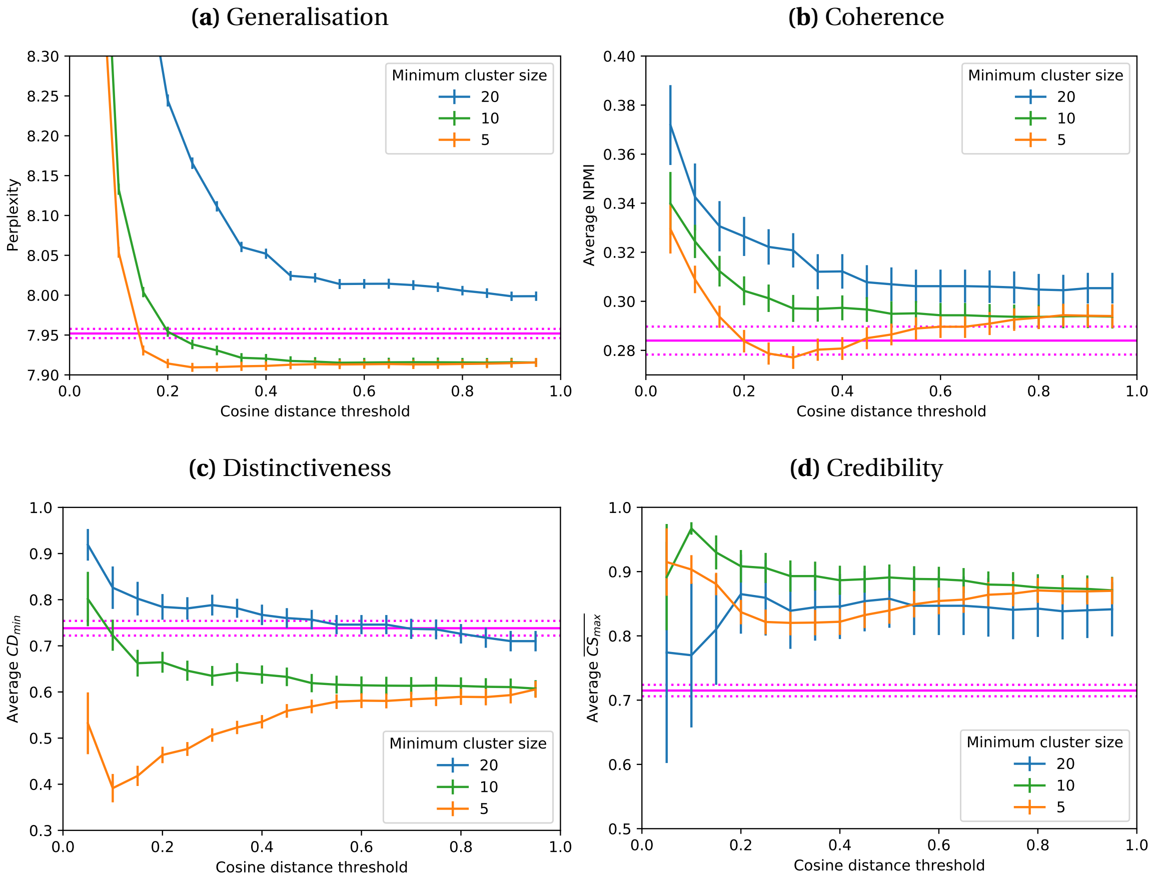
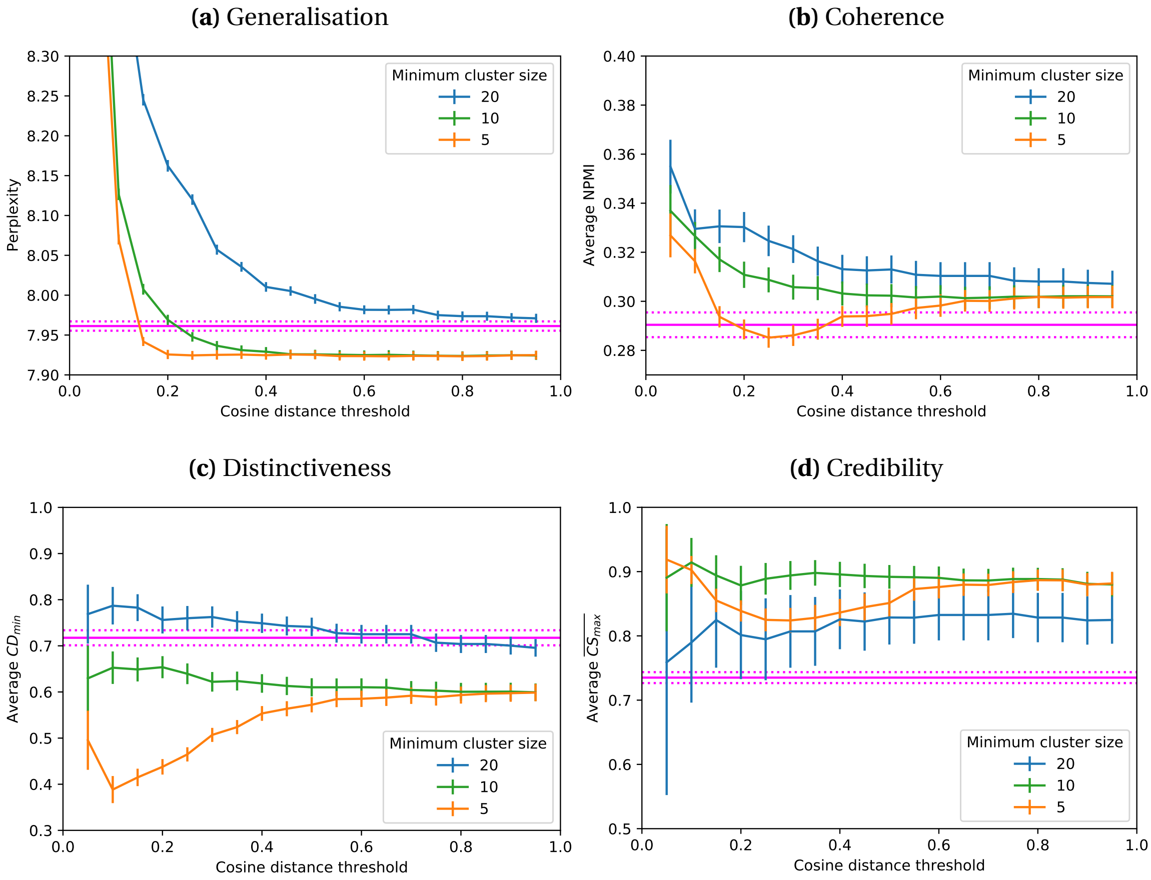


Figure 6.2: Evaluation of subsets of SeqSTM clustered topics. Subsets are formed with combinations of minimum cluster size and cosine distance thresholds. Horizontal and dotted lines show the average measures and ± standard error of the SeqSTM posterior samples.

STM-100, and HC-SeqSTM-100) show better performance in generalisation, coherence and credibility than the non-clustered topic models; and the three clustered models have the same level of credibility. On the other hand, LDA, STM and SeqSTM present larger distinctiveness. [134] notes that that posterior samples of topic models may include topics with some degree of similarity and highly distinctive but non-recurrent topics. Thereby, the distinctiveness of posterior samples tends to be larger than those of subsets of clustered topics since the latter excludes highly distinctive but non-recurrent topics.

The subset of clustered SeqSTM topics (HC-SeqSTM-100) shows the lowest perplexity and the greatest credibility; the subset of LDA clustered topics (HC-LDA-100) shows greater coherence, and the SeqSTM topics present the largest



(a) Generalisation

(b) Coherence

(c) Distinctiveness

(d) Credibility

Figure 6.3: Evaluation of subsets of STM clustered topics. Subsets are formed with combinations of minimum cluster size and cosine distance thresholds. Horizontal and dotted lines show the average measures and ± standard error of the STM posterior samples.

distinctiveness. As we will show in the following section, through exploiting temporal structure (i.e., transactions grouped by months) and temporal sequence (i.e., monthly dependent priors), SeqSTM identifies temporal topics that are overlooked by LDA. We further explore and interpret SeqSTM clustered topics and their monthly-specific topical mixtures.

6.4 Temporal British customer behaviours in grocery retail

In the previous section, we obtained a subset of clustered topics, which are used to recompute month-specific topical mixtures. We then rerun SeqSTM using the

Table 6.1: Generalisation, coherence, distinctiveness and stability metrics of LDA, STM and SeqSTM samples with 100 topics and subsets of LDA, STM and SeqSTM clustered topics (HC-LDA-100, HC-STM-100 and HC-SeqSTM-100), which are obtained from clustering the aforementioned topic models with 100 topics

.

Model	Topics	Generalisation	Coherence	Distinctiveness	Credibility
		Perplexity	NPMI	CD_{min}	$\overline{\text{CS}_{\text{max}}}$
		Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)
LDA-100	100	8.131 (0.003)	0.319 (0.006)	0.674 (0.016)	0.716 (0.009)
HC-LDA-100	96	8.076 (0.006)	**0.333 (0.005)**	0.565 (0.021)	**0.890 (0.010)**
STM-100	100	7.961 (0.002)	0.290 (0.005)	0.717 (0.016)	0.735 (0.008)
HC-STM-100	97	7.931 (0.002)	0.305 (0.004)	0.623 (0.020)	**0.898 (0.010)**
SeqSTM-100	100	7.951 (0.002)	0.284 (0.005)	**0.738 (0.016)**	0.715 (0.009)
HC- SeqSTM-100	98	**7.921 (0.003)**	0.296 (0.005)	0.642 (0.020)	**0.893 (0.011)**

identified 98 clustered topics (we do not recompute topic distributions) to obtain posterior samples of the time-specific topical mixtures. We run the block Gibbs sampler for 10,000 iterations with a burn-in period of 500 iterations and record samples every 500 iterations. MCMC trace plots are shown in Appendix D.3 where the convergence is satisfactory. Samples are recorded every 500 iterations to ensure little autocorrelation as shown in Appendix D.4. Month-specific topical mixtures are then obtained by averaging the posterior samples.

Monthly topic probabilities are transformed to topic ratios with respect to the monthly average topic probability. Figure 6.4 shows the top 20 topics with the highest topic proportions, identifying the topics with the largest monthly variations. The first 9 topics indicate the strongest temporal patterns that illustrate few months with large topic proportions (at least twice their monthly average topic probability). For instance, the probability of the Christmas topic in December is 4.5 times larger than the monthly average Christmas topic probability. The largest topic proportions of the following 11 clustered topics are at least 1.4 times their monthly average.

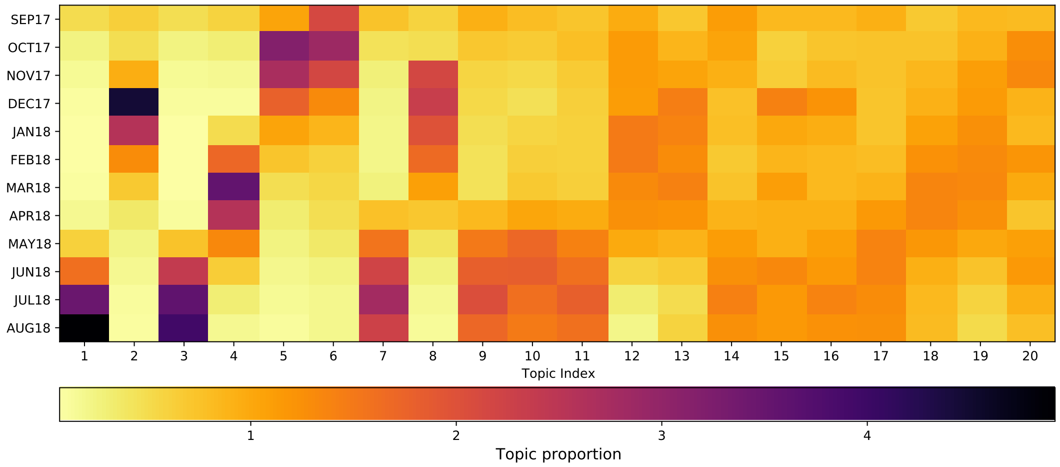


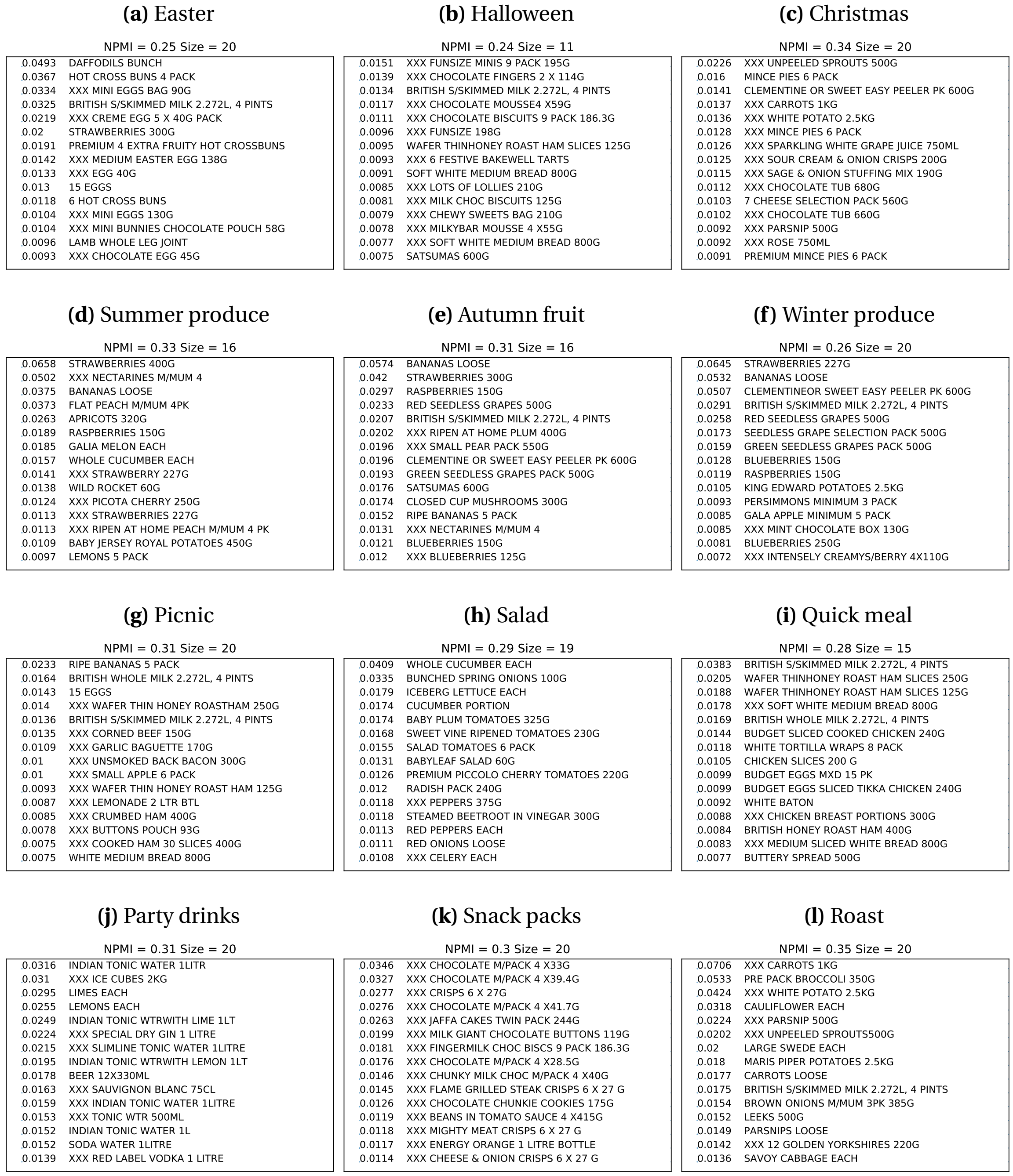
Figure 6.4: HC-SeqSTM (clustered) topics sorted by monthly proportions. Topics in sorted ordered: (1) picnic, (2) Christmas, (3) summer produce, (4) Easter, (5) Halloween, (6) autumn fruit, (7) ice cream, (8) winter produce, (9) barbecue, (10) early summer, (11) salad, (12) quick meal, (13) roast, (14) prepared fruit, (15) Crisps, (16) party drinks, (17) yoghurt and fruit, (18) spring produce, (19) cooked breakfast, (20) snack packs.

6.4.1 Temporal topics

We observe three types of temporal topics: *festive* topics such as Christmas and Easter, *seasonal* topics such as ‘summer produce’ or ‘salad’, which respond to seasonal harvest or seasonal demand; and *periodic* topics such as ‘snack packs’, which increase/decrease consumption regularly.

6.4.1.1 Festive topics

As Figure 6.4 clearly shows, the ‘Christmas’ topic becomes more likely during November and December, the ‘Easter’ topic stands out from February to April, being the most likely in March, and the ‘Halloween’ topic is more likely in October and November. The ‘Christmas’ topic is characterised by chocolate tubs, mince pies, sprouts and drinks as illustrated in Figure 6.5b; and the Easter topic is characterised by gathering chocolate eggs, daffodils, ‘hot cross buns’ and lamb as illustrated in Figure 6.5d. The ‘Halloween’ topic detailed in Figure 6.5e shows various chocolate and confectionery products (i.e., fun-size minis and chocolate fingers) which is a sign of the Halloween campaign. Further inspection of this topic finds a Halloween icon, the pumpkin, in the top 100 most likely products; The pumpkin does not rank in the top 100 in any other topic, confirming the



(a) Easter

NPMI = 0.25 Size = 20

0.0493	DAFFODILS BUNCH
0.0367	HOT CROSS BUNS 4 PACK
0.0334	XXX MINI EGGS BAG 90G
0.0325	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0219	XXX CREME EGG 5 X 40G PACK
0.02	STRAWBERRIES 300G
0.0191	PREMIUM 4 EXTRA FRUITY HOT CROSSBUNS
0.0142	XXX MEDIUM EASTER EGG 138G
0.0133	XXX EGG 40G
0.013	15 EGGS
0.0118	6 HOT CROSS BUNS
0.0104	XXX MINI EGGS 130G
0.0104	XXX MINI BUNNIES CHOCOLATE POUCH 58G
0.0096	LAMB WHOLE LEG JOINT
0.0093	XXX CHOCOLATE EGG 45G

(b) Halloween

NPMI = 0.24 Size = 11

0.0151	XXX FUNSIZE MINIS 9 PACK 195G
0.0139	XXX CHOCOLATE FINGERS 2 X 114G
0.0134	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0117	XXX CHOCOLATE MOUSSE4 X59G
0.0111	XXX CHOCOLATE BISCUITS 9 PACK 186.3G
0.0096	XXX FUNSIZE 198G
0.0095	WAFER THINHONEY ROAST HAM SLICES 125G
0.0093	XXX 6 FESTIVE BAKEWELL TARTS
0.0091	SOFT WHITE MEDIUM BREAD 800G
0.0085	XXX LOTS OF LOLLIES 210G
0.0081	XXX MILK CHOC BISCUITS 125G
0.0079	XXX CHEWY SWEETS BAG 210G
0.0078	XXX MILKYBAR MOUSSE 4 X55G
0.0077	XXX SOFT WHITE MEDIUM BREAD 800G
0.0075	SATSUMAS 600G

(c) Christmas

NPMI = 0.34 Size = 20

0.0226	XXX UNPEELED SPROUTS 500G
0.016	MINCE PIES 6 PACK
0.0141	CLEMENTINE OR SWEET EASY PEELER PK 600G
0.0137	XXX CARROTS 1KG
0.0136	XXX WHITE POTATO 2.5KG
0.0128	XXX MINCE PIES 6 PACK
0.0126	XXX SPARKLING WHITE GRAPE JUICE 750ML
0.0125	XXX SOUR CREAM & ONION CRISPS 200G
0.0115	XXX SAGE & ONION STUFFING MIX 190G
0.0112	XXX CHOCOLATE TUB 680G
0.0103	7 CHEESE SELECTION PACK 560G
0.0102	XXX CHOCOLATE TUB 660G
0.0092	XXX PARSNIP 500G
0.0092	XXX ROSE 750ML
0.0091	PREMIUM MINCE PIES 6 PACK

(d) Summer produce

NPMI = 0.33 Size = 16

0.0658	STRAWBERRIES 400G
0.0502	XXX NECTARINES M/MUM 4
0.0375	BANANAS LOOSE
0.0373	FLAT PEACH M/MUM 4PK
0.0263	APRICOTS 320G
0.0189	RASPBERRIES 150G
0.0185	GALIA MELON EACH
0.0157	WHOLE CUCUMBER EACH
0.0141	XXX STRAWBERRY 227G
0.0138	WILD ROCKET 60G
0.0124	XXX PICOTA CHERRY 250G
0.0113	XXX STRAWBERRIES 227G
0.0113	XXX RIPEN AT HOME PEACH M/MUM 4 PK
0.0109	BABY JERSEY ROYAL POTATOES 450G
0.0097	LEMONS 5 PACK

(e) Autumn fruit

NPMI = 0.31 Size = 16

0.0574	BANANAS LOOSE
0.042	STRAWBERRIES 300G
0.0297	RASPBERRIES 150G
0.0233	RED SEEDLESS GRAPES 500G
0.0207	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0202	XXX RIPEN AT HOME PLUM 400G
0.0196	XXX SMALL PEAR PACK 550G
0.0196	CLEMENTINE OR SWEET EASY PEELER PK 600G
0.0193	GREEN SEEDLESS GRAPES PACK 500G
0.0176	SATSUMAS 600G
0.0174	CLOSED CUP MUSHROOMS 300G
0.0152	RIPE BANANAS 5 PACK
0.0131	XXX NECTARINES M/MUM 4
0.0121	BLUEBERRIES 150G
0.012	XXX BLUEBERRIES 125G

(f) Winter produce

NPMI = 0.26 Size = 20

0.0645	STRAWBERRIES 227G
0.0532	BANANAS LOOSE
0.0507	CLEMENTINEOR SWEET EASY PEELER PK 600G
0.0291	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0258	RED SEEDLESS GRAPES 500G
0.0173	SEEDLESS GRAPE SELECTION PACK 500G
0.0159	GREEN SEEDLESS GRAPES PACK 500G
0.0128	BLUEBERRIES 150G
0.0119	RASPBERRIES 150G
0.0105	KING EDWARD POTATOES 2.5KG
0.0093	PERSIMMONS MINIMUM 3 PACK
0.0085	GALA APPLE MINIMUM 5 PACK
0.0085	XXX MINT CHOCOLATE BOX 130G
0.0081	BLUEBERRIES 250G
0.0072	XXX INTENSELY CREAMYS/BERRY 4X110G

(g) Picnic

NPMI = 0.31 Size = 20

0.0233	RIPE BANANAS 5 PACK
0.0164	BRITISH WHOLE MILK 2.272L, 4 PINTS
0.0143	15 EGGS
0.014	XXX WAFER THIN HONEY ROASTHAM 250G
0.0136	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0135	XXX CORNED BEEF 150G
0.0109	XXX GARLIC BAGUETTE 170G
0.01	XXX UNSMOKED BACK BACON 300G
0.01	XXX SMALL APPLE 6 PACK
0.0093	XXX WAFER THIN HONEY ROAST HAM 125G
0.0087	XXX LEMONADE 2 LTR BTL
0.0085	XXX CRUMBED HAM 400G
0.0078	XXX BUTTONS POUCH 93G
0.0075	XXX COOKED HAM 30 SLICES 400G
0.0075	WHITE MEDIUM BREAD 800G

(h) Salad

NPMI = 0.29 Size = 19

0.0409	WHOLE CUCUMBER EACH
0.0335	BUNCHED SPRING ONIONS 100G
0.0179	ICEBERG LETTUCE EACH
0.0174	CUCUMBER PORTION
0.0174	BABY PLUM TOMATOES 325G
0.0168	SWEET VINE RIPENED TOMATOES 230G
0.0155	SALAD TOMATOES 6 PACK
0.0131	BABYLEAF SALAD 60G
0.0126	PREMIUM PICCOLO CHERRY TOMATOES 220G
0.012	RADISH PACK 240G
0.0118	XXX PEPPERS 375G
0.0118	STEAMED BEETROOT IN VINEGAR 300G
0.0113	RED PEPPERS EACH
0.0111	RED ONIONS LOOSE
0.0108	XXX CELERY EACH

(i) Quick meal

NPMI = 0.28 Size = 15

0.0383	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0205	WAFER THINHONEY ROAST HAM SLICES 250G
0.0188	WAFER THINHONEY ROAST HAM SLICES 125G
0.0178	XXX SOFT WHITE MEDIUM BREAD 800G
0.0169	BRITISH WHOLE MILK 2.272L, 4 PINTS
0.0144	BUDGET SLICED COOKED CHICKEN 240G
0.0118	WHITE TORTILLA WRAPS 8 PACK
0.0105	CHICKEN SLICES 200 G
0.0099	BUDGET EGGS MXD 15 PK
0.0099	BUDGET EGGS SLICED TIKKA CHICKEN 240G
0.0092	WHITE BATON
0.0088	XXX CHICKEN BREAST PORTIONS 300G
0.0084	BRITISH HONEY ROAST HAM 400G
0.0083	XXX MEDIUM SLICED WHITE BREAD 800G
0.0077	BUTTERY SPREAD 500G

(j) Party drinks

NPMI = 0.31 Size = 20

0.0316	INDIAN TONIC WATER 1LITR
0.031	XXX ICE CUBES 2KG
0.0295	LIMES EACH
0.0255	LEMONS EACH
0.0249	INDIAN TONIC WTRWITH LIME 1LT
0.0224	XXX SPECIAL DRY GIN 1 LITRE
0.0215	XXX SLIMLINE TONIC WATER 1LITRE
0.0195	INDIAN TONIC WTRWITH LEMON 1LT
0.0178	BEER 12X330ML
0.0163	XXX SAUVIGNON BLANC 75CL
0.0159	XXX INDIAN TONIC WATER 1LITRE
0.0153	XXX TONIC WTR 500ML
0.0152	INDIAN TONIC WATER 1L
0.0152	SODA WATER 1LITRE
0.0139	XXX RED LABEL VODKA 1 LITRE

(k) Snack packs

NPMI = 0.3 Size = 20

0.0346	XXX CHOCOLATE M/PACK 4 X33G
0.0327	XXX CHOCOLATE M/PACK 4 X39.4G
0.0277	XXX CRISPS 6 X 27G
0.0276	XXX CHOCOLATE M/PACK 4 X41.7G
0.0263	XXX JAFFA CAKES TWIN PACK 244G
0.0199	XXX MILK GIANT CHOCOLATE BUTTONS 119G
0.0181	XXX FINGERMILK CHOC BISCS 9 PACK 186.3G
0.0176	XXX CHOCOLATE M/PACK 4 X28.5G
0.0146	XXX CHUNKY MILK CHOC M/PACK 4 X40G
0.0145	XXX FLAME GRILLED STEAK CRISPS 6 X 27 G
0.0126	XXX CHOCOLATE CHUNKIE COOKIES 175G
0.0119	XXX BEANS IN TOMATO SAUCE 4 X415G
0.0118	XXX MIGHTY MEAT CRISPS 6 X 27 G
0.0117	XXX ENERGY ORANGE 1 LITRE BOTTLE
0.0114	XXX CHEESE & ONION CRISPS 6 X 27 G

(l) Roast

NPMI = 0.35 Size = 20

0.0706	XXX CARROTS 1KG
0.0533	PRE PACK BROCCOLI 350G
0.0424	XXX WHITE POTATO 2.5KG
0.0318	CAULIFLOWER EACH
0.0224	XXX PARSNIP 500G
0.0202	XXX UNPEELED SPROUTS500G
0.02	LARGE SWEDE EACH
0.018	MARIS PIPER POTATOES 2.5KG
0.0177	CARROTS LOOSE
0.0175	BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0154	BROWN ONIONS M/MUM 3PK 385G
0.0152	LEEKS 500G
0.0149	PARSNIPS LOOSE
0.0142	XXX 12 GOLDEN YORKSHIRES 220G
0.0136	SAVOY CABBAGE EACH

Figure 6.5: HC-SeqSTM (clustered) topics in the UK grocery retail market baskets. Each topic is characterised by the 15 products with the largest probabilities. Probabilities and products are sorted in descending order. Brand names have been replaced by XXX for anonymisation purposes. NPMI is a measure of topic coherence; NPMI ≥ 0.30 indicates that the listed products are frequently bought together; NPMI ≤ 0.0 indicates highly incoherent topics where the listed products are frequently bought separately. Size refers to the cluster size; a size of 20 indicates that the topic is highly recurrent as it has appeared in every posterior sample.

topic's Halloween theme.

6.4.1.2 Seasonal topics

We observe various topics with seasonal patterns. These topics may not only respond to product availability (i.e., fruit available during warm months) but also to customer needs that respond to temperature (i.e., warm/highly calorific foods in cold months).

Spring/summer months show topics such as the 'picnic', 'summer produce', 'ice cream', 'barbecue', 'early summer fruit', 'salad', 'prepared fruit', 'yoghurt and fruit' and 'spring produce' as observed in Figure 6.4. The 'picnic' topic listed in Figure 6.5g is characterised by sandwich fillers, refreshments and seasonal fruits such as apricots and nectarines. The 'summer produce' topic also shows seasonal fruits and produce such as apricots, peaches, nectarines, melon and new jersey potatoes as depicted in Figure 6.5d, showing a more general grocery topic that is driven by product availability [147, 148]. A weather-oriented topic is depicted by the 'salad' topic listed in Figure 6.5h, which is more likely from April to August and less likely during the rest of the year.

Topics containing products that are more likely in colder months are the 'autumn fruit' and 'winter produce', as shown in Figure 6.4. In comparison to the 'summer produce', the 'autumn fruit' topic, depicted in Figure 6.5e, shows fruits such as pears, plums and satsumas that are in season from September through to January [147, 148]; and the 'winter produce', depicted in Figure 6.5f, highlights King Edward potatoes and persimmons that are in season from early autumn into spring [147, 148].

Other topics that are also more likely during cold months and illustrate high calorific foods are the 'quick meal', 'roast', and 'cooked breakfast'. The 'quick meal', characterised by bread, soda, cooked meat and sandwich fillers as depicted in Figure 6.5i and 'cooked breakfast' which contains sausages, bacon, butter and bread. The 'roast' topic, illustrated in Figure 6.5l, is characterised by vegetables, chicken and Yorkshire puddings.

6.4.1.3 Periodic topics

Apart from festive and seasonal topics, we observe that the ‘crisps’, ‘snack pack’, and ‘party drinks’ topics follow a periodic pattern during the year. As shown in Figure 6.4, the ‘snack pack’ topic, depicted in Figure 6.5k, is more probable during 3 pairs of months: October-November, January-February and May-June, suggesting that this topic might follow school terms. The ‘crisps’ topic is more likely during December, March and summer months, suggesting that this topic might follow school holidays. The ‘party drinks’ topic, characterised by alcoholic beverages and complementary items such as ice cubes, tonic water and limes depicted in Figure 6.5i, is slightly more likely in December and July. Note that the ‘crisps’ and ‘party drinks’ topics accompany the ‘Christmas’ topic as the three topics show larger topic proportions in December.

6.4.2 What does SeqSTM have that STM and LDA do not?

Figures 6.6 and 6.7 show the top 20 clustered STM/LDA topics with the largest monthly variations. Each topic is matched with the closest (in cosine similarity) clustered SeqSTM topic; indexes and cosine similarities are shown in the bottom labels. Cosine similarity larger than 0.7 indicates that the matching clustered topics are similar, i.e., the ‘Christmas’ topic.

As observed in Figure 6.6, clustered STM topics are also identified by SeqSTM with high similarities. However, clustered SeqSTM topics such as ‘picnic’, ‘Halloween’, and ‘winter produce’, are not found among the clustered STM topics. The STM fuses the ‘picnic’ and ‘Halloween’ topics with the ‘summer fruit’ and ‘autumn produce’ topics, respectively.

In Figure 6.7, we see that LDA identifies fewer topics with temporal patterns than the STM or SeqSTM. Only the first 5 clustered LDA topics show topic proportions that double their monthly average. In contrast, the first 9 clustered SeqSTM topics double their monthly average. The clustered SeqSTM topics such as ‘quick meal’, ‘cooked breakfast’, ‘prepared fruit’ and ‘yoghurt and fruit’, which have shown more popularity in cold and warm months, are captured by LDA but without a strong temporal pattern.

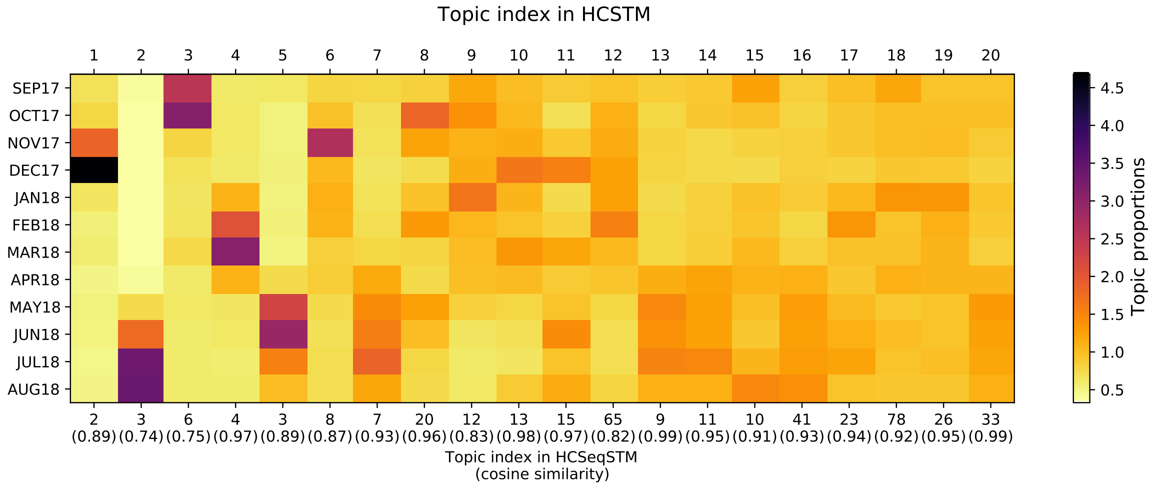


Figure 6.6: HC-STM (clustered) topics sorted by monthly proportions. Topics in sorted order: (1) Christmas, (2) summer fruit-picnic, (3) autumn produce and Halloween, (4) Easter, (5) summer produce, (6) autumn fruit, (7) ice cream, (8) snack packs, (9) quick meal, (10) roast, (11) crisps, (12) beef stew, (13) barbecue, (14) salad (15) late summer fruit, (16) meal promotion, (17) convenience, (18) yoghurt and fruit, (19) diet, (20) potato salad. Each STM clustered topic is matched with the SeqSTM clustered topic with the largest cosine similarity, e.g., the first STM clustered topic matches with the second SeqSTM clustered topic with 0.89 cosine similarity.

HC-SeqSTM clustered topics show smooth temporal patterns with a gradation of the topic proportions. For instance, the 'Halloween' topic in 6.5 spikes in October, and then, the topic proportions gradually decrease during November and December. In contrast, clustered STM/LDA topics show more month-specific patterns. For example, the topic 'autumn fruit' in Figure 6.6 spikes in November, drops in December and keeps low for the rest of the year.

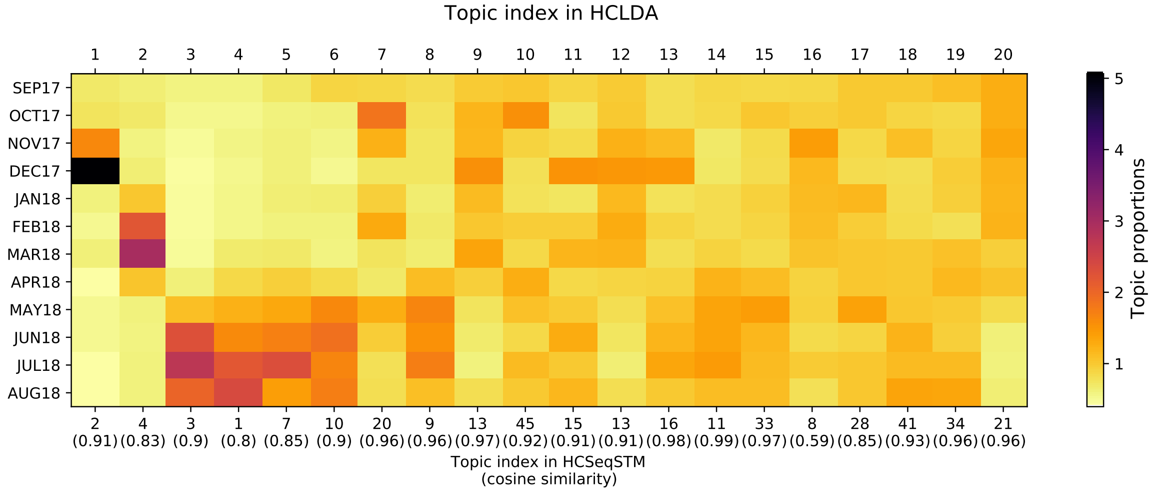


Figure 6.7: HC-LDA (clustered) topics sorted by monthly proportions. Topics in sorted order: (1) Christmas, (2) Easter, (3) summer produce, (4) picnic, (5) ice cream,(6) early summer, (7) snack packs, (8) barbecue, (9) roast, (10) milkshake, (11) crisps, (12) beef stew, (13) party drinks, (14) salad, (15) beef meal, (16) mixed basket, (17) vegetarian, (18) meal promotion, (19) Chocolate packs (20) budget. Each LDA clustered topic is matched with the SeqSTM clustered topic with the largest cosine similarity, e.g., the first LDA clustered topic matches with the second SeqSTM clustered topic with 0.91 cosine similarity.

6.5 Practical implications

Commercially speaking, identifying temporal patterns aid retailers to design marketing campaigns targeting seasonal customer behaviours, i.e., using highly ranked products in the 'salad' topic in summer or 'cooked breakfast' in winter to design promotions or recommendations. Analysing time-specific topical mixtures, retailers can track the growth or decline of topics, which aids the planning of product assortments. Transaction-specific topical mixtures (and customer-specific topical mixtures if customer's order history is available) could be exploited as features for customer segmentation and predictive modelling.

Quantifying temporal variations of customer behaviours in grocery retail transactions could be exploited in sociological research. For instance, tracking month-specific topic probabilities from the last five years may show a growth of 'organic' food consumption, showing, for example, how much on average that topic represents over an average transaction or market basket. Dietary studies could also exploit the outcomes of SeqSTM to quantify how eating habits change over time.

6.6 Summary

In this chapter, we demonstrated the application of topic models to identify temporal patterns in shopping behaviours in the grocery retail domain. Inspired by the dynamic mixture model, we propose a modification of the segmented topic model, named the sequential segmented topic model, to exploit the temporal sequence. We demonstrated that the sequential segmented topic model can identify time-driven topics that are fused or overlooked by STM or LDA. Moreover, the posterior summary of the SeqSTM achieves better model generalisation than the posterior summaries of LDA and STM.

We observe three types of temporal topics: festive, seasonal and periodic topics. Festive topics respond to customer behaviours driven by celebrations of Christmas, Easter, to name a few. Seasonal topics show consumption due to seasonal availability and seasonal demand such as 'Autumn fruit' or low/high-calorie foods. Periodic topics show increased popularity on regular basis such as 'crisps' or 'snack packs'. Analysing time-specific topical mixtures and topics' product composition could help retailers to customise product assortments and design marketing campaigns targeting time-specific popular topics. Topic models for the analysis of temporal patterns in food consumption could provide new venues for sociological research.

Chapter 7

Conclusions and Further Work

In this thesis, we investigated applications of topic models to the analysis of retail data. We conclude that topic modelling is a useful framework to identify customer behaviours through the analysis of large volumes of transactional data. We showed that LDA is capable of identifying combinations of products that are frequently bought together for the fulfilment of customer needs. A large variety of grocery topics have been identified, from topics that show a preference for specific types of foods, dishes and quality to topics that show events, activities, household composition, etc. Customer behaviours that respond to local demand and local supply are identified by the application of STM, which is suitable to accommodate store hierarchy over transactions. Finally, we showed that the application of SeqSTM finds customer behaviours that are driven by seasonal demand and product availability. Outcomes of this investigation have many practical implications with potential commercial and social impacts.

In more detail, we have identified:

Topic distributions: Inferred topics are not always the most coherent, may show product combinations that do not correspond to genuine customer behaviour. Topic distributions from a single posterior sample may not be the most distinctive; users may associate two or more topics with the same customer behaviour. Moreover, depending on their uncertainty, topics may appear and disappear across posterior samples or runs of the same model. Topic instability may reduce user confidence in the application of topic models.

Posterior summary: Our clustering methodology identifies a broader set of topic distributions from which users can select topics depending on their measures of topic uncertainty. Posterior summaries obtained through the clustering methodology show better performance of model generalisation and topic coherence than individual posterior samples of LDA, STM or SeqSTM. Moreover, posterior summaries of highly recurrent clustered topics are associated with measures of high credibility.

Spatial modelling: STM identifies topics that are relevant over the constituent countries of the UK and regions of England. Harnessing store structure allows topics to be constructed under store context, thereby identifying topics that may be relevant only in specific areas. Analysing product descriptions and mapping topic probabilities aid the analysis and identification of regional topics. LDA overlooks some of these regional behaviours, and even larger models of LDA fail to identify such regional topics. Linear Gaussian process regression complements the analysis of customer behaviours with spatial patterns by quantifying regional effects while capturing spatial dependence.

SeqSTM and temporal priors: SeqSTM identifies customer behaviours with temporal patterns that respond to seasonal product availability and seasonal demand. By accommodating time structure over transactions and temporal sequence over time slices, SeqSTM retrieves topics with temporal patterns that are fused or overlooked by STM or LDA. Analysing month-specific topic probabilities reveals temporal patterns associated with festive, seasonal and periodic customer behaviours.

This investigation has brought the following academic contributions:

- A clustering methodology that fuses topic distributions obtained from multiple samples to identify clusters of topics and quantify their uncertainty.

- An evaluation framework for topic models that includes four concepts: the generalisation of the model, topic coherence, *topic distinctiveness* and *topic credibility*, along with metrics for measuring topic distinctiveness and topic credibility.

- The demonstration of segmented topic model and linear Gaussian process regression to analyse grocery transactional data accounting for store hierarchy over transactions and to identify customer behaviours with spatial patterns.

- A topic model named 'sequential segmented topic model' (SeqSTM) which aims to identify grocery topics that respond to time-variant customer behaviours by exploiting temporal sequence and temporal hierarchy over transactions. SeqSTM allows the detection of customer behaviours that respond to seasonal availability and seasonal demand.

7.1 Limitations

The biggest limitation of this work is the computation time of MCMC methods, which were used to solve the inference of topic models. In the analysis of retail data, Gibbs sampler algorithms demand long periods of computation due to the sheer volume of transactions and high dimensionality of product assortments. For applications to retail data where a large number of customer behaviours coexist, large topic models are needed; thereby, large computational times are required to fit large topic models. Extremely long computational time hampered the analysis of large datasets and the application of topics models of large complexity.

Given the computational limitations of inference methods, we constrained our analysis by reducing the number of transactions and the size of the product assortment. We used a sample of 36,000 transactions and worked with the 10,000 most popular items. Transactions were randomly sampled from 100 nationwide stores. Item popularity was measured by the number of transactions containing the item divided by the number of months the item was available. Infrequent products are more unlikely to rank high among topics. Fitting topic models with full assortments and large data sets may provide a wider variety of customer behaviours, but at a significantly greater computational cost. Our results showed that our sampled data were enough to show numerous customer behaviours. However, modelling a larger set of transactions and product assortment may identify more detailed topics.

7.2 Future work

We applied topic models with generative processes that assume products as conditionally independent. This means that products could be sampled several times for the same transaction. This assumption may not represent transactions in which customers choose products once. In addition, LDA, STM and SeqSTM assume that topic assignments are sampled i.i.d from topical mixtures; this assumption does not fit highly correlated customer behaviours. Further work should explore methods such as CTM or PAM which account for topic correlations.

In our application of LDA, we ran symmetric Dirichlet priors in which the precision parameter is fixed a priori. As observed by [94], asymmetric priors can improve topic coherence by capturing highly frequent items in a small number of topics. Empirically, we found that the optimisation of Dirichlet parameters brings Markov chains to local modes, making the convergence of Markov chains difficult. Further work should explore sampling methods that consider asymmetric priors while aiming for good chain mixing.

When searching for spatial patterns in customer behaviours, we applied STM to accommodate store hierarchy over transactions and to obtain store-specific topical compositions. A potential extension of this model could account for topical correlation between store-specific topical mixtures, i.e., using geographical distance as a proxy of association between topical mixtures of nearby stores. Linear Gaussian process regression complemented our analysis modelling topic prevalence. We still observed residuals from this model with spatial patterns in small areas, which indicate a model's misfit. Further work should investigate regression methods that account for spatial autocorrelation with heterogeneous covariance.

SeqSTM assumes time-specific topical mixtures are Dirichlet distributed with parameters that derived from the previous time-specific topical mixture. The Dirichlet distribution does not conjugate with another Dirichlet distribution; thereby, inference has to be handled by an iterative procedure that updates time-specific topical mixtures as in [46]. An extension of this work could assume that time-specific topical mixtures are PDP distributed as in [67, 45]. Inference may be solved by coagulating PDPs as mentioned in [54].

Appendix A

Topic Modelling

A.1 Markov chain simulation

Markov chain simulation, also called Markov chain Monte Carlo, or MCMC, is a general method based on drawing values of random variables, say θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution $p(\theta \mid y)$. The samples are drawn sequentially from a *transition distribution*, $T_t(\theta^t \mid \theta^{t-1})$, that depends on the previous draw θ^{t-1}. The transition probability distributions must be constructed so that the Markov chain converges to a unique stationary distribution that is the posterior distribution, $p(\theta \mid y)$.

A.1.1 The Gibbs sampler

A particular Markov chain algorithm is the *Gibbs sampler.* The idea behind Gibbs sampling is that for each variable a sample is drawn in each turn, conditioned on the values of all the other variables in the distribution. Let the parameter vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)$. At each iteration t, the Gibbs sampler cycles through the d components in $\boldsymbol{\theta}$; each θ_j^t is sampled from the conditional distribution given all the other components of $\boldsymbol{\theta}$:

$$p(\theta_j \mid \theta_{-j}^{t-1}, y), \tag{A.1}$$

where θ_{-j}^{t-1} represents all the components of $\boldsymbol{\theta}$, except for θ_j, at their current values:

$$\theta_{-j}^{t-1} = (\theta_1^t, ..., \theta_{j-1}^t, \theta_{j+1}^{t-1}, ..., \theta_d^{t-1}). \tag{A.2}$$

Thus, each component θ_j is updated conditional on the latest values of the other components of $\boldsymbol{\theta}$.

A.1.2 Assessing convergence

One practical approach to assess convergence is to run multiple chains from different dispersed starting points and to plot the samples of some variables of interest in a trace plot. If the chains are well mixed, i.e., overlapping each other, then the trace plot suggests that the chains have converged to the same distribution.

A.1.2.1 Estimated potential scale reduction

Another method to assess convergence is to compute the *estimated potential scale reduction* (EPSR). EPSR compares the variance of a quantity ψ within each chain to its variance across chains.

$$\widehat{R} = \sqrt{\frac{\widehat{var}^+(\psi \mid y)}{W}}, \tag{A.3}$$

which declines to 1 as $n \to \infty$. Where the marginal posterior variance of the quantity ψ is given by:

$$\widehat{var}^+(\psi \mid y) = \frac{n-1}{n} W + \frac{1}{n} B, \tag{A.4}$$

where between-sequence variance B and within-sequence variance W are given by:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\psi}_{.j} - \overline{\psi}_{..}), \quad \overline{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}, \quad \overline{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \overline{\psi}_{.j} \tag{A.5}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \overline{\psi}_{.j})^2. \tag{A.6}$$

A.1.3 Difficulties with Markov chains

There are two main challenges when simulating Markov chains. Firstly, if the sequence has not run for long enough, the simulations may not be represented of the target distribution. On the other hand, early iterations are influenced by the staring approximation rather than the target distribution. In our applications of topic models, we run Markov chains until the trace plots show steady and well-mixed distributions. We disregard between 60% to 80% of samples that are still influenced by starting points.

Secondly, sequential draws tend to show high correlation and simulation inference from correlated draws is generally less precise than from the same number of independent draws. In our applications of topic models, we recorded the models' log-likelihood every 10 iterations; however, a thin of such length still shows a large correlation. As such, we evaluate the posterior performance of topic models using samples separated by 5,000 iterations (after burn-in) in Chapters 4 and 5 and 10,000 iterations (after burn-in) in Chapter 6. Such long thins were chosen to select samples with non-significant correlation.

Appendix B

Clustering and Evaluation of Topic Models: Identifying British Customer Behaviour in Grocery Retail

B.1 Hierarchical clustering

The hierarchical clustering algorithm takes a bag of topics, a list with sample indexes, and a cosine distant threshold. The bag of topics gathers topic distributions from various posterior samples from various MCMC. The list of sample indices records a sample index for each topic, i.e., assuming that the first 50 topics in the bag of topics come from posterior sample 1 and the next 50 topics come from posterior sample 2, then the first 50 elements in the list of samples indices are 1 and the next 50 elements are 2. The cosine distance threshold indicates the limit up to which topics would be merged.

The algorithm will start by forming clusters with each of the topics in the bag of topics. So, if there are N topics, there are N initial clusters. Then, a list L is created to record the cosine distance between two clusters. This list contains the indexes of the two compared clusters and the cosine distance between the clustered topics. A clustered topic is the average topic distributions of the cluster members.

At each step, the algorithm finds in L the pair of clusters with the minimum cosine distance. Then, the algorithm evaluates if the members of both clusters are from different posterior samples using the list of sample indices. If the evaluation is true (topics from different posterior samples), then a new cluster is created by merging the pair of clusters with the minimum cosine distance. Then, the algorithm removes from the L all comparisons that had any of the identified clusters and adds comparisons from all the remaining clusters to the new clusters. If the evaluation is false, then the algorithm would update the cosine distance in L with 1, so the algorithm would evaluate a new pair of clusters in the next step.

The algorithm will keep merging clusters until the minimum cosine distance is larger than the cosine distance threshold. The algorithm then retrieves all the remaining clusters (clusters that are not eliminated because they do not get merged).

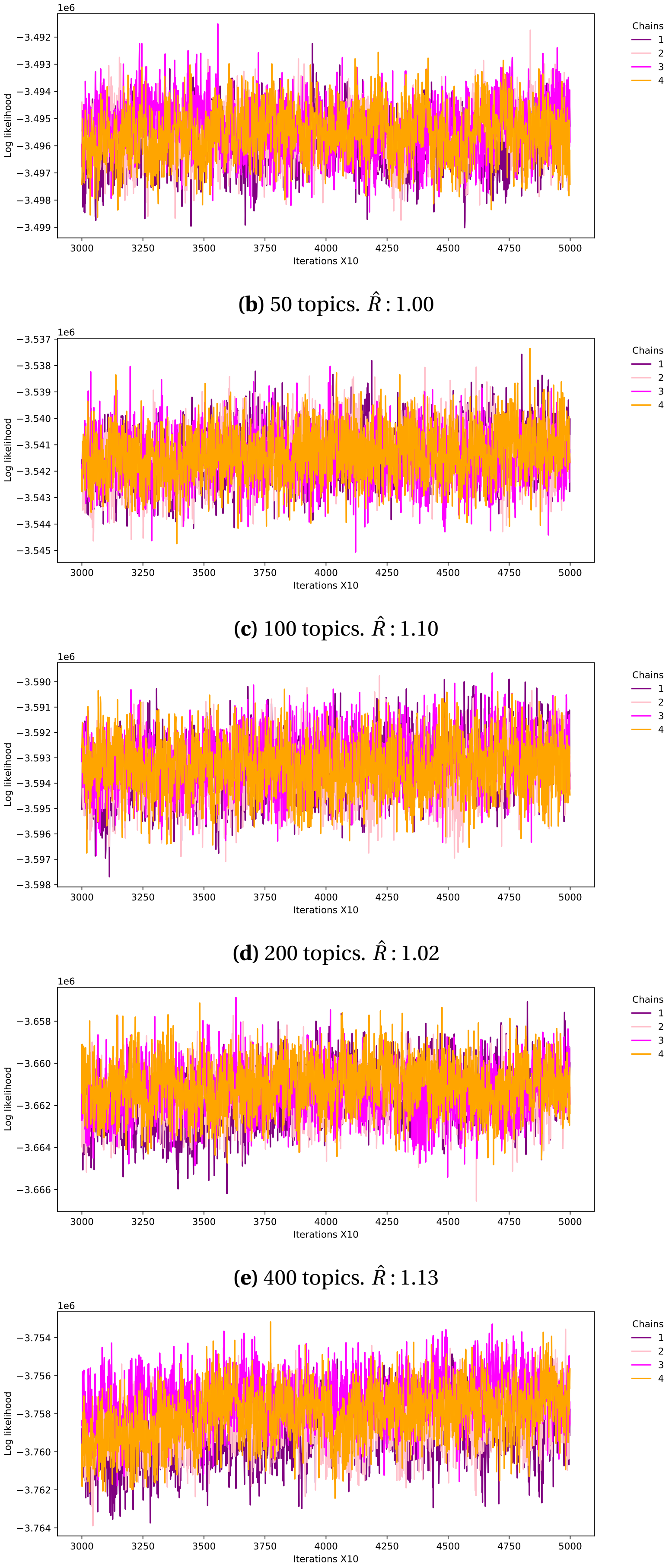
B.2 MCMC convergence of LDA

We evaluate LDA with 25, 50, 100, 200 and 400 topics. For each LDA model, 4 Markov chains are run for 50,000 iterations with a burn-in period of 30,000 iterations. Log-likelihood is measured at every 10 iterations. We calculate the potential scale reduction factor using 8000 samples.

We calculate the autocorrelation of the log-likelihood from a random taken chain using various lags. Samples with 5,000 iterations in between show non-significant autocorrelation.

Algorithm 1 Hierarchical clustering of topic distribtuions.

```
procedure CLUSTERING($\Phi, SampleIndexList, threshold$)
    n=0
    K = length of $\Phi$
    create an empty list $L$
    for i = 1 to K do
        create cluster $C_n = k$ containing index $k$
        n +=1
    end for
    for i = 1 to K do
        for j = i+1 to K do
            add to list $L$ [cluster index $i$,cluster index $j$, cosine distance between $\phi_i$ and $\phi_j$]
        end for
    end for
    find in $L$ the row $r$ with the minimum $cd$
    while $L[r,2] < threshold$ do
        differentSamples == true
        for $i \in C_{L[r,0]}$ do
            for $j \in C_{L[r,1]}$ do
                if $SampleIndexList[i] == SampleIndexList[j]$ then
                    differentSamples == false
                    Break
                end if
            end for
        end for
        if differentSamples == true then
            create $C_n = C_{L[r,0]} \cup C_{L[r,1]}$
            create $phi_n$ averaging $\phi_m \forall m \in C_n$
            delete in $L$ all rows with $ci = L[r,0]$ or $cj = L[r,0]$
            delete in $L$ all rows with $ci = L[r,1]$ or $cj = L[r,1]$
            delete $C_L[r,0]$ and $C_L[r,1]$
            for j = 1 to n-1 do
                if $C_j \exists$ then
                    add to list $L$ [cluster index $n$, cluster index $j$, cosine distance between $\phi_n$ and $\phi_j$ ]
                end if
            end for
            n +=1
        else
            update $L[r,2] = 1$
        end if
    end while
end procedure
```


(**a**) 25 topics. $\hat{R}: 1.00$

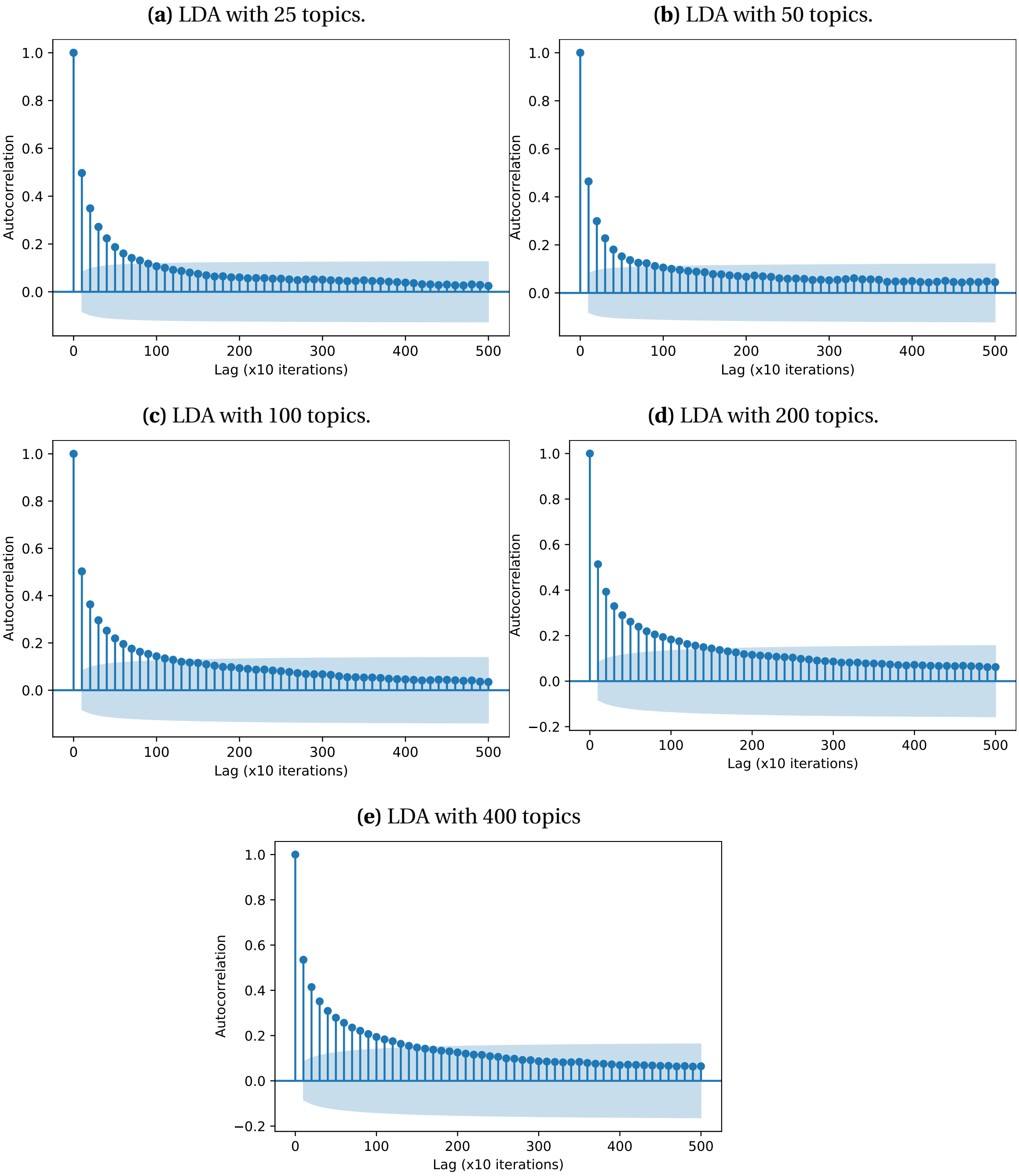
(**b**) 50 topics. $\hat{R}: 1.00$

(**c**) 100 topics. $\hat{R}: 1.10$

(**d**) 200 topics. $\hat{R}: 1.02$

(**e**) 400 topics. $\hat{R}: 1.13$

Figure B.1: Markov chains of LDA with 25, 50, 100, 200, and 400 topics. $\hat{R}$ is the potential scale reduction factor.



(a) LDA with 25 topics.

(b) LDA with 50 topics.

(c) LDA with 100 topics.

(d) LDA with 200 topics.

(e) LDA with 400 topics

Figure B.2: Log likelihood autocorrelation with 25, 50, 100, 200, and 400 LDA topics. Autocorrelations under the shaded are not significant.

Appendix C

Identifying Regional Behaviours and Modelling Spatial Prevalence

C.1 MCMC convergence of STM with 100 topics

We evaluate four Markov chains of STM with 100 topics. Markov chains are run for 100,000 iterations with a burn-in period of 80,000 iterations. Log-likelihood is measured at every 10 iterations. We calculate the potential scale reduction factor using 8,000 samples.

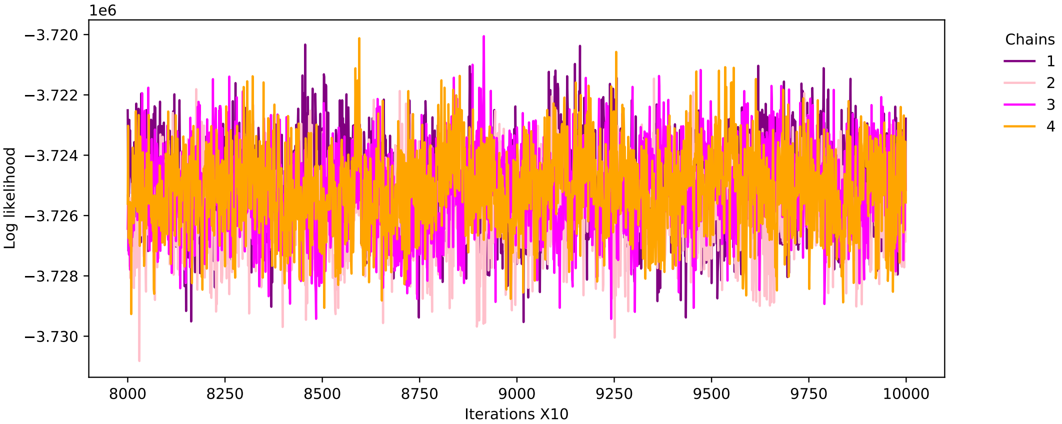


Figure C.1: Markov Chains of STM with 100. Potential scale reduction factor $\hat{R}$: 1.07.

We calculate the autocorrelation of the log-likelihood from a random taken chain using various lags. Samples with 5,000 iterations in between show non-significant autocorrelation.

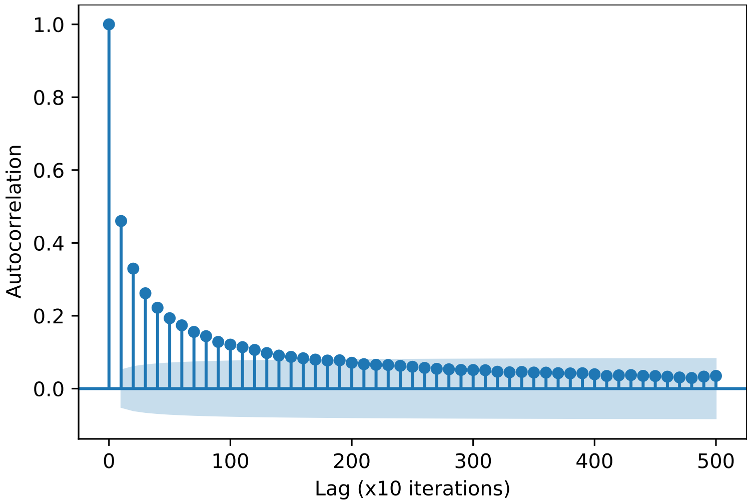


Figure C.2: Log likelihood autocorrelation with 100 STM topics known a priori. Autocorrelations under the shaded are not significant.

C.2 MCMC convergence of clustered STM topics

We run STM with 104 clustered topics known a priori for 1,500 iterations and burn-in period of 1,000 iterations.

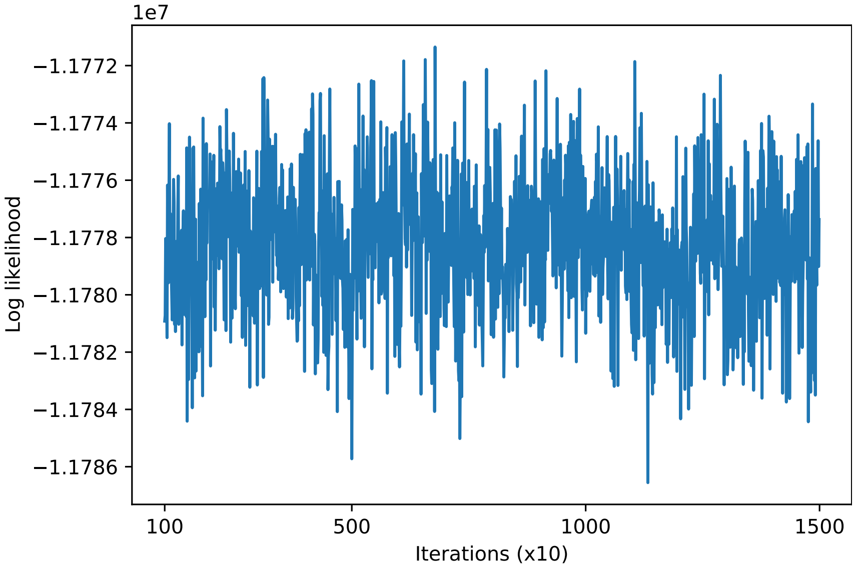


Figure C.3: Markov Chain of STM with 104 clustered topics known a priori. Potential scale reduction factor $\hat{R}: 0.998$.

We calculate the autocorrelation of the log-likelihood from MCMC with 104 clustered STM topics. Samples with 500 iterations in between show non-significant autocorrelation.

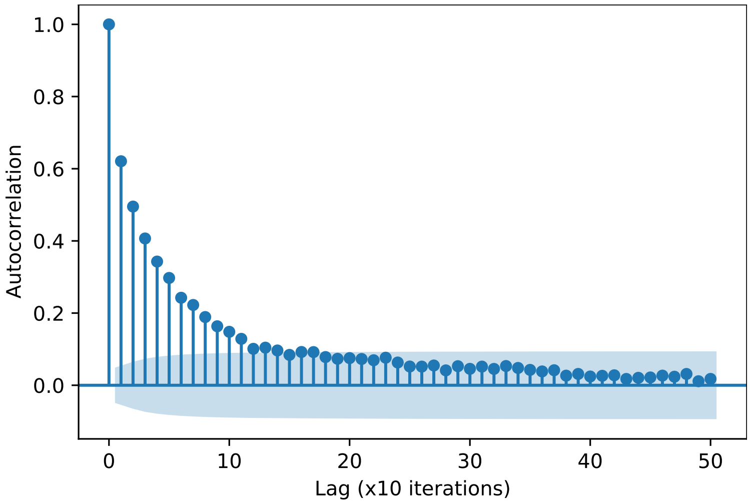


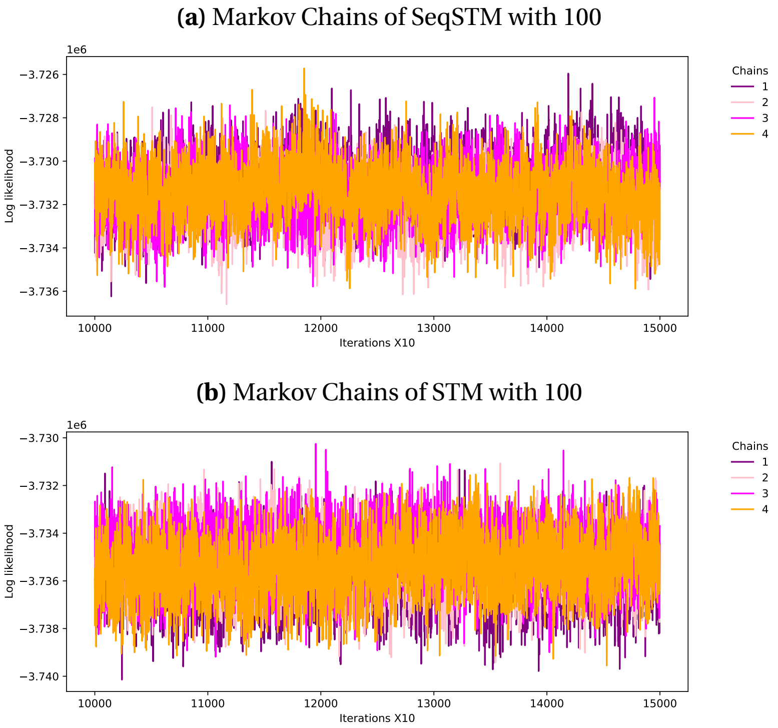
Figure C.4: Log likelihood autocorrelation with 104 clustered STM topics known a priori. Autocorrelations under the shaded are not significant.

Appendix D

Finding Temporal Behaviours and the Sequential Segmented Topic Model

D.1 Convergence of SeqSTM and STM

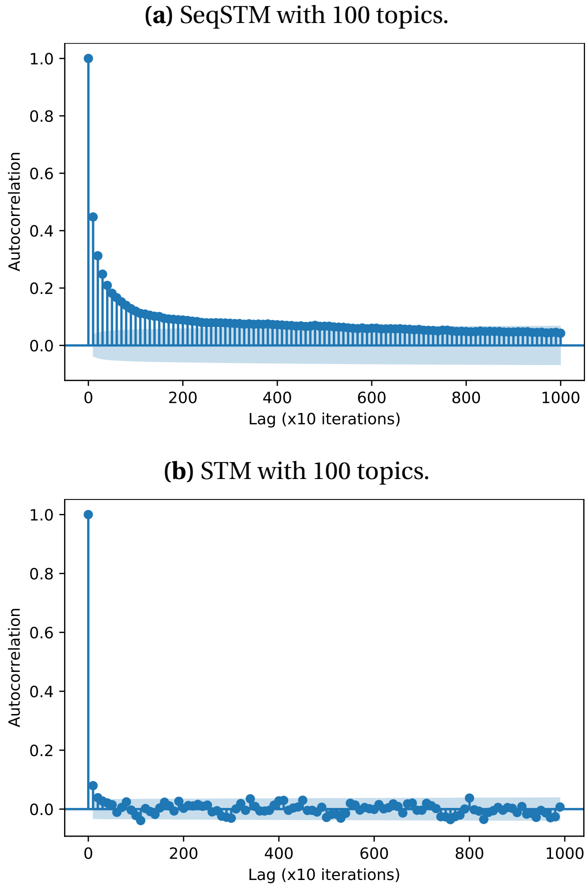
We evaluate four Markov chains of SeqSTM and STM with 100 topics. Markov chains are run for 150,000 iterations with a burn-in period of 100,000 iterations. Log-likelihood is measured at every 10 iterations. We calculate the potential scale reduction factor using 20,000 samples. Trace plots of LDA are shown in Appendix B.2.



(a) Markov Chains of SeqSTM with 100

(b) Markov Chains of STM with 100

Figure D.1: D.1a: Potential scale reduction factor $\hat{R}$: 1.04. D.1b: Potential scale reduction factor $\hat{R}$: 1.09.



(a) SeqSTM with 100 topics.

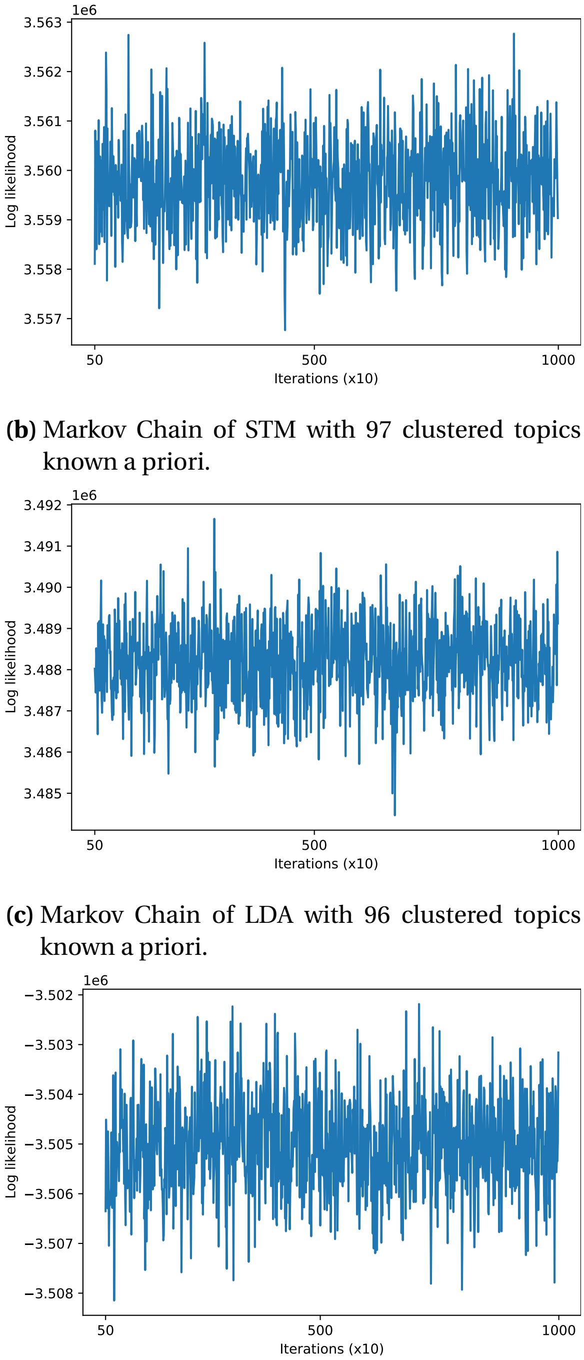
(b) STM with 100 topics.

Figure D.2: D.2a: autocorrelation of log likelihood with 100 SeqSTM topics.D.2b: autocorrelation of log likelihood with 100 STM topics. Autocorrelations under the shaded area are not significant.

We calculate the autocorrelation of the log-likelihood from a random chain using various lags. Samples with 10,000 iterations in between show non-significant autocorrelation.

D.2 MCMC convergence of clustered topics

We run SeqSTM with 198 clustered topics known a priori, STM with 97 clustered topics and LDA with 96 clustered topics for 10,000 iterations with a burn-in period of 500 iterations.



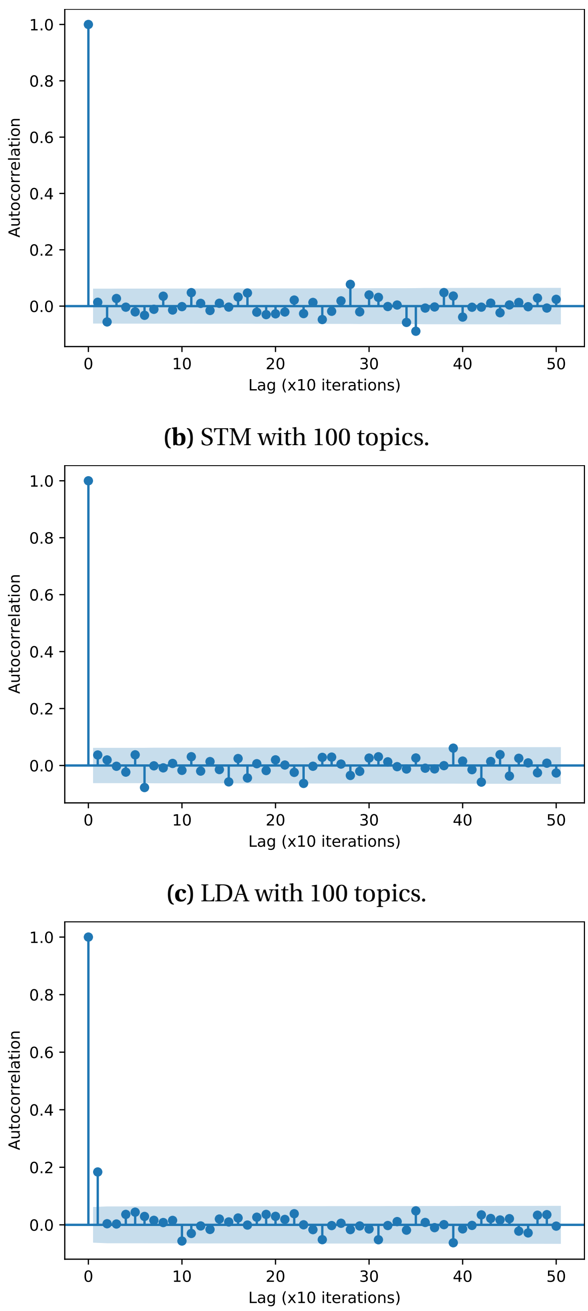
(a) Markov Chain of SeqSTM with 98 clustered topics known a priori.

(b) Markov Chain of STM with 97 clustered topics known a priori.

(c) Markov Chain of LDA with 96 clustered topics known a priori.

Figure D.3: D.3a: Potential scale reduction factor $\hat{R}: 0.996$. D.3b: Potential scale reduction factor $\hat{R}: 0.996$. D.3c: Potential scale reduction factor $\hat{R}: 0.997$.

We calculate the autocorrelation of the log-likelihood from MCMC with aforementioned clustered topics. Samples with 500 iterations in between show non-significant autocorrelation.



(a) SeqSTM with 100 topics.

(b) STM with 100 topics.

(c) LDA with 100 topics.

Figure D.4: D.4a: autocorrelation of the log likelihood with 98 clustered SeqSTM topics. D.4b: autocorrelation of the log likelihood with 97 clustered STM topics. D.4c: autocorrelation of the log likelihood with 96 clustered LDA topics. Autocorrelations under the shaded area are not significant.

Bibliography

[1] Teresa Bianchi-Aguiar, Elsa Silva, Luis Guimarães, Maria Antónia Carravilla, José F Oliveira, João Günther Amaral, Jorge Liz, and Sérgio Lapela. Using analytics to enhance a food retailer's shelf-space management. *Interfaces*, 46(5):424–444, 2016.

[2] Karel H Van Donselaar, Vishal Gaur, Tom Van Woensel, Rob ACM Broekmeulen, and Jan C Fransoo. Ordering behavior in retail stores and implications for automated replenishment. *Management Science*, 56(5):766–784, 2010.

[3] Marshall Fisher, Kumar Rajaram, and Ananth Raman. Optimizing inventory replenishment of retail fashion products. *Manufacturing & Service Operations Management*, 3(3):230–241, 2001.

[4] Teresa Bianchi-Aguiar, Elsa Silva, Luis Guimarães, Maria Antónia Carravilla, and José F Oliveira. Allocating products on shelves under merchandising rules: Multi-level product families with display directions. *Omega*, 76:47–62, 2018.

[5] H Neil Geismar, Milind Dawande, BPS Murthi, and Chelliah Sriskandarajah. Maximizing revenue through two-dimensional shelf-space allocation. *Production and Operations Management*, 24(7):1148–1163, 2015.

[6] Francis Buttle. Retail space allocation. *International Journal of Physical Distribution & Materials Management*, 14(4):3–23, 1984.

[7] Kris Johnson Ferreira and Joel Goh. Assortment rotation and the value of concealment. *Management Science*, 67(3):1489–1507, 2020.

[8] Manthan. Top 3 grocery retail strategies using better shopper insights. `https://www.manthan.com/blogs`. Visited on: 2021-06-14.

[9] Mümin Kurtuluş and Alper Nakkas. Retail assortment planning under category captainship. *Manufacturing & Service Operations Management*, 13(1):124–142, 2011.

[10] Flavian Vasile, Elena Smirnova, and Alexis Conneau. Meta-prod2vec: Product embeddings using side-information for recommendation. In *RecSys '16*, pages 225–232, 2016.

[11] Konstantinos Christidis, Dimitris Apostolou, and Gregoris Mentzas. Exploring customer preferences with probabilistic topics models. In *ECML PKDD '10*, pages 12–24, 2010.

[12] Mu-Chen Chen, Ai-Lun Chiu, and Hsu-Hwa Chang. Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4):773–781, 2005.

[13] Office for national statistics. `https://www.ons.gov.uk/businessindustryandtrade/retailindustry/datasets/poundsdatatotalretailsales`. Visited on: 2021-06-14.

[14] Department for Environment, Food and Rural Affairs. Family food 201718. `https://www.gov.uk/government/publications/family-food-201718/family-food-201718`. Visited on: 2021-06-14.

[15] Kantar. `https://www.kantarworldpanel.com/grocery-market-share/great-britain/snapshot/06.09.20`. Visited on: 2021-06-14.

[16] Statista. Grocery market share in great britain year on year comparison. https://www.statista.com/statistics/300656/grocery-market-share-in-great-britain-year-on-year-comparison/. Visited on: 2021-06-14.

[17] Tesco PLC. Annual report. https://www.tescoplc.com/media/474803/68336_tesco_ar_digital_interactive_250417.pdf. Visited on: 2021-06-14.

[18] Asda. Company facts. https://corporate.asda.com/our-story/company-facts. Visited on: 2021-06-14.

[19] USDA Foreign Agriculture Service. UK supermarket chain profiles 2016. https://www.fas.usda.gov/data/united-kingdom-uk-supermarket-chain-profiles-2016. Visited on: 2021-06-14.

[20] Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. *Advances in Data Analysis and Classification*, pages 1–19, 2020.

[21] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93*, pages 207–216, 1993.

[22] Yen-Liang Chen, Kwei Tang, Ren-Jie Shen, and Ya-Han Hu. Market basket analysis in a multiple store environment. *Decision Support Systems*, 40(2):339–354, 2005.

[23] Juan M Ale and Gustavo H Rossi. An approach to discovering temporal association rules. In *SAC' 00*, pages 294–300, 2000.

[24] Eliseo Clementini, Paolino Di Felice, and Krzysztof Koperski. Mining

multiple-level spatial association rules for objects with a broad boundary. *Data & Knowledge Engineering*, 34(3):251–270, 2000.

[25] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[26] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[27] Thomas Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR '99*, pages 50–57, 1999.

[28] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[29] Bruno JD Jacobs, Bas Donkers, and Dennis Fok. Model-based purchase predictions for large assortments. *Marketing Science*, 35(3):389–404, 2016.

[30] Harald Hruschka. Linking multi-category purchases to latent activities of shoppers: analysing market baskets by topic models. *Journal of Research and Management*, 36(4):267–273, 2014.

[31] Harald Hruschka. Hidden variable models for market basket data. University of Regensburg Working Papers in Business, Economics and Management Information Systems 489, University of Regensburg, Department of Economics, December 2016.

[32] Adam N Hornsby, Thomas Evans, Peter S Riefer, Rosie Prior, and Bradley C Love. Conceptual organization is revealed by consumer activity patterns. *Computational Brain & Behavior*, pages 1–12, 2019.

[33] Francisco JR Ruiz, Susan Athey, David M Blei, et al. Shopper: A probabilistic model of consumer choice with substitutes and complements. *Annals of Applied Statistics*, 14(1):1–27, 2020.

[34] Bruno Jacobs, Dennis Fok, and Bas Donkers. Understanding large-scale dynamic purchase behavior. *Marketing Science*, (0), 2020.

[35] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[36] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, 28(1):1–38, 2010.

[37] Jason Chuang, Margaret E Roberts, Brandon M Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. Topiccheck: Interactive alignment for assessing topic model stability. In *NAACL HLT'15*, pages 175–184, 2015.

[38] Nikolaos Aletras and Mark Stevenson. Measuring the similarity between automatically generated topics. In *ACL '14*, volume 2, pages 22–27, 2014.

[39] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *NIPS '09*, pages 288–296, 2009.

[40] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of LDA generative models. In *ECML PKDD '09*, pages 67–82. Springer, 2009.

[41] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *ICML '09*, pages 1105–1112. ACM, 2009.

[42] Wray Buntine. Estimating likelihoods for topic models. In *ACML '09*, pages 51–64. Springer, 2009.

[43] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *NAACL HLT' 10*, pages 100–108. Association for Computational Linguistics, 2010.

[44] Lan Du, Wray Buntine, and Huidong Jin. A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine learning*, 81(1):5–19, 2010.

[45] Changyou Chen, Lan Du, and Wray Buntine. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *ECML PKDD '11*, pages 296–311. Springer, 2011.

[46] Xing Wei, Jimeng Sun, and Xuerui Wang. Dynamic mixture models for multiple time-series. In *IJCAI '07*, volume 7, pages 2909–2914, 2007.

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

[48] Charles E Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

[49] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

[50] Thomas Minka. Estimating a dirichlet distribution, 2000.

[51] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[52] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

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

[54] Wray Buntine and Marcus Hutter. A Bayesian view of the Poisson-Dirichlet process. `https://arxiv.org/abs/1301.6705`.

[55] Yee Whye Teh. Dirichlet process. `http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/dp.pdf`.

[56] Sharon Goldwater, Mark Johnson, and Thomas L Griffiths. Interpolating between types and tokens by estimating power-law generators. In *NIPS '06*, pages 459–466, 2006.

[57] David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[58] Leetsch C Hsu and Peter Jau-Shyong Shiue. A unified approach to generalized Stirling numbers. *Advances in Applied Mathematics*, 20(3):366–384, 1998.

[59] Jordan Boyd-Graber, David Mimno, and David Newman. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and their Applications*, 225255, 2014.

[60] Jordan L Boyd-Graber, Yuening Hu, David Mimno, et al. *Applications of topic models*, volume 11. now Publishers Incorporated, 2017.

[61] David M Blei. Probabilistic topic models. *ACM*, 55(4):77–84, 2012.

[62] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *NIPS '05*, pages 1385–1392, 2005.

[63] David Blei and John Lafferty. Correlated topic models. *NIPS '06*, 18:147, 2006.

[64] Wei Li and Andrew McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML '06*, pages 577–584. ACM, 2006.

[65] Wei Li, David Blei, and Andrew McCallum. Nonparametric Bayes pachinko allocation. *arXiv preprint arXiv:1206.5270*, 2012.

[66] Do-kyum Kim, Geoffrey M Voelker, and Lawrence K Saul. Topic modeling of hierarchical corpora. `https://arxiv.org/abs/1409.3518`, 2014.

[67] Lan Du, Wray Buntine, Huidong Jin, and Changyou Chen. Sequential latent Dirichlet allocation. *Knowledge and Information Systems*, 31(3):475–503, 2012.

[68] David M Blei and John D Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120. ACM, 2006.

[69] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. `https://arxiv.org/abs/1206.3298`.

[70] Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *SIGKDD '06*, pages 424–433, 2006.

[71] David M Blei and Jon D McAuliffe. Supervised topic models. In *NIPS '07*, pages 121–128, 2007.

[72] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP'09*, pages 248–256. Association for Computational Linguistics, 2009.

[73] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI '04*, pages 487–494. AUAI Press, 2004.

[74] Li Fei-Fei and Pietro Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR '05*, volume 2, pages 524–531. IEEE, 2005.

[75] Marco Cristani, Alessandro Perina, Umberto Castellani, and Vittorio Murino. Geo-located image analysis using latent representations. In *CVPR '08*, pages 1–8. IEEE, 2008.

[76] Yang Wang and Greg Mori. Max-margin latent Dirichlet allocation for image classification and annotation. In *BMVC '11*, volume 2, page 7, 2011.

[77] Tomoyasu Nakano, Kazuyoshi Yoshii, and Masataka Goto. Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity. In *ICASSP '14*, pages 5202–5206. IEEE, 2014.

[78] Susana Zoghbi, Ivan Vulić, and Marie-Francine Moens. Latent dirichlet allocation for linking user-generated content and e-commerce data. *Information Sciences*, 367:573–599, 2016.

[79] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.

[80] David Hall, Dan Jurafsky, and Christopher D Manning. Studying the history of ideas using topic models. In *EMNLP '08*, pages 363–371, 2008.

[81] Bing Liu, Lin Liu, Anna Tsykin, Gregory J Goodall, Jeffrey E Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. Identifying functional miRNA–mRNA regulatory modules with correspondence latent Dirichlet allocation. *Bioinformatics*, 26(24):3105–3111, 2010.

[82] Căcilia Zirn and Heiner Stuckenschmidt. Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90:38–53, 2014.

[83] David Mimno. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(1):1–19, 2012.

[84] Stacy K Lukins, Nicholas A Kraft, and Letha H Etzkorn. Source code retrieval for bug localization using latent Dirichlet allocation. In *WCRE '08*, pages 155–164. IEEE, 2008.

[85] Stacy K Lukins, Nicholas A Kraft, and Letha H Etzkorn. Bug localization using latent Dirichlet allocation. *Information and Software Technology*, 52(9):972–990, 2010.

[86] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR '13*, pages 889–892, 2013.

[87] Yu Wang, Eugene Agichtein, and Michele Benzi. TM-LDA: efficient online modeling of latent topic transitions in social media. In *SIGKDD '12*, pages 123–131, 2012.

[88] Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. Empatweet: annotating and detecting emotions on Twitter. In *LREC '12*, volume 12, pages 3806–3813. Citeseer, 2012.

[89] Filippo Valle, Matteo Osella, and Michele Caselle. A topic modeling analysis of TCGA breast and lung cancer transcriptomic data. *Cancers*, 12(12):3799, 2020.

[90] David Andrzejewski. Modeling protein–protein interactions in biomedical abstracts with latent Dirichlet allocation, 2006.

[91] Hongning Wang, Minlie Huang, and Xiaoyan Zhu. Extract interaction detection methods from the biological literature. *BMC bioinformatics*, 10(1):1–13, 2009.

[92] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent Dirichlet allocation. *NIPS '10*, 23:856–864, 2010.

[93] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *UAI' 02*, UAI'02, page 352–359. Morgan Kaufmann Publishers Inc., 2002.

[94] Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In *NIPS '09*, pages 1973–1981, 2009.

[95] Hanna M Wallach. *Structured topic models for language*. PhD thesis, University of Cambridge, 2008.

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

[97] Harald Hruschka. Comparing unsupervised probabilistic machine learning methods for market basket analysis. *Review of Managerial Science*, pages 1–31, 2019.

[98] Fanglin Chen, Xiao Liu, Davide Proserpio, Isamar Troncoso, and Feiyu Xiong. Studying product competition using representation learning. In *SIGIR '20*, SIGIR '20, page 1261–1268, New York, NY, USA, 2020. Association for Computing Machinery.

[99] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[100] Gilles Celeux. Bayesian inference for mixture: The label switching problem. In *Compstat*, pages 227–232. Springer, 1998.

[101] Matthew Stephens. Bayesian methods for mixtures of normal distributions. 1997.

[102] Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

[103] Merrilee Hurn, Ana Justel, and Christian P Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.

[104] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, 6:355–378, 2019.

[105] Ajay Jasra, Chris C Holmes, and David A Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20:50–67, 2005.

[106] Matthew Sperrin, Thomas Jaki, and Ernst Wit. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3):357–366, 2010.

[107] Stuart J Blair, Yaxin Bi, and Maurice D Mulvenna. Increasing topic coherence by aggregating topic models. In *KDD'16*, pages 69–81, 2016.

[108] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *EMNLP '11*, pages 262–272. Association for Computational Linguistics, 2011.

[109] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *GSCL '09*, pages 31–40, 2009.

[110] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL '14*, pages 530–539, 2014.

[111] David Newman, Edwin V Bonilla, and Wray Buntine. Improving topic coherence with regularized topic models. In *NIPS '11*, pages 496–504, 2011.

[112] Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *IWCS '13*, pages 13–22, 2013.

[113] Allison June-Barlow Chaney and David M Blei. Visualizing topic models. In *ICWSM '12*, 2012.

[114] Matt Taddy. On estimation and selection for topic models. In *AISTATS'12*, pages 1184–1193, 2012.

[115] Jason Chuang, Christopher D Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *AVI '12*, pages 74–77. ACM, 2012.

[116] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *ACL '14*, pages 63–70, 2014.

[117] Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen. Mining common topics from multiple asynchronous text streams. In *WSDM '09*, pages 192–201. ACM, 2009.

[118] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(Aug):1801–1828, 2009.

[119] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *CIKM '09*, pages 957–966, 2009.

[120] Linzi Xing and Michael J Paul. Diagnosing and improving topic models by analyzing posterior variability. In *AAAI '18*, 2018.

[121] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, 2013.

[122] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Lucia Del Prete. Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Science*, 8(1):14, 2019.

[123] Farshideh Einsele, Leila Sadeghi, Rolf Ingold, and Helena Jenzer. A study about discovery of critical food consumption patterns linked with lifestyle diseases using data mining methods. In *BIOSTEC '15*, pages 239–245. SCITEPRESS - Science and Technology Publications, Lda, 2015.

[124] Xia Wang, Yingying Ouyang, Jun Liu, Minmin Zhu, Gang Zhao, Wei Bao, and Frank B Hu. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. *BMJ*, 349:44–90, 2014.

[125] J Wardle. Eating behaviour and obesity. *Obesity Reviews*, 8:73–75, 2007.

[126] Angela Groves. The local and regional food opportunity. *Institute of Grocery Distribution*, 2005.

[127] Sharron Kuznesof, Angela Tregear, and Andrew Moxey. Regional foods: a consumer perspective. *British Food Journal*, 99(6):199–206, 1997.

[128] Charlotte Sturley, Andy Newing, and Alison Heppenstall. Evaluating the potential of agent-based modelling to capture consumer grocery retail store choice behaviours. *The International Review of Retail, Distribution and Consumer Research*, 28(1):27–46, 2018.

[129] Alec Davies, Les Dolega, and Daniel Arribas-Bel. Buy online collect in-store: exploring grocery click&collect using a national case study. *International Journal of Retail & Distribution Management*, 2019.

[130] Andy Newing, Graham P Clarke, and Martin Clarke. Developing and applying a disaggregated retail location model with extended retail demand estimations. *Geographical Analysis*, 47(3):219–239, 2015.

[131] Thomas BP Waddington, Graham P Clarke, Martin Clarke, and Andy Newing. Open all hours: spatiotemporal fluctuations in UK grocery store sales

and catchment area demand. *The International Review of Retail, Distribution and Consumer Research*, 28(1):1–26, 2018.

[132] Tom Berry, Andy Newing, Deborah Davies, and Kirsty Branch. Using workplace population statistics to understand retail store performance. *The International Review of Retail, Distribution and Consumer Research*, 26(4):375–395, 2016.

[133] Nadine Schröder. Using multidimensional item response theory models to explain multi-category purchases. *Journal of Research and Management*, 39(2):27–37, 2017.

[134] Mariflor Vega-Carrasco, Jason O'sullivan, Rosie Prior, Ioanna Manolopoulou, and Mirco Musolesi. Modelling grocery retail topic distributions: Evaluation, interpretability and stability. `https://arxiv.org/abs/2005.10125`.

[135] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. Finding expert users in community question answering. In *WWW '12*, pages 791–798, 2012.

[136] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. Multi-aspect sentiment analysis with topic models. In *IEEE'11*, pages 81–88, 2011.

[137] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data.* CRC Press, 2014.

[138] Noel Cressie and Christopher K Wikle. *Statistics for Spatio-Temporal Data.* John Wiley & Sons, 2015.

[139] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning.* MIT Press Cambridge, MA, 2006.

[140] Office for National Statistics. Postcode lookup May 2019. `https://geoportal.statistics.gov.uk/datasets/ons::`

national-statistics-postcode-lookup-may-2019/about. Visited on: 2021-06-14.

[141] Waldo R Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1):234–240, 1970.

[142] C Carl Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[143] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

[144] Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. https://arxiv.org/abs/1701.02434.

[145] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[146] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[147] Food seasonal calendar. Food seasonal calendar. https://www.bbcgoodfood.com/seasonal-calendar. Visited on: 2021-06-14.

[148] Food glossary. Food glossary. https://www.bbcgoodfood.com/glossary. Visited on: 2021-06-14.