Received: 28 March 2020 | Accepted: 26 January 2022

DOI: 10.1111/rssc.12546

ORIGINAL ARTICLE

Posterior summaries of grocery retail topic models: Evaluation, interpretability and credibility

Mariflor Vega Carrasco[1] | **Ioanna Manolopoulou**[1] | **Jason O'Sullivan**[2] | **Rosie Prior**[2] | **Mirco Musolesi**[1]

[1]University College London, London, UK

[2]Dunnhumby Ltd, London, UK

Correspondence
Mariflor Vega Carrasco, University College London, London, UK.
Email: mariflor.vega.15@ucl.ac.uk

Abstract

Understanding the shopping motivations behind market baskets has significant commercial value for the grocery retail industry. The analysis of shopping transactions demands techniques that can cope with the volume and dimensionality of grocery transactional data while delivering interpretable outcomes. Latent Dirichlet allocation (LDA) allows processing grocery transactions and the discovering of customer behaviours. Interpretations of topic models typically exploit individual samples overlooking the uncertainty of single topics. Moreover, training LDA multiple times show topics with large uncertainty, that is, topics (dis)appear in some but not all posterior samples, concurring with various authors in the field. In response, we introduce a clustering methodology that post-processes posterior LDA draws to summarise topic distributions represented as recurrent topics. Our approach identifies clusters of topics that belong to different samples and provides associated measures of uncertainty for each group. Our proposed methodology allows the identification of an unconstrained number of customer behaviours presented as recurrent topics. We also establish a more holistic framework for model evaluation, which assesses topic models based not only on their predictive likelihood but also on quality aspects such as coherence and distinctiveness of single topics and credibility of a set of topics. Using the outcomes of a tailored survey, we set thresholds that aid in interpreting quality aspects in grocery retail data. We demonstrate that selecting recurrent topics not only improves predictive likelihood but also outperforms interpretability and credibility. We illustrate our methods with an example from a large British supermarket chain.

KEYWORDS

customer behaviours, grocery shopping motivations, latent Dirichlet allocation, topic credibility, topic distinctiveness, topic model evaluation

1 | INTRODUCTION

In the grocery retail industry, millions of transactions are generated every day by customers that choose and buy products to fulfil one or more needs. Transactions contain few products out of thousands of available items, reflecting the unseen customer behaviours. For instance, customers go to the grocery retailers to buy foods for breakfast, ingredients to cook a roast dinner or popular products for a barbecue. 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.

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 customer while non-linear models, in general, are difficult to interpret (Ramon et al., 2020).

Topic modelling (TM) offers a new scalable statistical framework that can process large volumes of transactions while maintaining the explanatory power to discover, analyse and understand customer behaviours. Latent Dirichlet allocation (LDA) (Blei et al., 2003), the vanilla topic model, was originally introduced to uncover topics that summarise the semantic structure in a large collection of text data. In the retail context, LDA facilitates interpretations of customer behaviours and provides a simple model to summarise a sheer volume of transactions. Topics, which are distributions over a product assortment, reveal products that are frequently bought together to fulfil a specific need. Transactions are then no longer summarised by individual items but as mixtures of customer behaviours.

LDA provides a simple and interpretable model; however, the inference is computationally intractable. There are various approaches for estimating LDA's posterior distribution, such as gradient descent (Hoffman et al., 2010), Gibbs sampling (Griffiths & Steyvers, 2004), variational inference (Blei et al., 2003), and expectation propagation (Minka & Lafferty, 2002). In this paper, we use the collapsed Gibbs sampling (Griffiths & Steyvers, 2004), a Markov chain Monte Carlo algorithm, to sample from the posterior distribution and learn topic distributions since this method has shown coherent and useful topics with potential commercial value in the field of our application.

Applications of LDA for exploratory purposes usually relies on one posterior sample, ignoring variability within topic distributions. Since the likelihood of a topic model is, in essence, a mixture model, there is no guaranteed correspondence between individual topics across samples (Griffiths & Steyvers, 2004), akin to the label-switching problem. Thus, a posterior summary that exploits multiple posterior samples to characterise the posterior distribution requires a relabelling algorithm to ensure identifiability between topics. Various relabelling methodologies assume one-to-one matches, that is, algorithms that minimise an overall loss function (Celeux, 1998; Stephens, 2000; Stephens & Phil, 1997), with identifiability constraints (McLachlan et al., 2019), with probabilistic approach (Jasra et al., 2005; Sperrin et al., 2010). However, one-to-one matches across samples may merge topics with a large distributional dissimilarity. This compromises the meaning behind topic distributions disrupting the interpretations of customer behaviours.

In addition, applications of LDA may exhibit significant variations across iterations of the same model as observed in (Chuang et al., 2015; Rosen-Zvi et al., 2010; Steyvers & Griffiths, 2007). For instance, a topic associated with a customer behaviour may appear in 8 of 10 model iterations showing some uncertainty. Note that we call a model iteration to an MCMC chain, while posterior samples are taken from one MCMC chain. Empirically, we have found that topics show useful customers behaviours with various levels of uncertainty. Thus, applications of LDA require the analysis over multiple model iterations otherwise insightful topics may be overlooked.

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 (model iterations). 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 (Aletras & Stevenson, 2014). 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*, that is, the number of topics within the same cluster (the number of posterior samples exhibiting the same topic). Depending on the domain of interest, users can set thresholds of minimum recurrence to select clustered topics of low uncertainty. Hierarchical clustering has been used previously to interactively align topics (Chuang et al., 2015) and to aggregate topic models with small and large numbers of topics (Blair et al., 2016). In comparison to these works, we aim to identify topics that illustrate different customer behaviours while measuring their uncertainty.

LDA is the topic model with the largest number of applications (Boyd-Graber et al., 2014; Jelodar et al., 2019), however, various authors have pointed out some flaws on inferred topics. For example, topics may not correspond to genuine and meaningful themes (AlSumait et al., 2009), affecting the user's confidence in the application of the topic model (Mimno et al., 2011). 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 (Boyd-Graber et al., 2014). And as mentioned before, topics may also show significant variations across multiple model iterations. Evaluation of topics models should account for quality aspects that favour models with larger interpretability, distinctiveness and credibility.

Evaluation of topic models is typically based on model fit metrics such as held-out-likelihood or perplexity (Buntine, 2009; Wallach et al., 2009b) that measure the generalisation capability by computing the model likelihood on unseen data. These held-out metrics may be suitable for applications that ultimately aim to predict new topical mixtures, but they are not sufficient for applications where the value lies on the topics themselves. As pointed by Chang et al. (2009), held-out metrics may lead to topic models with less semantically meaningful topics; thereby, disagreeing with human annotators that would prefer topic models with interpretable topics. Thus, evaluation of topic models should be more holistic assessing model generalisation along with the aforementioned quality aspects.

Topic coherence Newman et al. (2010) measures the interpretability of individual topics, typically quantified by co-occurrence metrics such as pointwise mutual information (PMI) and normalised pointwise mutual information (NPMI) (Bouma, 2009). NPMI and PMI have been shown to correlate with human annotators in Newman et al. (2010) and Lau et al. (2014). Various methods have been proposed to improve topic coherence. For example, Wallach et al. (2009a) used asymmetric priors over document distributions to capture highly frequent terms in few topics; Newman et al. (2011) proposed the applications of regularisation methods; Mimno et al. (2011) 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.

Topic distinctiveness and topic credibility measure the semantic dissimilarity among topics of the same posterior sample and the distributional similarity among posterior samples (from multiple MCMC chains) respectively. 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 from MCMC chains, topic credibility is defined as the average maximum cosine similarity; where the maximum cosine similarity is with respect to 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 (Aletras & Stevenson, 2014). Thus, high-quality topics are not only coherent but also distinctive among them and identifiable in other posterior samples.

In a nutshell, we propose a post-processing methodology to summarise topical posterior distribution and a more holistic framework for evaluating topic models. We demonstrate our methods using a large collection of transactions from a major retailer in the United Kingdom and identify customer behaviours. 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. 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.

This paper is organised as follows: we discuss related work in Section 2. LDA is described in Section 3. Section 4 presents the definitions of model generalisation, topic coherence, topic distinctiveness and topic credibility. Section 5 introduces our proposed methodology for clustering and selecting recurrent topics. Sections 6–8 show the application of grocery retail data from a major retailer in the United Kingdom. More specifically, Section 6 discusses thresholds for interpretability and similarity obtained from a user study with experts in grocery retail analytics and exhibits the pitfalls of LDA topics. Section 7 demonstrates the advantages of selecting clustered topics of high posterior recurrence. Section 8 displays identified grocery topics and indicates commercial implications in the grocery retail sector. Finally, we summarise our findings in Section 9.

2 | RELATED WORK

Topic modelling, in particular LDA, has already been used to identify latent shopping motivations in retail data. For instance, Christidis et al. (2010) applied LDA to grocery transactions from a major European supermarket to identify latent topics of product categories, intending to support an item recommendation system. In this study, 102 thousand unique products were aggregated into 473 synthetic categories with no distinction between brands or package sizes. Hruschka (2014) sketched the core of a recommender system to illustrate the managerial relevance of estimated topics, which were obtained from training LDA and the correlated topic model on market baskets from a medium-sized German supermarket. The study only accounted for the 60 product categories with the highest univariate purchase frequencies. Jacobs et al. (2016) 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. Again, the authors aggregated products to a category-brand level, that is, different fragrances/flavours of the same product and brand are aggregated into one category, reducing more than 3 thousand unique products to 394 categories. Hruschka (2016, 2021) compared topic models 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 collections of product categories and not the full product resolution, thereby reducing the dimensionality of the problem. Hornsby et al. (2020) provided a direct application of a 25-topic LDA model on transactional data from a major British retailer to identify shopping goals.

Beyond LDA, other approaches have been applied to market baskets. For instance, Schröder (2017) applied Multidimensional Item Response Theory (MIRT) models on a limited data set with 31 product categories collected by a house panel from a single supermarket in the United States, and found that MIRT models outperformed LDA according to the Akaike information criterion and its corrected form. MIRT may be an option to analyse small data sets of discrete grouped data. Hruschka (2016, 2021) also compared topic models such as LDA and correlated topic model (CTM) to alternative methods such as binary factor analysis, restricted Boltzmann machine (RBF) and deep belief net (DBN). It was shown that the alternative methods outperform topic models in model generalisation. However, the number of topics was restricted to a range from 2 to 6, while networks of much larger architectures were explored. Moreover, the DBN and RBF outcomes are far less interpretable than LDA topics. Ruiz et al. (2020) introduced 'SHOPPER', a sequential probabilistic model, that captures interaction among items and answers counterfactual queries about changes in prices. Chen et al. (2020) 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. Jacobs et al. (2020) combined the correlated topic model with the vector autoregression to account for product, customer and time dimensions present in purchase history data.

Within LDA, various methods have been proposed to improve topic coherence. For example, Wallach et al. (2009a) used asymmetric priors over document distributions to capture highly frequent terms in few topics; Newman et al. (2011) introduced two regularisation methods, and Mimno et al. (2011) generalised the Pòlya urn model aiming to reduce the number of low-quality topics. In this paper, we do not try to improve LDA to render more coherent topics, but we will show that our proposed methodology retrieves groups of clustered topics with higher coherence.

Hierarchical clustering has been used previously to interactively align topics (Chuang et al., 2015) and to aggregate topic models (Blair et al., 2016). The former work assumes that topics align with up to one topic from a different posterior sample. The latter work merges topics from posterior samples with small and large numbers of topics aiming to improve topic coherence. However, these works do not assess other aspects of topic quality, such as topic distinctiveness and topic credibility nor consider the likelihood of the resulting models.

With regards to the label-switching problem, which also affects LDA since it is inherently a mixture model, Stephens (2000), Celeux (1998) and Stephens and Phil (1997) developed relabelling algorithms to perform a k-means type clustering of the MCMC samples. Hastie et al. (2015) followed a k-medoid strategy to obtain an optimal partition that takes advantage of the whole MCMC output rather than taking a maximum a posteriori partition. Other relabelling strategies consider label invariant loss functions (Celeux et al., 2000; Hurn et al., 2003), identifiability constraints (McLachlan et al., 2019) and probabilistic relabelling (Jasra et al., 2005; Sperrin et al., 2010). Note that these techniques assume that topics are present (but with switched labels) across samples. Thus, we cannot use relabelling techniques to summarise topic models, since topics may (dis)appear across a Markov chain. Instead, we propose a methodology to group topics using similarity measures.

3 | LATENT DIRICHLET ALLOCATION

Here, we interpret LDA (Blei et al., 2003) in terms of retail data, where transactions are interpreted as *bags of products*. This is a natural assumption of in-store transactions where products are registered without an inherited order. In addition, transactions are assumed to be independent and exchangeable, so metadata such as timestamps and coordinate location are disregarded.

Within the LDA framework, transactions are represented as mixtures over a finite number of topics K and topics are distributions over products from a fixed product assortment of size V. More formally, LDA is a generative process in which the topics $\Phi = [\phi_1, \dots, \phi_K]$ are sampled from a Dirichlet distribution governed by hyperparameters $\boldsymbol{\beta} = [\beta_1, \dots, \beta_V]$ and the topical mixtures $\Theta = [\theta_1, \dots, \theta_D]$ are sampled from a Dirichlet distribution governed by hyperparameters $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_K]$. For each transaction (equivalent to a basket) d, product i is sampled through a two-step process. First, a topic assignment $z_{i,d}$ is chosen from the transaction-specific topical mixture θ_d. Second, a product is sampled from the assigned topic $\phi_{z_{i,d}}$. Mathematically,

$$\begin{aligned} \phi_k &\sim Dirichlet(\boldsymbol{\beta}) \\ \theta_d &\sim Dirichlet(\boldsymbol{\alpha}) \\ z_{i,d}|\theta_d &\sim Multinomial(\theta_d) \\ w_{i,d}|\phi_{z_{i,d}} &\sim Multinomial(\phi_{z_{i,d}}). \end{aligned} \tag{1}$$

The data then correspond to the observed set of products $w_{i,d}$ within each transaction d. The posterior distribution of the topic distributions Φ and topical mixtures $\mathit{\Theta}$ are given by the posterior conditional probability:

$$P(\Phi, \Theta, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\Phi, \Theta, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}, \tag{2}$$

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

There are various approaches for estimating LDA's posterior distribution, such as gradient descent (Hoffman et al., 2010), Gibbs sampling (Griffiths & Steyvers, 2004), variational inference

(Blei et al., 2003) and expectation propagation (Minka & Lafferty, 2002). Empirically, we have found that the variational Bayes (VB) algorithm of (Blei et al., 2003) leads to the lower interpretability of learnt topics show, that is, showing products that do not convey a clear shopping purpose. On the other hand, Gibbs sampler (GS) requires longer training times, but learns more coherent and useful topics, that is, showing products that can be easily associated with a shopping purpose. In this paper, we chose the Gibbs sampler. Although more recent VB developments using amortised inference such as Srivastava and Sutton (2017) may be a promising alternative, they are beyond the scope of this work.

In this paper, we used LDA with symmetric Dirichlet priors governed by a scalar concentration parameter and a uniform base measure, so that topics are equally likely a priori. Wallach et al. (2009a) showed that an optimised asymmetric Dirichlet prior over topical mixtures improves model generalisation and topic interpretability by capturing highly frequent terms in a few topics. However, we empirically found that LDA with an asymmetric prior may lead to poor convergence of the Gibbs sampler in the context of our application. On the other hand, a fixed symmetric prior not only has shown satisfactory mixing of MCMC chains but also coherent topics that have been acknowledged by experts in the field of our application.

3.1 | Gibbs sampling

The Gibbs sampling algorithm starts with a random initialisation of topic assignments **z** to values $1, 2, \ldots, K$. In each iteration, topic assignments are sampled from the full conditional distribution, defined as:

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{N_{k,v}^{-i} + \beta_v}{N_k^{-i} + \beta} \frac{N_{d,k}^{-i} + \alpha_k}{N_d^{-i} + \alpha}, \tag{3}$$

where the notation N^{-i} is a count that does not include the current assignment of z_i. $N_{k,v}$ is the number of assignments of product v to topic k. $N_{d,k}$ is the number of assignments of topic k in transaction d. N_k is the total number of assignments of topic k. N_d is the size of transaction d. $\alpha = \sum_k^K \alpha_k$ and $\beta = \sum_v^V \beta_v$. This full conditional distribution can be interpreted as the product of 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.

After a burn-in period, states of the Markov chain (topic assignments) are recorded with 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 their conditional posterior means:

$$\hat{\phi}_{k,v}^s = E(\phi_{k,v}^s|\mathbf{z}^s, \boldsymbol{\beta}) = \frac{N_{k,v}^s + \beta_v^s}{N_k^s + \beta^s}, \quad k = 1 \ldots K, v = 1 \ldots V, \tag{4}$$

$$\hat{\theta}_{d,k}^s = E(\theta_{d,k}^s|\mathbf{z}^s, \boldsymbol{\alpha}) = \frac{N_{d,k}^s + \alpha_k^s}{N_d^s + \alpha^s}, \quad d = 1 \ldots D, k = 1 \ldots K. \tag{5}$$

4 | TOPIC MODEL EVALUATION

Topic model evaluation is typically based on model fit metrics such as held-out-likelihood or perplexity (Buntine, 2009; Wallach et al., 2009b), 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 (Chang et al., 2009). The evaluation of topic models should therefore not be exclusively based on likelihood metrics, but also include topic quality metrics such as topic coherence, 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 paper.

4.1 | Model generalisation

Model fit metrics such as perplexity or held-out-likelihood of unseen documents (transactions) estimate the 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,

$$\text{Perplexity} = -\frac{\log P(\mathbf{w}'|\Phi, \boldsymbol{\alpha})}{N'}, \tag{6}$$

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 (Buntine, 2009; Wallach et al., 2009b). In this paper, we use the left-to-right algorithm with 30 particles to approximate the log-likelihood on held-out documents (Wallach, 2008; Wallach et al., 2009b). 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 | Topic coherence

A topic is said to be coherent when its most likely terms can be interpreted and associated with a single semantic concept (Newman et al., 2010). 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 normalised pointwise mutual information (NPMI) (Bouma, 2009) shown in Newman et al. (2010) and Lau et al. (2014). PMI measures the probability of seeing two products within the same topic in comparison to the probability of seeing them individually. NPMI standardises PMI, providing a score in the range of $[-1, 1]$. NPMI towards 1 corresponds to high co-occurrence.

$$\text{PMI}(w_i, w_j) = \log\left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right); \quad i \neq j, \quad 1 \leq i, j \leq 15. \tag{7}$$

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log P(w_i, w_j)}; \quad i \neq j, \quad 1 \leq i, j \leq 15. \tag{8}$$

In the literature, average NPMI and PMI are usually measured using the top 10 terms Lau et al. (2014), Aletras and Stevenson (2013), Chaney and Blei (2012) and Newman et al. (2010). However, we choose to use 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 Taddy (2012), Chuang et al. (2012) and Sievert and Shirley (2014), 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 (Lau et al., 2014).

4.3 | 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 distinctive from each other. 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 the ‘flavoured sparkling water’ since both topics exhibit the same theme.

Several measures have been used to identify similar topics: KL-divergence (Li & McCallum, 2006; Newman et al., 2009; Wang et al., 2009), the average log odds ratio (Chaney & Blei, 2012), the cosine distance (Chuang et al., 2015; He et al., 2009; Ramage et al., 2009; Xing & Paul, 2018). Aletras and Stevenson (2014) and Xing and Paul (2018) showed that cosine distance outperforms other distributional similarity measures, such as KL-divergence, Jensen Shannon Divergence, Euclidean distance, 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 $\Phi^t \backslash \phi_i^t$ within the same posterior sample, denoted by :

$$\mathrm{CD}_{\min}\left(\phi_i^t, \Phi^t \backslash \phi_i^t\right) = \min[\mathrm{CD}(\phi_i^t, \phi_1^t), \dots, \mathrm{CD}(\phi_i^t, \phi_{i-1}^t), \mathrm{CD}(\phi_i^t, \phi_{i+1}^t), \dots, \mathrm{CD}(\phi_i^t, \phi_K^t)], \tag{9}$$

where

$$\mathrm{CD}\left(\phi_i, \phi_j\right) = 1 - \frac{\phi_i \cdot \phi_j}{||\phi_i||\,||\phi_j||}. \tag{10}$$

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.

4.4 | Topic credibility

When comparing different LDA posterior draws, 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 draw $s \neq t$ indicates uncertainty about ϕ_i^t. To measure topic credibility of topic ϕ_i^t in posterior draw t, we compute the maximum cosine similarity between ϕ_i^t and all topics within posterior draw Φ^s, for $s \neq t$, and average across all posterior draws $s \neq t$. If a topic is highly credible, then we expect a very similar topic to appear in every single posterior draw, 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,

$$\mathrm{CS}_{\max}\left(\phi_i^t, \Phi^s\right) = \max[\mathrm{CS}(\phi_i^t, \phi_1^s), \dots, \mathrm{CS}(\phi_i^t, \phi_K^s)], \tag{11}$$

where

$$\mathrm{CS}\left(\phi_i, \phi_j\right) = \frac{\phi_i \cdot\ \phi_j}{||\phi_i||\,||\phi_j||}. \tag{12}$$

Averaging across all other posterior draws,

$$\overline{\mathrm{CS}_{\max}}(\phi_i^t, \Phi^{1:S}) = \frac{\sum_{s \neq t} \mathrm{CS}_{\max}\left(\phi_i^t, \Phi_{.}^s\right)}{S - 1}. \tag{13}$$

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.

5 | POSTERIOR SUMMARY OF TOPIC DISTRIBUTIONS

Here 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 LDA 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 topical similarity. The resulting clusters are used to quantify topic posterior recurrence of a clustered topic, which is ultimately used to identify and filter out topics of high uncertainty.

We choose the hierarchical clustering method over other clustering techniques for three reasons: (a) hierarchical clustering automatically provides a solution for any desired number of clusters, allowing the user to interact with the clustering, (b) it is more flexible at allowing the user to set different distance thresholds, (c) it gave consistently sensible and transparent results in our experiments. In addition, clustering algorithms that require specifying a number of clusters a priori showed lower coherence, since topics are forced to be merged even when they do not share a similar theme.

5.1 | 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 (Aletras & Stevenson, 2014) and human rating of posterior variability (Xing & Paul, 2018). We opt for the *average* linkage method since, empirically, it has worked better than *single* and *complete* linkage methods, that is, 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.

5.2 | Recurrent topics

The AHC retrieves a collection of clusters $C_1, \dots, C_N$, which are represented by a *clustered topic* $\overline{\phi_k}$ with a *cluster size* $|C_k|$, where $k = 1, \dots N$. The clustered topic is the average distribution of the topics that share the same membership. The cluster size is the number of members, for example, 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 large cluster size, indicating that the topic consistently reappears across multiple samples. Thus, we measure the recurrence of a topic by its cluster size:

$$\text{recurrence}(\overline{\phi_i}) = |C_i|. \tag{14}$$

Then, 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.

6 | APPLICATION TO GROCERY RETAIL DATA

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 United Kingdom. Transactions are sampled randomly, covering 100 nationwide superstores between September 2017 and August 2018. The training data set contains 36 thousand 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 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 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. We disregard transactions with fewer than three products assuming that smaller transactions do not have enough products to exhibit a customer need. No personal customer data were used for this research.

6.1 | Human judgement on interpretability and similarity of topics

To aid interpretation of topics within the context of the application, meaningful NPMI and cosine similarity thresholds need to be set. 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 hyper-parameters $\alpha = [0.1, 0.01]$ 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. One hundred and eighty-nine and 935 evaluations for topic distinctiveness and topic coherence were collected respectively.

Figure 1a 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 1b compares human judgment on topic similarity against cosine distance. Unsurprisingly, the lower the cosine distance, the more similar the topic distributions are. We observe that 70% of the pairs with CD ≤ 0.1 were interpreted as 'Similar' or 'Highly similar', and 95%

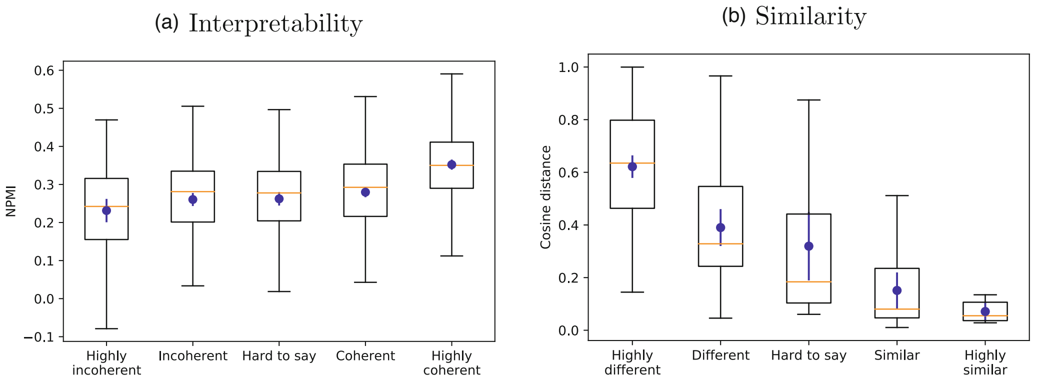


FIGURE 1 Human evaluation on interpretability of individual topics and similarity between pairs of topics. Figure 1a shows coherence scores against topic normalised pointwise mutual information (NPMI). Figure 1b shows similarity scores against the cosine distance between compared topic distributions. 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 [Colour figure can be viewed at wileyonlinelibrary.com]

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 ≥ 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.

6.2 | LDA performance

We trained five LDA models with $K = 25, 50, 100, 200, 400$ topics, with a symmetric Dirichlet hyperparameters $\alpha_k = 3/K$ and $\beta_v = 0.01$. Note that $\sum_k alpha_k = 3$, which reflects the minimum transaction size. $\beta_v = 0.01$ is commonly used in the literature (Mimno et al., 2011; Newman et al., 2011). For each model, four Markov chains are run for 50,000 iterations with a burn-in of 30,000 iterations; samples were recorded every 10 000 iterations obtaining 20 samples in total. As shown in Appendix A, convergence of the Markov chains is satisfactory.

LDA models are assessed on the four aforementioned quality aspects. 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 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 3a posterior draws with 25 and 50 topics show larger average NPMI, however, 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 Chang et al. (2009), posterior samples with higher coherence do not necessarily have the best likelihood, which is the case of 25-topic LDA samples. Figure 3b illustrates two topics with low/high coherence. 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 that are frequently bought together due to promotional discounts.

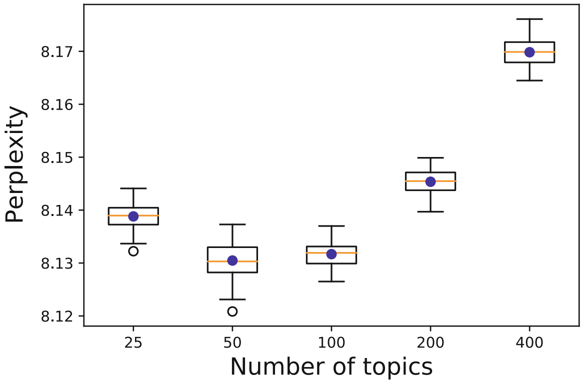
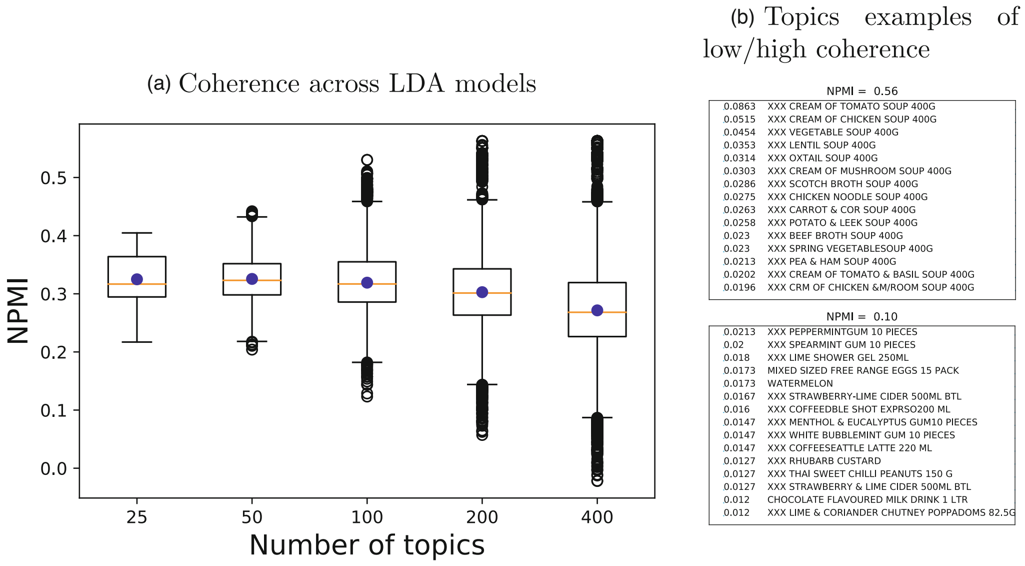


FIGURE 2 The perplexity of latent Dirichlet allocation models with 25/50/100/200/400 topics. Each boxplot represents the perplexity distribution over the 20 samples. Blue circles indicate the average perplexity; standard errors are smaller than the marker size [Colour figure can be viewed at wileyonlinelibrary.com]



(a) Coherence across LDA models

(b) Topics examples of low/high coherence

NPMI = 0.56

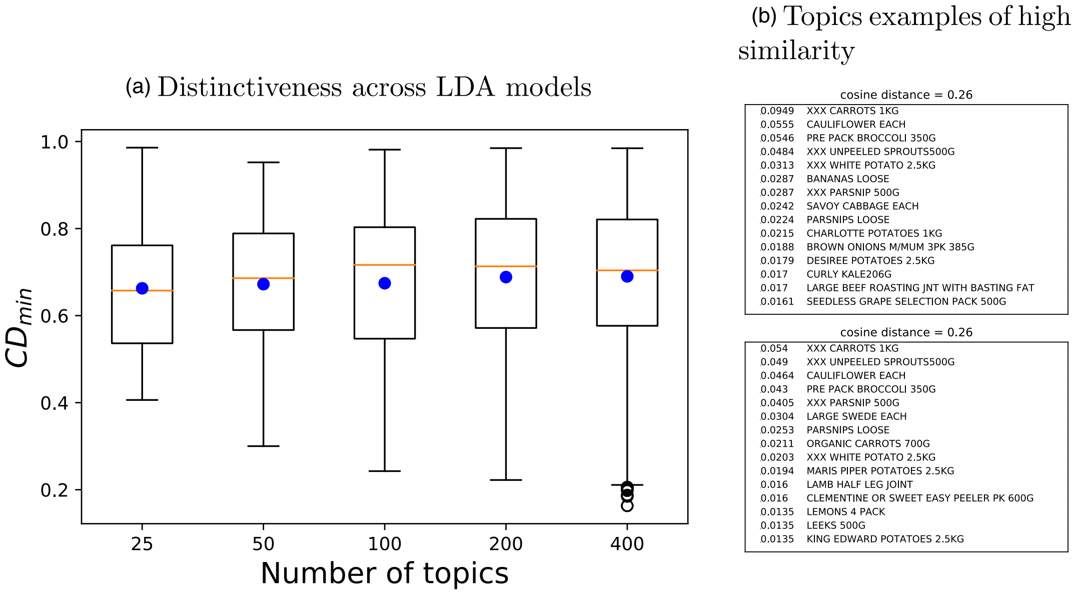
0.0863	XXX CREAM OF TOMATO SOUP 400G
0.0515	XXX CREAM OF CHICKEN SOUP 400G
0.0454	XXX VEGETABLE SOUP 400G
0.0353	XXX LENTIL SOUP 400G
0.0314	XXX OXTAIL SOUP 400G
0.0303	XXX CREAM OF MUSHROOM SOUP 400G
0.0286	XXX SCOTCH BROTH SOUP 400G
0.0275	XXX CHICKEN NOODLE SOUP 400G
0.0263	XXX CARROT & COR SOUP 400G
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

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

FIGURE 3 (3a) Topic-specific normalised pointwise mutual information (NPMI) of 25/50/100/200/400-topic latent Dirichlet allocation model. Blue circles indicate the average NPMI; standard errors are smaller than the marker size. (3b) shows (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 [Colour figure can be viewed at wileyonlinelibrary.com]

In Figure 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 $\geq$ 0.5) within their posterior draw. However, as expected, the larger the model, the more topics with some degree of similarity (CD $\leq$ 0.3) as seen in LDA models with 100 to 400 topics. Figure 4b shows an example of two topics with some degree of similarity, both show collections of produce and red meat.



(a) Distinctiveness across LDA models

(b) Topics examples of high similarity

cosine distance = 0.26

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

cosine distance = 0.26

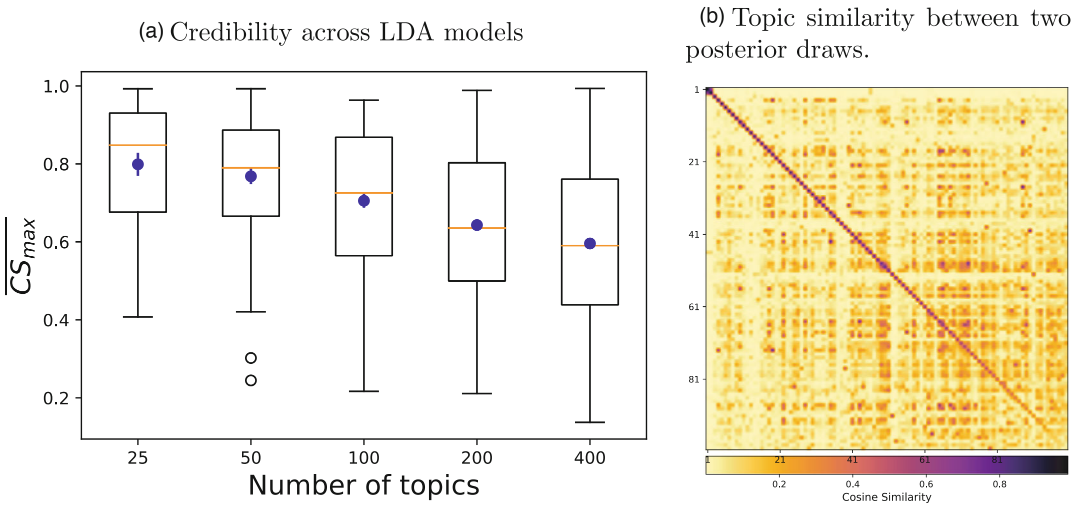
0.054	XXX CARROTS 1KG
0.049	XXX UNPEELED SPROUTS500G
0.0464	CAULIFLOWER EACH
0.043	PRE PACK BROCCOLI 350G
0.0405	XXX PARSNIP 500G
0.0304	LARGE SWEDE EACH
0.0253	PARSNIPS LOOSE
0.0211	ORGANIC CARROTS 700G
0.0203	XXX WHITE POTATO 2.5KG
0.0194	MARIS PIPER POTATOES 2.5KG
0.016	LAMB HALF LEG JOINT
0.016	CLEMENTINE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

FIGURE 4 (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. (4b) shows 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 [Colour figure can be viewed at wileyonlinelibrary.com]

In Figure 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 each sample tends to 0, so does its average maximum cosine similarity. We observe up to 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 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 (Rosen-Zvi et al., 2010). This plot indicates that around one-fifth of the topics do not appear with some similarity CS ≥ 0.5 in the other posterior draw. This implies that applying label-switching algorithms to resolve labelling for each posterior draw would inevitably ‘match-up’ topics which 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.

7 | CLUSTERING AND SELECTION OF RECURRENT TOPICS

In this section, we apply our methodology to summarise LDA posterior 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. Since our goal is not to predict new baskets, but


(a) Credibility across LDA models

(b) Topic similarity between two posterior draws.

FIGURE 5 (5a) Topic-specific average maximum cosine similarity. For each topic, the maximum cosine similarity is calculated over the topics of a different posterior draw from two MCMC chains. 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. (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 [Colour figure can be viewed at wileyonlinelibrary.com]

to understand the customer motivations; we do not aim to maximise perplexity, but to identify recurrent and coherent topics while preserving reasonable perplexity.

We conduct three experiments with LDA samples with 50, 100 and 200 topics. In each experiment, a bag of topics is formed from 20 samples that come from four separate Gibbs samplers. From each chain, samples are obtained after a burn-in period (30,000 iterations) and recorded every 5000 iterations to reduce autocorrelation. Computing perplexity is a computationally expensive. Thus, we do not record the evaluation metrics at each clustering step. Instead, we evaluate subsets of clustered topics obtained at different distance thresholds (cosine distance from 0 to 0.55 and every 0.05). We assume that topics with cosine distance ≥0.55 are too different, which would render cluster merging meaningless. 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 topic and the latter show worse perplexities.

Figure 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 the 5%, 25%, 50% and 100% of the samples. As observed in the perplexity plot (top left), the subset with the lowest perplexity is the one at minimum cluster size 1 and cosine distance 0, this is the original bag of 1000 topics before merging. This subset has the lowest performance in distinctiveness; thereby, 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 the topics that have not been clustered. Since a symmetric prior is used to compute perplexity, the

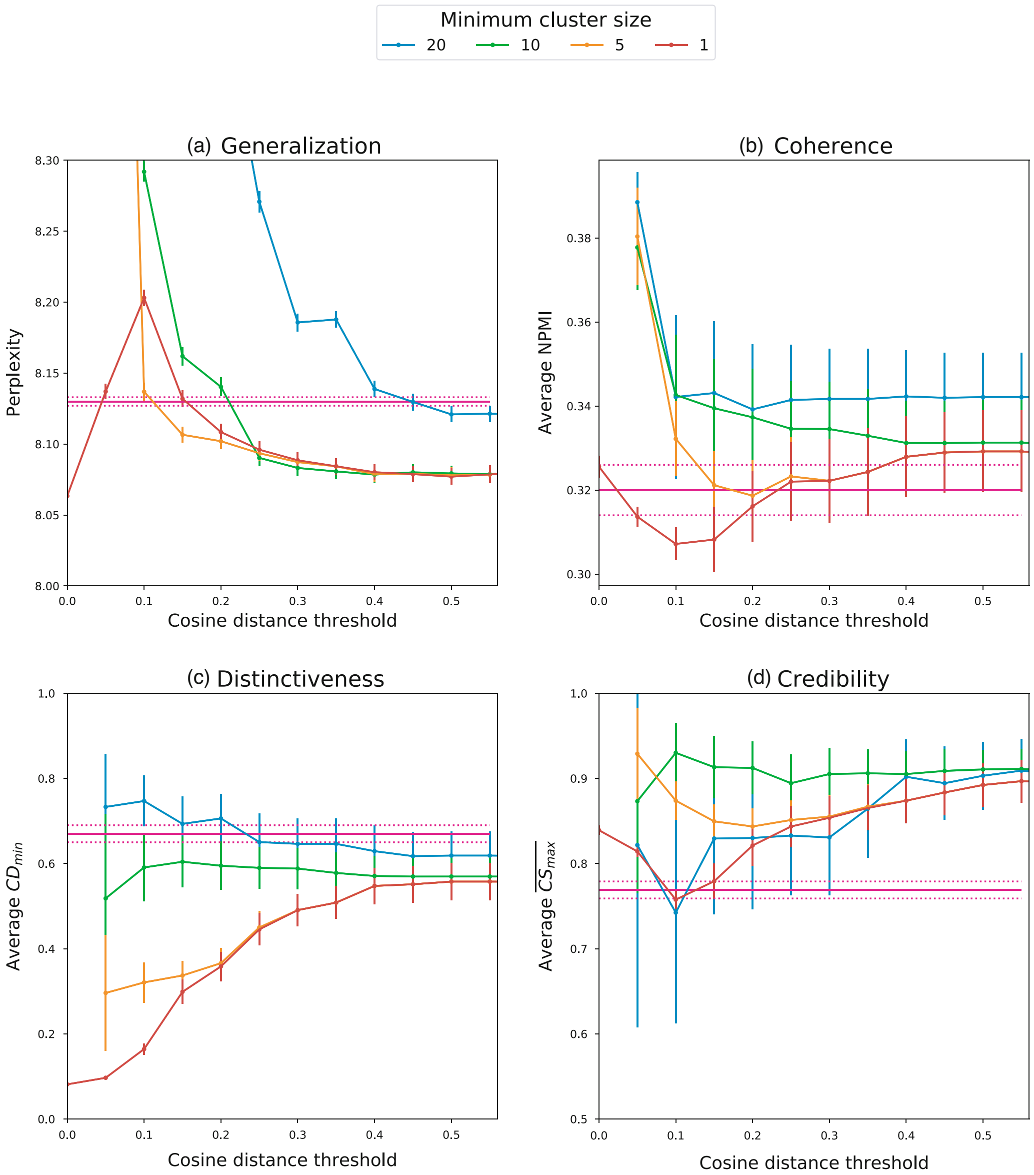


FIGURE 6 Subset evaluation using cosine distance (varying from 0 to 1 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). Clustered topics were obtained from clustering 20 samples of latent Dirichlet allocation (LDA) with 50 topics. Vertical lines represent one standard error. Magenta lines show the average measures (± one standard error) of the LDA samples [Colour figure can be viewed at wileyonlinelibrary.com]

uncertain topics outweigh the credible topics. More interestingly, subsets of minimum clusters size 5, 10 or 20 show significantly better perplexity, depending on the cosine distance threshold, for instance, the subset of cluster topics with a minimum cluster size of 5 and at cosine distance larger than 0.15. The coherence plot (top right) and distinctiveness plot (bottom left) 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

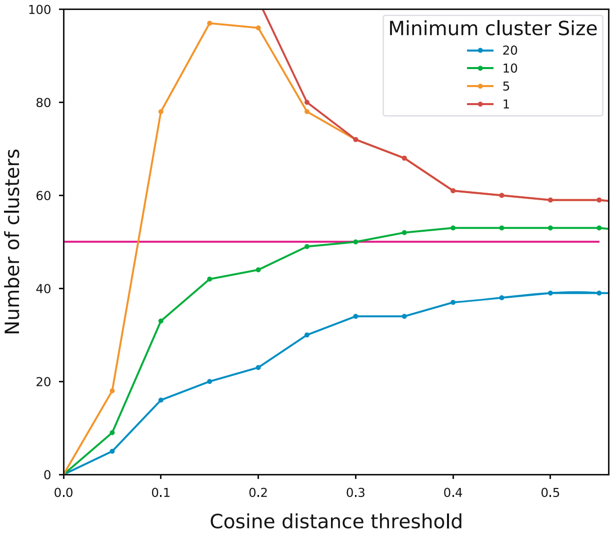


FIGURE 7 The number of clusters obtained at cosine distance (varying from 0 to 1 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). The magenta line shows the number of topics in the latent Dirichlet allocation samples. For visualisation purposes, subsets larger than 100 clusters are not shown (subsets with minimum size 1 and $CD < 0.25$) [Colour figure can be viewed at wileyonlinelibrary.com]

the new cluster grows in size). Interestingly, the credibility plot (bottom right) 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 7. These topics may not repeat 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 Appendix B.

Figure 7 shows the number of clustered topics obtained by varying cosine distance thresholds and minimum cluster size. Subsets with a minimum cluster size of 1 report a large number of clusters (more than 100), which for visualisation purposes are not shown. Topics that reappear in 20 samples are always fewer than 100 (number of topics of the LDA samples), confirming the uncertainty and low credibility of some topics. Note that the number of clusters does not get reduced up to 1 because the hierarchical clustering only merge clusters if their topics come from different samples.

Based on this analysis, we select a subset generated by minimum cluster size 10 and 0.35 CD threshold. Minimum cluster size 20 may lead to greater coherence but lower perplexity and vice-versa minimum cluster size 1 or 5 leads to better perplexity but worse coherence. After the 0.35 CD threshold, perplexity is no longer improved. 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 three experiments with LDA samples with 50, 100 and 200 topics, but this time, we allow merging of topics within the same posterior sample. This implies that the clustering is no longer just a summary of the posterior distribution, but it is also, in effect, informing the number of topics within LDA. This allows us to compare and interpret some of the behaviour of the clustered topics from models with a large number of topics, as gathering similar topics from the same and different samples will form more distinctive subsets of clustered topics.

TABLE 1 Generalisation, coherence, distinctiveness and stability metrics of latent Dirichlet allocation (LDA) samples and subsets of clustered topics (HC-LDA and HC-LDA-WS) obtained from clustering LDA samples with 50, 100 and 200 topics

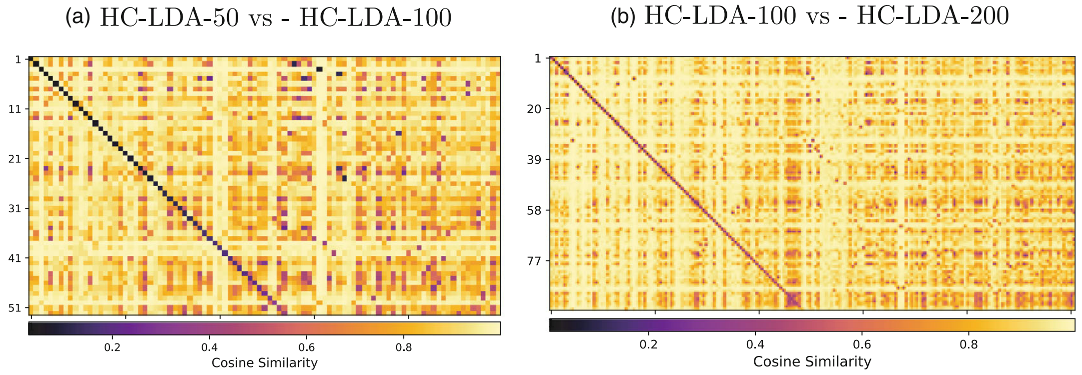
		Generalisation	**Coherence**	**Distinctiveness**	**Credibility**
		Perplexity	**NPMI**	$\mathbf{CD}_{min}$	$\overline{\mathbf{CS}_{\max}}$
Model	**Topics**	**Mean ± SE**	**Mean ± SE**	$\mathbf{Mean}_{min}$ **± SE**	**Mean ± SE**
LDA-50	50	8.130 ± 0.003	0.325 ± 0.006	0.672 ± 0.020	0.769 ± 0.011
HC-LDA-50	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

In Table 1, we compare the performance of the selected subsets when topics from different samples form a cluster (HC-LDA), and when topics from the same and different samples form a cluster (HC-LDA-WS), against the average performance of the LDA models. As observed, subsets of clustered topics show significantly lower measures of generalisation, larger topic coherence and larger topic credibility than LDA inferred topics. Note that topic distinctiveness is not improved, which might result from excluding highly distinctive non-recurrent topics. Allowing the merging of topics from the same samples retrieves fewer topics, does not significantly improve perplexity but increases the subset distinctiveness.

Different numbers of topics may retrieve similar performance. For example, Table 1 shows that the 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 9h is only found in LDA samples with 200 topics. Besides, clustered topics may be included in a subset derived from larger LDA samples. For instance, Figure 8 shows that the clustered topics 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-parameterised model.

8 | TOPICS IN BRITISH GROCERY RETAIL

The analysis of topics and the products that together fulfil customers' motivations convey customer insights, that is, diet orientations, cooking from scratch, preference for specific drinks or


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

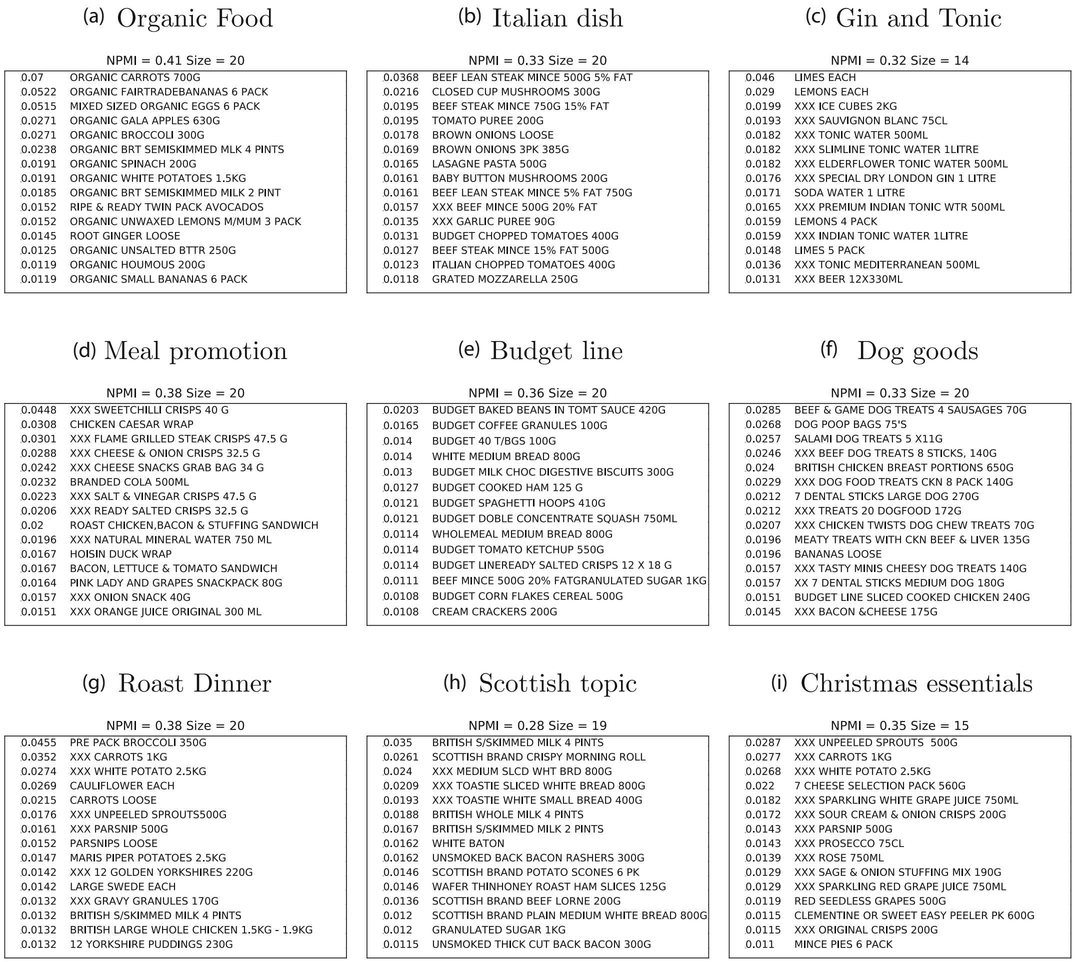
(b) HC-LDA-100 vs - HC-LDA-200

FIGURE 8 Clustered topics correspondence between clustering of latent Dirichlet allocation samples with 50, 100 and 200 topics [Colour figure can be viewed at wileyonlinelibrary.com]

dishes, etc. For instance, Figure 9a presents the topic of 'Organic Food', Figure 9b shows ingredients to cook an 'Italian dish' and Figure 9c highlights ingredients to prepare 'Gin and Tonic'. Along with these topics, other identified topics show vegetarian-friendly foods, free-from lactose/gluten foods, ingredients for cooking Asian, Mexican or Indian recipes. In these examples, topics gather products from different categories, that is, ice and tonic water are two categories while tonic water and soda water are in the same category. Identifying combinations of products from different categories may have useful and commercial implications in improving product recommendations, developing promotional campaigns, optimising assortments and planning shelf space, etc.

In contrast to cooking from scratch, customers may prefer convenience foods such as ready-to-eat meal promotions. For example, Figure 9d represents a 'meal promotion' composed of a sandwich, a bottle of soda or water and a package of prepared fruit or crisps. Topics also show that customers tend to choose products within the supermarket's budget line or premium line, for example, Figure 9e gathers products from a 'budget line' which offers products of a lower price than branded substitutes. Pet-ownership or household composition can be illustrated by topics, for instance, Figure 9f lists 'dog goods', including food, meat and cleaning items. Other topics illustrate baby-related foods and large size items indicating household composition. Topics reveal customer's decision drivers, which can aid further customer analysis such as customer segmentation and customer profiling, to improve customer experience and to build brand loyalty.

Topics reveal customer motivations that are driven by specific events, geography or seasonality. For instance, Figure 9g depicts the 'roast dinner' which is a traditional British main meal that is typically served on Sunday. Other event-specific topics manifest customers' motivations such, as having a picnic, buying a gift (flowers and chocolates), or barbecue. Topics also exhibit specific shopping themes that are driven by products that are available or highly preferred in certain locations or at specific periods. For example, Figure 9h reveals Scottish-branded products in the 'Scottish topic'. Similarly, a Northern Irish topic includes packed and locally supplied foods. Figure 9i shows the 'Christmas essentials' topic which is characterised by mince pies, sparkling grape juice, produce and snacks. Easter and Halloween are also depicted by topics that contain the icons: chocolate egg and pumpkin respectively. Commercially speaking, identifying events and geographical/seasonal patterns may inform marketing campaigns and support the optimisation of product assortment.



(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) Meal 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 topic

NPMI = 0.28 Size = 19

0.035	BRITISH S/SKIMMED MILK 4 PINTS
0.0261	SCOTTISH BRAND CRISPY MORNING ROLL
0.024	XXX MEDIUM SLCD WHT BRD 800G
0.0209	XXX TOASTIE SLICED WHITE BREAD 800G
0.0193	XXX TOASTIE WHITE SMALL BREAD 400G
0.0188	BRITISH WHOLE MILK 4 PINTS
0.0167	BRITISH S/SKIMMED MILK 2 PINTS
0.0162	WHITE BATON
0.0162	UNSMOKED BACK BACON RASHERS 300G
0.0146	SCOTTISH BRAND POTATO SCONES 6 PK
0.0146	WAFER THINHONEY ROAST HAM SLICES 125G
0.0136	SCOTTISH BRAND BEEF LORNE 200G
0.012	SCOTTISH BRAND PLAIN MEDIUM WHITE BREAD 800G
0.012	GRANULATED SUGAR 1KG
0.0115	UNSMOKED THICK CUT BACK BACON 300G

(i) Christmas essentials

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 9 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 anonymity purposes. NPMI > 0 is associated with coherent topics. Size is the number of posterior samples the topic has been found in. Topics reflect a variety of shopping motivations, that is, diet orientations, cooking from scratch, ready-to-eat meals, preference for budget/premium product lines, pet ownership/household composition, specific events, geography and seasonality. Topics may also be associated with consumption of alcohol/fat/salt/sugar

Our approach allows us to provide measures of uncertainty for each inferred topic. For example, the topics 'Organic food', '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 essentials' appeared 19 of 20 and 15 of 20 times, respectively, within the 20 LDA posterior draws. The lower frequency of these topics might be explained by the small representation of them 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 an irrelevant topic.

Understanding grocery consumption not only assists marketing practices but also opens up new avenues for social research. Uncovering consumption patterns related to alcohol (Figure 9c)/fat/sugar/salt through topic modelling is scalable, low-cost and allows the identification of specific products and their characteristics. Thus, topic modelling may help the conduction of dietary studies that are typically limited to survey data such as food frequency questionnaires and open-ended dietary assessment (Aiello et al., 2019; Einsele et al., 2015; Wang et al., 2014; Wardle, 2007).

9 | CONCLUSION

In this paper, we expand the evaluation process of LDA to include qualitative aspects such as topic coherence, topic distinctiveness and topic credibility along with model generalisation. In addition, we propose a methodology that post-processes LDA models, to summarise the entire posterior distribution of an LDA model into a single set of topical modes. Our approach identifies recurrent topics using meaningful distance criteria and allows the user to assess topic credibility. The distance criteria were developed through a customised survey which we carried out with experts in the field of grocery retailing; these helped us evaluate and set thresholds that assist the evaluation of interpretability and similarity of grocery retail topics. Empirically, we showed the advantages of the proposed methodology in terms of capturing topic uncertainty and enhancing coherence and credibility. We identified credible and coherent topics that exhibit a variety of shopping motivations, that is, diet orientations, cooking from scratch, specific events, pet ownership, geography, seasonality, etc. Topics can be associated with alcohol/fat/salt/sugar consumption, which may provide new venues for sociological research. Finally, our methods focused on the context of LDA models. Summarising multiple posterior draws from a mixture model, however, is a challenge that extends beyond LDA. Our methods can be implemented beyond LDA by replacing the cosine distance with other measures relevant to each context.

ACKNOWLEDGEMENTS

This work has been carried out with the financial support of the Economic and Social Research Council, the Alan Turing Institute and dunnhumby ltd, our industrial partner. Access to the anonymized data servers was granted to us by dunnhumby ltd.

DATA AVAILABILITY STATEMENT

Data used in this research is commercially sensitive and protected by dunnhumby; thereby, our dataset containing grocery transactions is not available to the public. Datasets can be requested directly from the company.

ORCID

Mariflor Vega Carrasco https://orcid.org/0000-0003-1839-2287
Ioanna Manolopoulou https://orcid.org/0000-0002-5379-2916

REFERENCES

Aiello, L.M., Schifanella, R., Quercia, D. & Del Prete, L. (2019) Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Science*, 8, 14.

Aletras, N. & Stevenson, M. (2013) Evaluating topic coherence using distributional semantics. In: *IWCS'13*, 13–22.

Aletras, N. & Stevenson, M. (2014) Measuring the similarity between automatically generated topics. In: *ACL'14*, vol. 2, 22–27.

AlSumait, L., Barbará, D., Gentle, J. & Domeniconi, C. (2009) Topic significance ranking of lda generative models. In: *ECML PKDD'09*, pp. 67–82. Springer.

Blair, S.J., Bi, Y. & Mulvenna, M.D. (2016) Increasing topic coherence by aggregating topic models. In: *KDD'16*, 69–81.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Bouma, G. (2009) Normalized (pointwise) mutual information in collocation extraction. GSCL'09, 31–40.

Boyd-Graber, J., Mimno, D. & Newman, D. (2014) Care and feeding of topic models: problems, diagnostics, and improvements. In: *Handbook of mixed membership models and their applications*, CRC Handbooks of Modern Statistical Methods. Boca Raton, FL: CRC Press.

Buntine, W. (2009) Estimating likelihoods for topic models. In: *ACML'09*, 51–64. Springer.

Celeux, G. (1998) Bayesian inference for mixture: The label switching problem. In: *COMPSTAT'98*, pp. 227–232. Springer.

Celeux, G., Hurn, M. & Robert, C.P. (2000) Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95, 957–970.

Chaney, A.J.-B. & Blei, D.M. (2012) Visualizing topic models. In: *ICWSM'12*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L. & Blei, D.M. (2009) Reading tea leaves: How humans interpret topic models. In *NIPS'09*, 288–296.

Chen, F., Liu, X., Proserpio, D., Troncoso, I. & Xiong, F. (2020) Studying product competition using representation learning. In: *SIGIR'20*, 1261–1268.

Christidis, K., Apostolou, D. & Mentzas, G. (2010) Exploring customer preferences with probabilistic topics models. In: *ECML-PKDD'10*, pp. 12–24.

Chuang, J., Manning, C.D. & Heer, J. (2012) Termite: Visualization techniques for assessing textual topic models. In: *AVI'12*, 74–77. ACM.

Chuang, J., Roberts, M.E., Stewart, B.M., Weiss, R., Tingley, D., Grimmer, J. et al. (2015) Topiccheck: Interactive alignment for assessing topic model stability. In: NAACL HLT'15, 175–184.

Einsele, F., Sadeghi, L., Ingold, R. & Jenzer, H. (2015) A study about discovery of critical food consumption patterns linked with lifestyle diseases using data mining methods. In: *BIOSTEC'15*, vol. 5, pp. 239–245. Setubal, PRT.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013) *Bayesian data analysis*. Boca Raton: CRC press.

Griffiths, T.L. & Steyvers, M. (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101, 5228–5235.

Hastie, D.I., Liverani, S. & Richardson, S. (2015) Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, 25, 1023–1037.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P. & Giles, L. (2009) Detecting topic evolution in scientific literature: how can citations help? In: *CIKM '09*, pp. 957–966.

Hoffman, M., Bach, F. & Blei, D. (2010) Online learning for latent Dirichlet allocation. *NIPS '10*, 23, 856–864.

Hornsby, A.N., Evans, T., Riefer, P.S., Prior, R. & Love, B.C. (2020) Conceptual organization is revealed by consumer activity patterns. *Computational Brain & Behavior*, 3, 162–173.

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

Hruschka, H. (2016) Hidden variable models for market basket data. statistical performance and managerial implications. *University of Regensburg Working Papers in Business, Economics and Management Information Systems 489*, University of Regensburg, Department of Economics.

Hruschka, H. (2021) Comparing unsupervised probabilistic machine learning methods for market basket analysis. *Review of Managerial Science*, 15, 497–527.

Hurn, M., Justel, A. & Robert, C.P. (2003) Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12, 55–79.

Jacobs, B., Fok, D. & Donkers, B. (2020) Understanding large-scale dynamic purchase behavior. *ERIM Report Series Research in Management Erasmus Research Institute of Management*.

Jacobs, B.J., Donkers, B. & Fok, D. (2016) Model-based purchase predictions for large assortments. *Marketing Science*, 35, 389–404.

Jasra, A., Holmes, C.C. & Stephens, D.A. (2005) Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20, 50–67.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. et al. (2019) Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78, 15169–15211.

Lau, J.H., Newman, D. & Baldwin, T. (2014) Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: *EACL'14*, 530–539.

Li, W. & McCallum, A. (2006) Pachinko allocation: Dag-structured mixture models of topic correlations. In: *ICML'06*, 577–584. ACM.

McLachlan, G.J., Lee, S.X. & Rathnayake, S.I. (2019) Finite mixture models. *Annual Review of Statistics and Its Application*, 6, 355–378.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M. & McCallum, A. (2011) Optimizing semantic coherence in topic models. In: *EMNLP'11*, 262–272. Association for Computational Linguistics.

Minka, T. & Lafferty, J. (2002) Expectation-propagation for the generative aspect model. In *UAI' 02*, UAI'02, 352–359. Morgan Kaufmann Publishers Inc.

Newman, D., Asuncion, A., Smyth, P. & Welling, M. (2009) Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10, 1801–1828.

Newman, D., Lau, J.H., Grieser, K. & Baldwin, T. (2010) Automatic evaluation of topic coherence. In *NAACL HLT'10*, 100–108. Association for Computational Linguistics.

Newman, D., Bonilla, E.V. & Buntine, W. (2011) Improving topic coherence with regularized topic models. In *NIPS'11*, 496–504.

Ramage, D., Hall, D., Nallapati, R. & Manning, C.D. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: *EMNLP'09*, 248–256. Association for Computational Linguistics.

Ramon, Y., Martens, D., Provost, F. & Evgeniou, T. (2020) A comparison of instancelevel counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. *Advances in Data Analysis and Classification*, 1–19.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P. & Steyvers, M. (2010) Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, 28, 1–38.

Ruiz, F.J., Athey, S., & Blei, D.M. (2020) Shopper: A probabilistic model of consumer choice with substitutes and complements. *Annals of Applied Statistics*, 14, 1–27.

Schröder, N. (2017) Using multidimensional item response theory models to explain multi-category purchases. *Marketing: ZFP–Journal of Research and Management*, 39, 27–37.

Sievert, C. & Shirley, K. (2014) LDAvis: A method for visualizing and interpreting topics. In: *ACL'14*, 63–70.

Sperrin, M., Jaki, T. & Wit, E. (2010) Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20, 357–366.

Srivastava, A. & Sutton, C. (2017) Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

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

Stephens, M. & Phil, D. (1997) Bayesian methods for mixtures of normal distributions.

Steyvers, M. & Griffiths, T. (2007) Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427, 424–440.

Taddy, M. (2012) On estimation and selection for topic models. In: *AISTATS'12*, 1184–1193.

Wallach, H.M. (2008) Structured topic models for language. Ph.D. thesis, Cambridge: University of Cambridge.

Wallach, H.M., Mimno, D.M. & McCallum, A. (2009a) Rethinking LDA: Why priors matter. In *NIPS'09*, 1973–1981.

Wallach, H.M., Murray, I., Salakhutdinov, R. & Mimno, D. (2009b) Evaluation methods for topic models. In: *ICML'09*, 1105–1112. ACM.

Wang, X., Zhang, K., Jin, X. & Shen, D. (2009) Mining common topics from multiple asynchronous text streams. In: *WSDM'09, 192–201*. ACM.

Wang, X., Ouyang, Y., Liu, J., Zhu, M., Zhao, G., Bao, W. et al. (2014) Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. *The BMJ*, 349, 4490.

Wardle, J. (2007) Eating behaviour and obesity. *Obesity Reviews*, 8, 73–75.

Xing, L. & Paul, M.J. (2018) Diagnosing and improving topic models by analysing posterior variability. In: *AAAI'18*.

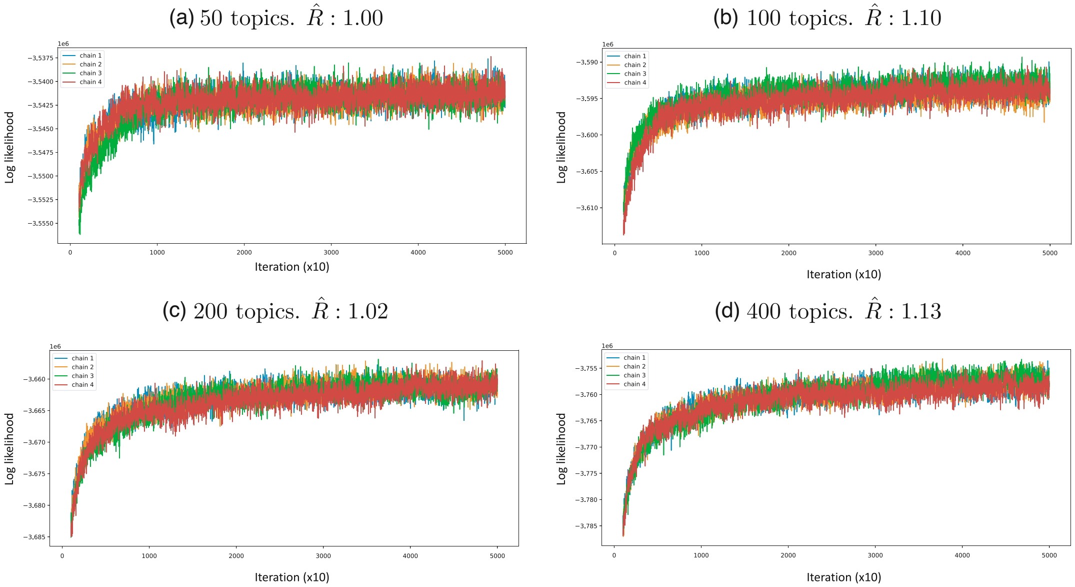
How to cite this article: Vega Carrasco, M., Manolopoulou, I., O'Sullivan, J., Prior, R. & Musolesi, M. (2022) Posterior summaries of grocery retail topic models: Evaluation, interpretability and credibility. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(3), 562–588. Available from: https://doi.org/10.1111/rssc.12546

APPENDIX A. MCMC CONVERGENCE

For each LDA model, four Markov chains are run for 50,000 iterations with a burn-in period of 30,000 iterations. We evaluate convergence using the potential scale reduction factor $\widehat{R}$ (Gelman et al., 2013). When $\widehat{R}$ is near 1, we can assume that samples approximate the posterior distribution. Values of $\widehat{R}$ below 1.1 are acceptable. Figure A1 shows the trace plot for the log-likelihood (measured at every 10 iterations) of LDA with 50, 100, 200 and 400 topics. We calculate the potential scale reduction factor using four chains and 8000 samples. Chains for LDA with 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.

APPENDIX B. CLUSTERING OF TOPICS

Figures A2 and A3 show the evaluation of subsets of clustered topics obtained from 20 LDA posterior samples with 100 and 200 topics.

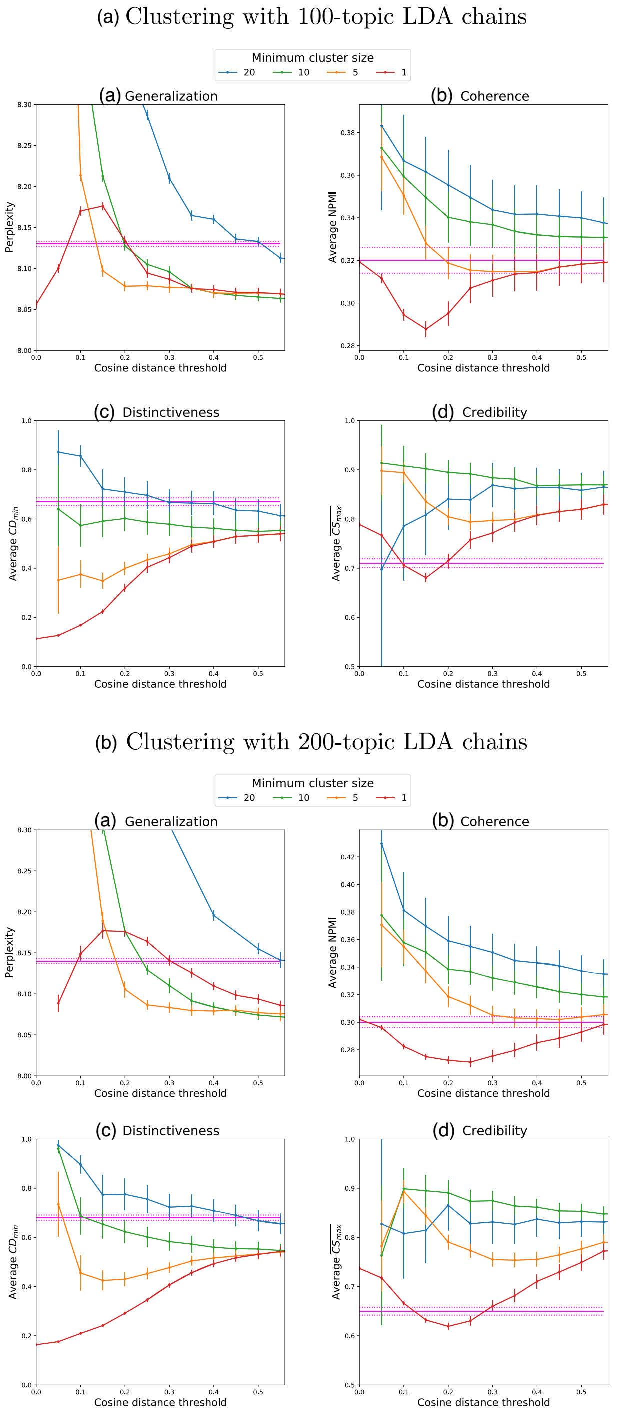

(a) 50 topics. $\hat{R}: 1.00$

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

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

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

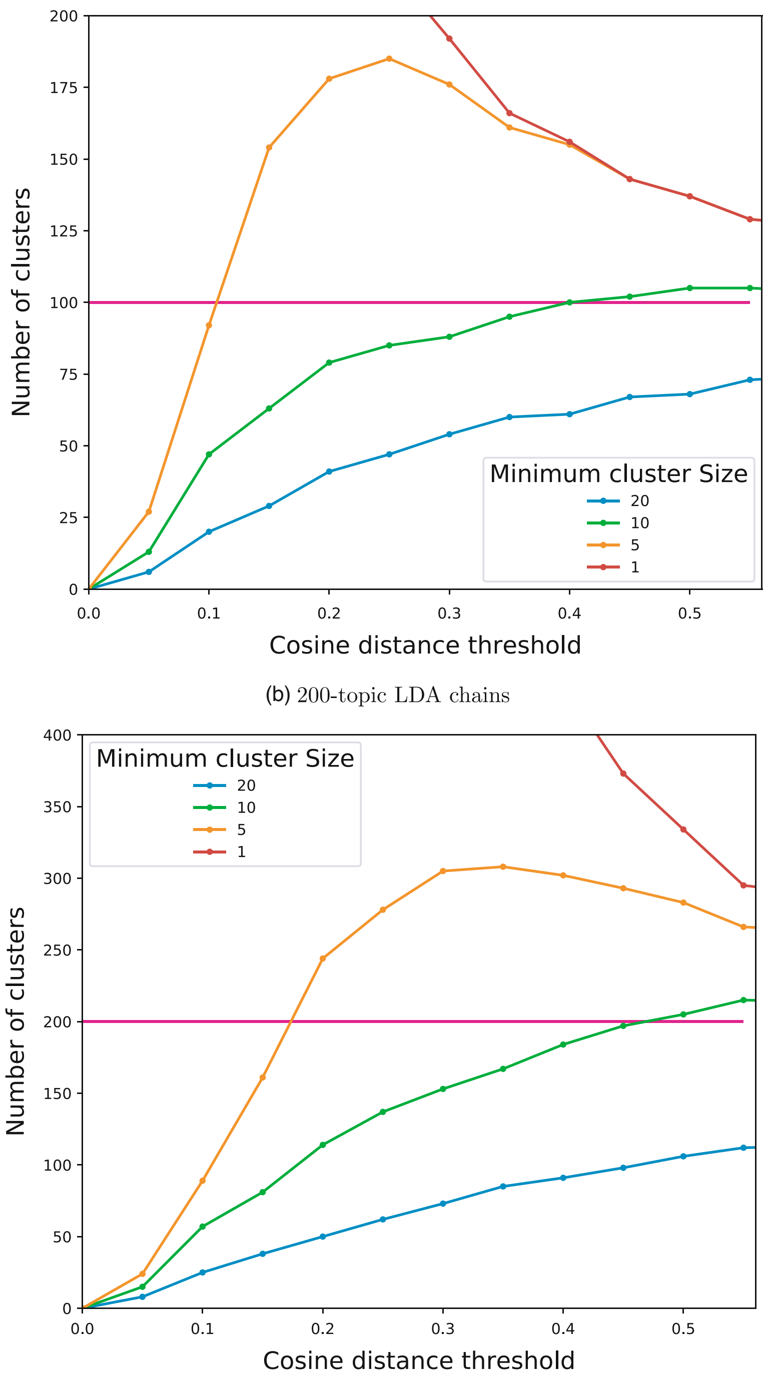
FIGURE A1 Markov Chains of latent Dirichlet allocation with 50, 100, 200 and 400 topics. $\hat{R}$ is the potential scale reduction factor [Colour figure can be viewed at wileyonlinelibrary.com]



(a) Clustering with 100-topic LDA chains

(b) Clustering with 200-topic LDA chains

FIGURE A2 Subset evaluation using cosine distance (varying from 0 to 1 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). Clustered topics were obtained from clustering 20 samples of latent Dirichlet allocation with 100/200 topics. Vertical lines represent one standard error. Magenta lines show the average measures (± one standard error) of the LDA samples [Colour figure can be viewed at wileyonlinelibrary.com]



(a) 100-topic LDA chains

(b) 200-topic LDA chains

FIGURE A3 The number of clusters obtained at cosine distance (varying from 0 to 1 with increments of 0.05) and minimum cluster size (20, 10, 5 and 1). The magenta line shows the number of topics in the latent Dirichlet allocation samples. For visualisation purposes, subsets larger than 200/400 clusters are not shown [Colour figure can be viewed at wileyonlinelibrary.com]