Improving Text Analysis Using Sentence Conjunctions and Punctuation

by

Joachim Büschken
School of Management
Catholic University of Eichstätt-Ingolstadt
joachim.bueschken@ku.de

Greg M. Allenby
Fisher College of Business
Ohio State University
allenby.1@osu.edu

January 31, 2017

Improving Text Analysis Using Sentence Conjunctions and Punctuation

Abstract

User generated content in the form of customer reviews, blogs or tweets is an emerging and rich source of data for marketers. Topic models have been successfully applied to such data, demonstrating that empirical text analysis benefits greatly from a latent variable approach which summarizes high-level interactions among words. We propose a new topic model that allows for serial dependency of topics in text. That is, topics may carry over from word to word in a document, violating the bag-of-words assumption in traditional topic models. In our model, topic carry-over is informed by sentence conjunctions and punctuation. Typically, such observed information is eliminated prior to analyzing text data (i.e., "pre-processing") because words such as "and" and "but" do not differentiate topics. We find that these elements of grammar contain information relevant to topic changes. We examine the performance of our model using multiple data sets and establish boundary conditions for when our model leads to improved inference about customer evaluations. Implications and opportunities for future research are discussed.

Keywords: LDA, Autocorrelated Topics, Syntactic Covariates, User-generated Content, Bayesian Analysis

1 Introduction

Text data in the form of customer reviews, blogs and tweets is a fast-growing and rich source of data for marketing researchers. Websites such tripadvisor.com or yelp.com offer a growing range of products and services for which customers can post reviews. An important and fruitful area of model-based empirical text analysis is the application of latent topic models (Blei et al. 2003; Tirunillai and Tellis 2014) to such data. Topic models identify sets of words that frequently co-occur, giving rise to the ability to account for high-level interaction among words. Essentially, topic models are devices to detect latent clusters of co-occurring multinomial variables. The clusters emerging from these models can be used to analyze the relationship between topics and variables of interest such as purchase intention and customer satisfaction (Büschken and Allenby, 2016), providing insights into consumer preferences and behavior.

A challenge in applying topic models to customer reviews is the limited amount of data contained in any one review. The number of words in a review is typically less than 100, making it difficult to assess topic and word probabilities without imposing additional structure. An assumption typically present in topic models is that topics exhibit zero autocorrelation in that the probability of the topic assignment to word $t+1$ is independent of the topic assignment to word t. This assumption gives rise to the "bag-of-words" property and to word counts being sufficient statistics for the standard model. Recently, Büschken and Allenby (2016) propose a model in which topics are constrained to not change within a sentence. They show that this restriction leads to better fit to the data and more interpretable topic word probabilities in customer review data. A similar approach was proposed by Nallapati and Allan (2002) who use sentence boundaries as structural information to a unigram-type probabilistic language model. The common element to both these models is to impose a sentence-based constraint to the model, imposing common topics within observed boundaries of text.

In this paper we propose an autocorrelated topic model that relaxes the assumption that topics remain fixed throughout a sentence. Our model assumes that a reviewer stays with a topic for some time before switching to another, and this switch may occur midway through the sentence or paragraph. The probability of a topic change is parameterized as a binary logit model with covariates, and we find that punctuation (e.g., periods, exclamation marks, commas) and conjunctions (e.g., and, but, because) are predictive of topic carryover. These grammatical elements are frequently discarded as part of data cleaning in text analysis, and have not been previously analyzed for their value in predicting topic changes within sentences and inferences about the topics themselves.

We apply our model to five datasets and find improvements relative to an unconstrained, standard-type latent Dirichlet allocation (LDA) model in all cases, and improvement relative to a sentence constrained model (SC-LDA) for longer reviews. That is, the sentence-constrained model predicts customer evaluations better than the proposed autocorrelated model when reviews are short in length and worse when reviews are longer. Sentence punctuation and conjunctions are shown to signal the start of new topics in addition to full stops previously found to be useful in the SC-LDA model.

The remainder of this paper is organized as follows. In section 2, we develop our autocorrelated topic model with covariates that allow topic transitions to be dependent on conjunctions and punctuation in sentences. In Section 3, we present a summary of the data we use in our empirical analysis. In section 4, we present results from applying our model to the data and compare it to various benchmark models. Section 5 presents a summary of our results and concluding comments.

2 Autocorrelated Topic Model

2.1 Previous research

The traditional latent Dirichlet allocation (LDA) model assumes that each word observed in a text document is generated from latent topics characterized by topic-specific word probabilities across a fixed vocabulary. Each document (d) is described by a vector of topic probabilities θ_d of dimension T and each topic is characterized by vector ϕ_t that specifies the word probabilities associated with that topic. Words in a document are generated by first drawing a latent indicator variable z from a discrete (multinomial) distribution with probability vector θ_d and then drawing a word from the vocabulary list with probability vector ϕ_z. Büschken and Allenby (2016) propose a constrained version of this model by restricting all words within a sentence to be generated from the same topic. This is accomplished by drawing the latent variable z once for all words in a sentence.

LDA and sentence-constrained LDA represent two extremes in topic generation. Topics generated from the LDA are assumed independent and identically distributed (IID) across all words, while the sentence-constrained LDA is a model of deterministic dependency within an observed locale and with only a sentence's period (e.g. full stop, exclamation point) allowing for topic variation within a document. It would be straightforward to extend this constraint to any locale of interest (clauses of sentences, paragraph, chapter etc.) as long as its boundaries are observed. In this paper, we use a different approach in that we allow topics to carry over probabilistically from word-to-word instead of sentence-to-sentence or paragraph-to-paragraph, introducing a more flexible model of serial dependence for the topics. This approach seems a more realistic representation of speech. Typically, when discussing a particular service experience, a reviewer stays with one topic for some time before switching to another. Such switches may occur within a sentence such as in the statement in a hotel review: "Although the staff were friendly and helpful, the rooms and facilities were really not clean." The use of a comma in this

sentence indicates a possible change in topic.

The issue of correlated topics has been examined previously in the literature. Griffiths et al. (2004) propose a model in which words differ in syntactic (i.e., related to placement or arrangement) and semantic (i.e., related to meaning) content. A hidden Markov model (HMM) is used to generate sentence syntax and an LDA (topic) model is used to generate its content. The HMM model introduces serial correlation between syntactic vocabularies and topics, but not among the topics themselves. Wallach (2006) proposes a model where topics generate words conditional on the previous word, introducing first-order autocorrelation in word generation, but not in topic generation. That is, the topic indicator variable is still assumed to be IID across words. Blei and Lafferty (2007) employ a logistic Normal distribution instead of a Dirichlet distribution to allow for correlations in the prior for the topics. Their model affects the topic probabilities θ_d, but does not induce autocorrelation of topics within the document indicated by the latent indicator variables z. Trusov et al. (2016) propose a model with correlated topics for website visitation data to account for latent interests (or, as they are called by the authors, "roles") simultaneously driving the number of times different websites are visited up (or down). The common element to these approaches is that topics may exhibit a priori dependence (e.g. if reviewers talk extensively about service problems in restaurant reviews, they might also talk more about inflated prices). Finally, in a version of their sentence-constrained LDA model, Büschken and Allenby (2016) allow for a probabilistic carry-over of topics from sentence to sentence, but do not find empirical support for this model.

The model applied in this research allows for autocorrelation in the draws of z from word to word. Essentially, this model is a word-based version of the sentence-based model investigated in Büschken and Allenby (2016). A word-based approach allows for topic carry-over from word to word. Thus, the difference to models with a priori correlated topics (Blei and Lafferty 2007; Trusov et al. 2016) is that we consider dependency of topics on the word level, not the level of the prior. We use observed structural information

in text to affect a possible topic change. An important structural feature of written romanic and germanic languages is the use of conjunctions such as "and" and "but" that play a syntactic role by joining parts of the sentence. As such, they do not represent topics or semantic content. A typical example for this role is present in the hotel review: "Comfy beds but not your usual large American beds" in which the conjunction "but" links two different perspectives of evaluation, one personal, the other more general ("not American"). In comparison, punctuation typically does not join, but separates part of speech. A full stop, for example, indicates the end of a sentence and introduces a pause to the flow of thoughts. Such a pause is a natural candidate for a topic change. Other examples of punctuation are exclamation or question marks, both of which introduce structure to text in a similar way, but also add weight or a interrogative notion to a statement. Structural punctuation for the purpose of this analysis are marks that act on parts of documents typically not larger than a sentence and not smaller than a word (e.g. hyphens) (Say and Akman 1996, Meyer 1987). The central idea of our model is to use the observed structural information in text presented by punctuation and conjunctions for inference regarding the dynamics of topics in text.

It is interesting to note that, in empirical applications of topics models, conjunctions as well as incidents of punctuation are typically removed from the data prior to analysis (i.e., pre-processing). Conjunctions are removed because they are stopwords. Stopwords typically carry very little power to discriminate topics. This is evident in the nearly uniform probabilities of stopwords, if included in the data, to appear under any topic. However, we propose that conjunctions and punctuation, as carriers of structural information, present information to topic change and introduce this information to our model. The challenge is in retaining the structural information without compromising inference about topics. In Figure (1), we present a stylized way of using these data in our model.

In the review at the top of Figure (1), we highlight conjunctions (green), punctuation (red) and stopwords other than conjunctions (blue). In two versions of this review, num-

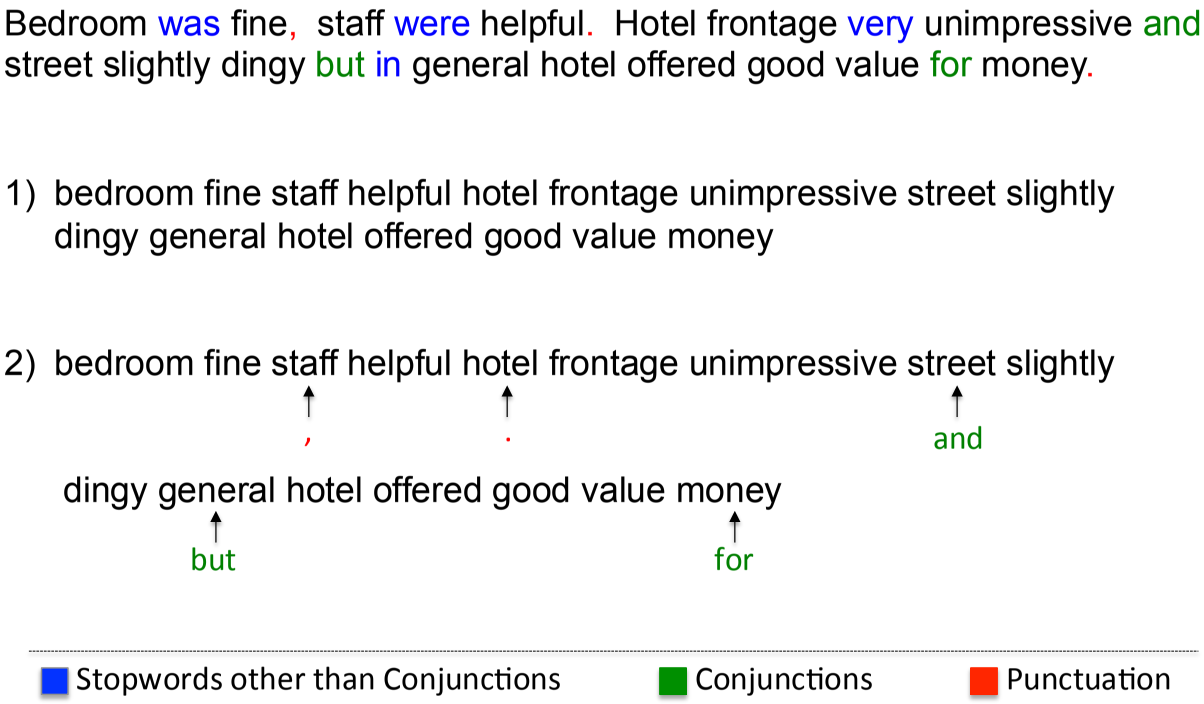


Figure 1: Using syntactic covariates in topic analysis for actual example of a hotel review. Top: Original review. Middle: Review after removing stopwords, including all conjunctions and all punctuation. Bottom: Review after removal of stopwords other than conjunctions and positional assignment of conjunctions and punctuation as covariates. Arrows indicate role of covariates as prior information to topic assignments.

bered 1) and 2), we present different ways of exploiting structural information. Version 1) results from removal of all stopwords, including conjunctions, and all punctuation. This pre-processing of data is consistent with the "bag-of-words" assumption and uses no structural information other than the remaining words. For version 2), which is applied here, we view conjunctions (green) and punctuation (red) as prior information to a topic carry-over between consecutive words. For example, the conjunction "but" is used as a covariate to the probability of a topic carry over from the word (street slighty) "dingy" to the word "general" (hotel offered good value). In a similar fashion, the full stop preceding the word "hotel" is covariate to the probability of a carry over from (staff) "helpful" to "hotel" (frontage). The difference between the two approaches lies in the use of otherwise ignored data as observed covariates to topic change. In the empirical application of our model we find topic carry-over to be heavily driven by structural elements of text.

2.2 Model development

We propose a topic model in which the topic assignments of words in a document may exhibit serial dependency by way of carry-over from word to word and in which structural covariates are prior information to topic carry over. Our model is based on the LDA topic model (Blei et al., 2003) which proposes the following joint distribution of knowns and unknowns:

$$p\left(w_n, z_n, \theta_d, \phi, \alpha, \beta\right) = p\left(w_n|\phi_{z_n}\right) \times p\left(z_n|\theta_d\right) \times p\left(\theta_d|\alpha\right) \times p\left(\phi|\beta\right) \times p\left(\alpha\right) \times p\left(\beta\right) \quad (1)$$

where:

w_n is the n-th word of document d,

z_n is the topic assignment of word w_n,

θ_d is a vector of prior topic probabilities for document d,

ϕ_t is a vector of word probabilities given topic t, and

α, β are fixed priors of θ_d and ϕ_t, respectively.

We extend this model so that the topic z_{n-1} assigned to word w_{n-1} may carry over to word w_n independent of θ_d, so that topics are autocorrelated. We define carry-over of a topic as: $z_n = z_{n-1}$, and introduce the latent binary variable ζ_n to indicate whether the topic assignment to word w_n is the result of carry-over:

$$\begin{aligned} &\zeta_n = 1 : z_n = z_{n-1} \\ &\zeta_n = 0 : z_n \sim \text{Multinomial}(\theta_d) \end{aligned} \tag{2}$$

In the LDA model, $\zeta_n = 0 \ \forall n$, implying that this model is a special case of the AT-LDA. The same holds for the sentence constrained LDA which imposes $\zeta_{n,s} = 1 \ \forall n_s > 1$ where s is a sentence. We assume ζ_n to be distributed Binomial with probability ψ_n:

$$\begin{aligned} &\zeta_n \sim \text{Binomial}(\psi_n | z_{n-1}) \\ &\psi_n | z_{n-1} = \frac{\exp[\delta_{0,z_{n-1}} + \tilde{x}_n' \delta]}{1 + \exp[\delta_{0,z_{n-1}} + \tilde{x}_n' \delta]} \end{aligned} \tag{3}$$

where $\tilde{x}_n$ is a vector of dummy variables that indicate conjunctions and punctuation to the current word (Figure 1), and δ are estimated coefficients that affect the probability of topic change. Negative values of δ increase the likelihood of an IID topic draw, while positive values indicate that a topic carry-over is more likely. We allow for intercepts $\delta_{0,z_{n-1}}$ that depend on the previous word's topic z_{n-1}. Eq. (3) specifies common coefficients δ for the conjunctions and punctuation in $\tilde{x}_{n-1}$. In our empirical analysis, we also consider interaction effects of the latent topics and covariates giving rise to topic-specific effects of conjunctions and punctuations. The generative model of the AT-LDA with covariates to topic change and fixed priors α, β, $\mu_\delta, \Sigma_\delta$ is as follows:

1. Draw δ from MV Normal$(\mu_\delta, \Sigma_\delta)$
2. Draw ϕ_t from Dirichlet(β) $\forall t$ iid
3. Draw θ_d from Dirichlet(α) $\forall d$ iid
4. For the first word in document d, w_1:
   (a) Draw z_1 from Multinomial(θ_d)
   (b) Draw w_1 from Multinomial$(\phi_{t=z_1})$
   (c) compute $p(\psi_2|x_2, \delta, z_1)$ using Eq. (3), draw ζ_2 given ψ_2
5. For words w_n $n \in 2 : N_d$:
   (a) if $\zeta_n = 0$: draw z_n from Multinomial(θ_d); if $\zeta_n = 1$: set $z_n = z_{n-1}$
   (b) Draw w_n from Multinomial$(\phi_{t=z_n})$
   (c) compute $p(\psi_{n+1}|x_{n+1}, \delta, z_n)$, draw ζ_{n+1} given ψ_{n+1}
6. Repeat steps 4,5 for all documents $d \in D$ (except for draw of ζ_{N_d}).

The joint distribution of the knowns and unknowns of the AT-LDA model with covariates, given document d, factorizes as follows:

$$\begin{aligned} &p\left(\{w\}_d, \{z\}_d, \theta_d, \phi, \{\zeta\}_d, \delta, \alpha, \beta, x\right) \propto \\ &\qquad p\left(w_1|\phi, z_1\right) \times p\left(z_1|\theta_d\right) \times \\ &\prod_{n=2}^{N_d} \Big[p\left(w_n|\phi, z_n, z_{n-1}, \zeta_n\right) \times p\left(z_n|z_{n-1}, \theta_d, \zeta_n\right) \times p\left(\zeta_n|, x_n, \delta, z_{n-1}\right)\Big] \times \\ &\qquad p\left(\phi|\beta\right) \times p\left(\theta_d|\alpha\right) \times p\left(\beta\right) \times p\left(\alpha\right) \times p\left(\delta\right) \end{aligned} \tag{4}$$

where we, as usual, assume independent prior distributions. The likelihood of a word,

conditional on ζ_n:

$$p\left(w_n|\phi, z_n, z_{n-1}, \zeta_n = 0\right) = p\left(w_n|\phi, z_n\right)$$
$$p\left(w_n|\phi, z_n, z_{n-1}, \zeta_n = 1\right) = p\left(w_n|\phi, z_{n-1}\right)$$

The likelihood of a topic assignment, conditional on ζ_n:

$$p\left(z_n|z_{n-1}, \theta_d, \zeta_n = 0\right) = p\left(z_n|\theta_d\right)$$
$$p\left(z_n|z_{n-1}, \theta_d, \zeta_n = 1\right) = p\left(z_n = z_{n-1}\right) = 1$$

3 Data

We examine several data sets that differ with respect to size and complexity. Our purpose is twofold; first, we want to establish whether topic autocorrelation is a regular feature of text data typical to marketing-type applications; and second, we want to establish boundary conditions with respect to the need for a model that allows for autocorrelated topics.

Table (1) presents descriptive statistics of the data sets used in our study. Two of the five data sets contain restaurant reviews, one data set contains luxury hotel reviews and two data sets contain reviews of durable consumer products (camping tents, power drills). The luxury hotel data set consists of 3,214 reviews of 5-star hotels in Manhattan, NY, obtained from www.expedia.com. A second data sets contains 696 reviews of Italian restaurants obtained from we8there.com. From the same source, we also obtained 1,324 reviews of American restaurants. The fourth data set contains 2,100 reviews of camping tents obtained from www.amazon.com. We picked tents in the price range of $\$80 - 150$ which is the most popular range in terms of number of reviews posted. Lastly, the power drill data set consists of 4,438 reviews of power drills in the \$100-150 price range posted

Table 1: Descriptive statistics of data sets.

	Manhattan Hotels	Italian Restaurants	American Restaurants	Camping Tents	Power Drills
Number of reviews	3,214	696	1,324	2,100	4,438
Corpus size (words)	78,136	47,948	100,179	90,980	68,917
Number of unique terms	1,615	2,416	1,461	1,340	1,232
Number of words per review					
Mean	24.3	68.9	75.7	43.3	15.45
SD	19.1	72.1	90.3	64.2	24.35
Max	215	498	599	1331	252
Consumer rating					
Mean	4.42	3.82	3.74	3.91	4.64
SD	0.88	1.40	1.38	1.34	0.84

on www.amazon.com.

All reviews contain an overall evaluation of the customer experience on a 5-point rating scale. In our analyses of these data, we relate the topic probabilities to the overall rating as in Büschken and Allenby (2016). That is, we assume the observed rating for a review r_d is related to the latent topic probabilities using a cut-point model:

$$r_d = k \qquad \text{if} \qquad c_{k-1} \leq \tau_d \leq c_k \tag{5}$$

and

$$\tau_d \sim N(\theta_d'\beta, \sigma^2) \tag{6}$$

The fit of the cutpoint model provides a way of assessing the predictive plausibility of the competing models.

Table 1 shows that the our data sets differ greatly in terms of corpus size and complexity. The average number of words per review ranges from 24 (hotels) to 76 (American restaurants). The number of unique terms per data set ranges from 1,232 (power drills) to 2,416 (Italian restaurants). This implies Italian restaurant reviewers apply the most extensive vocabulary to describe their service experience in our analysis. Power drill reviewers, in comparison, write shorter reviews and typically use a smaller set of terms for this purpose. All data sets exhibit significant heterogeneity in terms of review length as indicated by the standard deviation and range of the number of words in each review. The coefficient of variation of the number of words exceeds one in four of the five datasets, and the distribution of the number of words in reviews in the American restaurants and camping tent datasets exhibit long tails, suggesting that many customers feel the need to report about their experience in great detail.

Table 2 reports counts of the use of conjunctions and various forms of punctuation in our data sets. From Table 2, it is clear that all data sets in our analysis are rich in syntactic content. The conjunction "and" appears on average 5.3 times (3,341/627) in the Italian

restaurant review and two times in each hotel review. Similarly, on average, full stops mark boundaries between sentences 10.7 times in Italian restaurant reviews and 3.3 times in hotel reviews. A special case of punctuation is presented by the use of (round) parentheses. We record the use of parentheses more than 500 times in the corpus of hotel reviews and about 580 times in the American restaurant reviews. Apparently, reviewers feel the need to structure some part of their narrative by placing words within parentheses. Typically, parentheses are used to clarify preceding text or, when combined with a full stop, as a side remark. Parentheses provide an observable signal of words belonging together suggesting some form on topical dependency. On average, a review in the American restaurant data set contains 39 structural elements in the form of conjunctions or punctuation. A review of luxury hotels contains 10. The frequency at which structural elements appear in our data raises the question why this information is typically ignored.

4 Empirical Analysis

We examine the performance of the proposed model by examining in-sample model fit and the prediction performance for customer ratings, topic carryover, and the direction of effects of the various conjunctions and punctuation.

4.1 Model fit and prediction

The predictive performance of the proposed autocorrelated topic model is examined and compared to other models:

1. Latent Dirichlet allocation (LDA)
2. Sentence-constrained LDA (SC-LDA)
3. Autocorrelated topic LDA without covariates (AT-LDA intercept only)
4. Autocorrelated topic LDA with common covariates

Table 2: Counts of conjunctions and punctuation in data sets.

	Manhattan Hotels	Italian Restaurants	American Restaurants	Camping Tents	Power Drills
Conjunctions					
for	1,859	1,098	2,387	2,932	2,621
and	5,696	3,341	7,277	5,693	4,431
but	1,020	791	1,723	1,479	985
or	279	261	489	577	366
yet	32	19	50	101	93
so	528	446	975	856	733
after	151	182	390	272	204
although	78	24	49	55	27
as	526	429	931	899	952
because	143	124	281	269	184
before	102	114	236	148	119
even	200	141	333	284	177
if	365	270	556	760	356
now	28	55	140	136	219
once	42	51	129	152	63
provided	31	9	14	25	14
since	56	96	183	131	105
than	215	148	339	362	439
that	897	916	2,110	1,477	1,196
though	80	52	112	136	65
unless	13	7	13	56	13
until	65	36	72	40	52
when	323	313	699	523	394
whenever	12	3	8	1	1
where	107	78	183	142	52
whether	4	5	20	6	13
which	294	223	624	300	230
while	78	90	213	211	108
who	79	95	203	99	75
why	27	28	65	44	47
what	137	226	488	209	333
Punctuation					
,	5,494	3,641	8,078	5,853	4,363
.	9,628	6,696	14,639	10,125	7,658
;	151	72	253	126	75
:	81	48	84	308	123
!	-	360	889	435	583
?	37	55	117	72	76
&	1	70	165	53	126
(	517	300	580	693	397
)	503	293	578	713	380
Total occurrences*	29,894	21,217	46,695	36,765	28,455
Number of documents*	2,912	627	1,196	1,890	4,438
Covariates per document*	10.3	33.8	39.0	19.5	6.4

Legend: * calibration data

5. Autocorrelated topic LDA with topic-specific covariates

The LDA model is commonly used to analyze text data. Our implementation of the LDA model includes the cutpoint regression model in equations (5) and (6), but retains the assumption that topics are associated with IID draws from the topic distribution indexed by θ_d for each word. The SC-LDA restricts all words within a sentence to originate from a single topic, while the AT-LDA models allow for topics to be autocorrelated within each sentence. The amount of autocorrelation is affected by covariates in equation (3) reflecting syntactic content. We fit two versions of the AT-LDA with covariates. One version specifies main effects for the covariates only (common covariates). In a second version, we allow for (all) interaction effects between covariates and topics (topic-specific covariates). This version of the AT-LDA allows the effect of, for example, a question mark on topic change to be different across topics whereas the main-effects models assumes this effect to be homogeneous. For model comparison, we report the (log) marginal likelihood of the data. To evaluate predictive performance, we report the log average likelihood of the hold-out data. To compute the likelihood of hold-out data, we first generate ζ_n, given observed covariates, z_{n-1} and $\beta^{(reg)}$. Then, given a realization of ζ_n, z_n is either carried over from z_{n-1} or independently generated from θ_d. θ_d for hold-out documents is generated from the prior. After assigning topics, we can compute the probabilities of words in hold-out documents using ϕ_z. Since the number of topics is not a parameter of our models, we estimate each model for a large range of T ($T \leq 50$) and choose the best-fitting model in terms of predictive fit for all further analysis.

Table 3 presents summary measures of model fit for the datasets. We find evidence that the standard assumption that words are generated IID from topic vocabularies is not supported by the data. The IID assumption is central to the LDA model. Instead, we find that the SC-LDA model and AT-LDA models fit the data better across all data sets. Across all data sets the improvement in in-sample fit of the AT-LDA without covariates over the SC-LDA is large (e.g. LMD of -558,165 as compared to -582,957 for

Table 3: Fit results.

Model	Manhattan Hotels	Italian Restaurants	American Restaurants	Camping Tents	Power Drills
LDA					
In sample	-430,532 (12)	-288,722 (8)	-581,127 (9)	-500,841 (10)	-366,756 (11)
Out of sample	-46,521 (12)	-30,390 (8)	-65,645 (9)	-59,970 (10)	-44,441 (11)
Sentence Constrained LDA					
In sample	-424,394 (12)	-287,485 (11)	-582,957 (11)	-508,884 (7)	-362,217 (9)
Out of sample	-46.479 (12)	-30,125 (11)	-65,398 (11)	-59,898 (7)	-43,855 (9)
AT-LDA intercept only					
In sample	-399,487 (19)	-281,379 (9)	-558,165(12)	-483,976 (7)	-351,623 (11)
Out of sample	-46,519 (19)	-30,323 (9)	-65,364 (12)	-59,929 (7)	-45,701 (11)
AT-LDA with common covariates					
In sample	-397,700 (19)	-281,078 (11)	-558,034 (12)	-481,066 (9)	-356,445 (12)
Out-of-sample	-46,492 (19)	-30,370 (11)	-65,407 (12)	-59,817 (9)	-45,571 (12)
AT-LDA with topic-specific covariates					
In sample	-398,362 (18)	-281,189 (10)	-558,676 (12)	-483,387 (8)	-350,966 (12)
Out-of-sample	-46,252 (18)	-30,118 (10)	-65,325 (12)	-59,587 (8)	-45,365 (12)

Legend: Numbers in parentheses indicate number of topics.

Table 4: Explained variance of customer rating.

Model	Manhattan Hotels	Italian Restaurants	American Restaurants	Camping Tents	Power Drills
LDA	0.558	0.786	0.737	0.628	0.603
SC-LDA	0.731	0.779	0.793	0.761	0.643
AT-LDA intercept only	0.669	0.835	0.800	0.710	0.547
AT-LDA common cov	0.626	0.819	0.818	0.682	0.511
AT-LDA topic-specific cov	0.645	0.837	0.829	0.717	0.581

American restaurant data). This suggests that a more flexible approach to modeling topic assignments than a sentence-constraint needs to be taken.

Given the role of structural covariates to our modeling approach, we note that introducing covariates leads to an improvement of in-sample and out-of-sample fit of the AT-LDA. For example, for the hotel data set, we obtain an in-sample LMD of the AT-LDA without covariates of -399,487. The introduction of covariates leads to an improvement to LMD of -397,700. The implied Bayes' factor of $e^{1,787}$ presents strong evidence in favor of the covariates model. We observe a similar improvement for the camping tents data set. It is interesting to note that, for the Italian restaurant data and the power drill data, an improvement in fit by introduction of covariates over the AT-LDA without covariates can only be observed for the model with interactions. It appears that accounting for topic-specific effects of covariates is critical for these reviews. Further below, we present a more extensive analysis of topic carry-over and the role of structural information therein.

Table 4 reports R^2 measures of fit for the cut-point regression of the overall rating on topic probabilities, θ_d (Eq. 5). Higher measures of R^2 are interpreted as reflecting truer topic probabilities and content. We find uniform improvement in the R^2 fit statistic relative to the LDA model, but that the AT-LDA model is not always superior to the SC-LDA model. That is, the AT-LDA outperforms the SC-LDA with respect to explaining the summary rating for the Italian and American restaurant data sets but results in a lower R^2

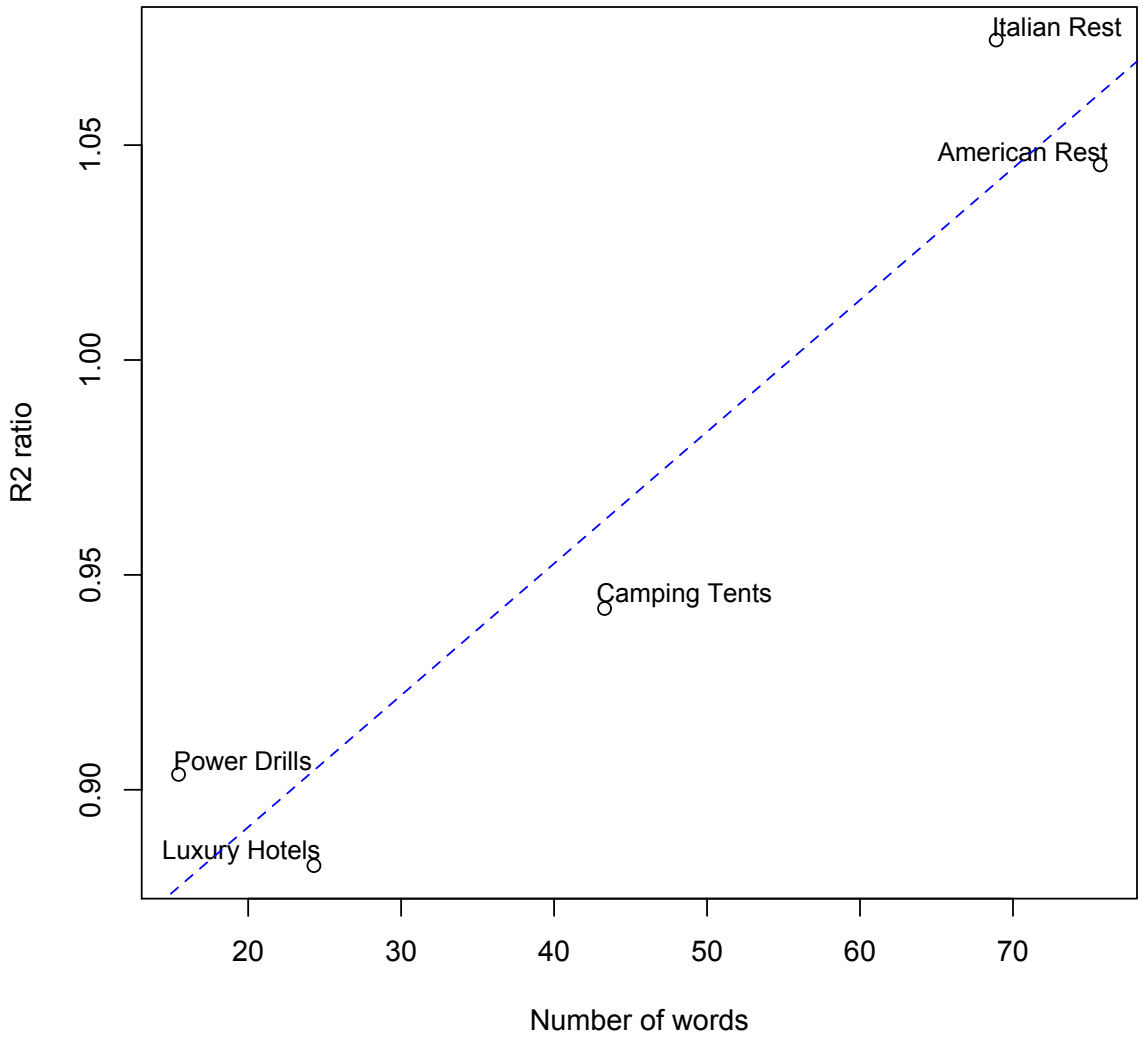
for the hotel, camping tents and power drill data. Across all data sets we find that allowing for topic-covariate interaction effects in the hierarchical regression leads to improved R^2 over the AT-LDA with main effects of the covariates. Figure 2 examines the improvement in predicting overall satisfaction for the autocorrelated topic LDA model with covariates to the sentence constrained LDA model (SC-LDA). Plotted is the ratio of R^2 measures relative to the average number of words in the review for each of the five datasets. We find that the R^2 ratio improves in an almost linear fashion as the number of words increases, indicating that the flexibility provided by our autocorrelated model is better realized in longer reviews than in shorter reviews. Text data that exhibits low complexity (short reviews, small vocabulary) are in greater agreement with the assumption that topics are sentence specific. For more complex text data (longer reviews, larger vocabulary), a more flexible approach may be necessary. In such data, topics may exhibit a more complex form of serial dependency than simply sentence-based.

4.2 Investigating topic carryover

We illustrate insights from our model regarding topic carryover using reviews from the American restaurant dataset. Four reviews are displayed in Figure 3. Words shown in grey are those that are eliminated from the review during pre-processing of the text. Those shown in black are words associated with a new topic, and those in color are have topics that carryover from the previous word. Words that are underlined are used as covariates $\tilde{x}_{n-1}$ affecting the degree of topic carryover. The final period (i.e., full stop) is shown in parenthesis and is ignored in each review because it inform neither the topics, nor their probabilities.

The top two reviews in figure 3 provide examples of topic generation and carryover within each sentence. In review 180, the words "pleasant" and "experience" are estimated to come from the same topic, while the words "lots-atmosphere-owner-working-visiting-people" were estimated to come from another topic in the same sentence. The first

Figure 2: Performance comparison of AT-LDA vs. SC-LDA. Models are compared with respect to the relative variance explained of the (latent) customer rating.



Legend: Vertical axis shows ratio of R^2 from AT-LDA with topic-specific covariates over R^2 from SC-LDA. A ratio smaller than 1 indicates that R^2 from the SC-LDA exceeds the R^2 from the AT-LDA and vice versa. Horizontal axis shows average number of words in reviews per data set. The dashed blue line indicates the trend obtained by regressing the R2 ratio on the (average) number of words in each data set.

sentence is estimated to come from two topics instead of one that would have been imposed by the sentence constrained LDA model. Review 264 also exhibits varying serial dependency in topic assignment, from "good-food" to "fine-dining-restaurants-area".

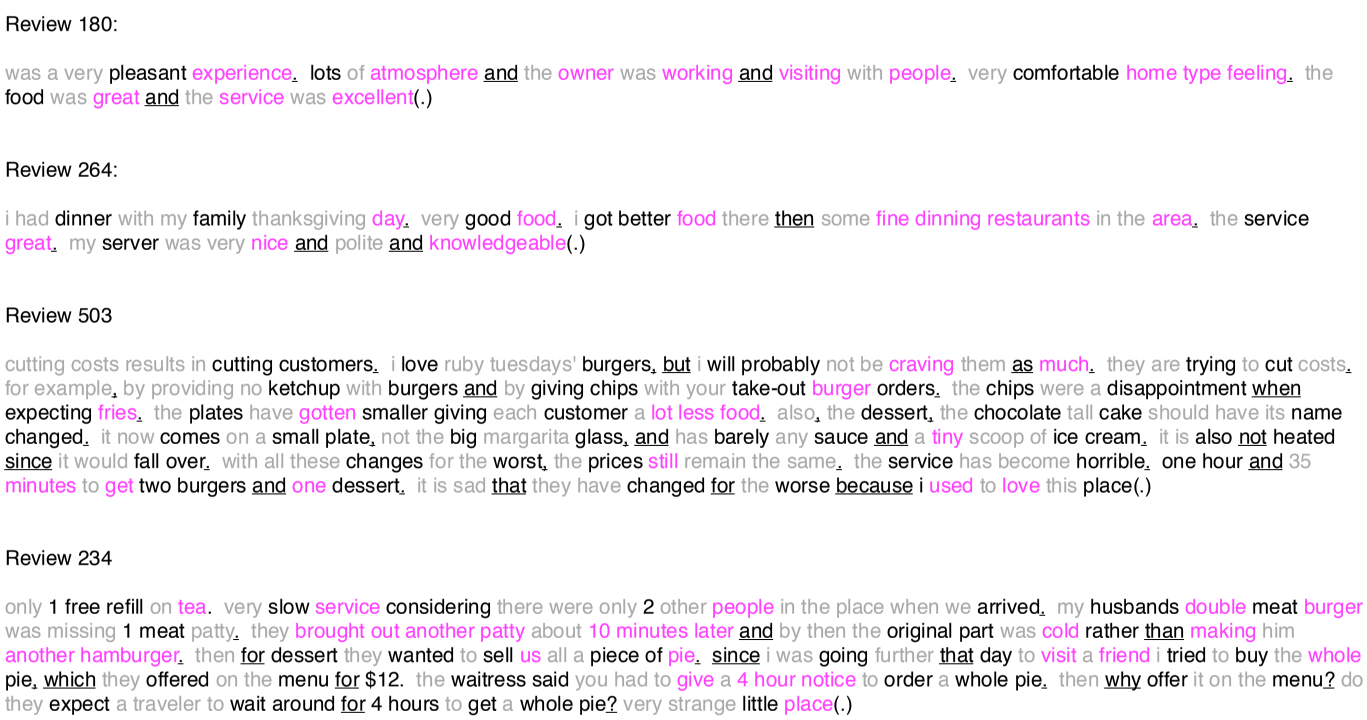
The bottom two reviews are more complex. In review 503, less than 30% of the topic assignments are due to carry-over suggesting that results approach those from a standard LDA. However, those instances of carryover observable in this review ("customer-lot-less-food" and "hour-minutes-get") tie words that clearly belong together. In review 234, we find that about half of topic assignments are due to carry-over with many of these being instances of a single topic carry over ("slow-service," "part-cold," "sell-us," "little-place"). The larger reviews exhibit different types of topic carry-over and a flexible model of topic assignment is needed. In some reviews, topics are consistently carried across complete sentences. In others, topic assignments are only locally dependent. Our model with autocorrelated topics can accommodate both situations.

4.3 Covariates affecting topic carryover

An important question in our analysis is the extent to which syntactic elements of text drive local topic dependency that we use to modify the probability of topic carryover. We start by considering the marginal probabilities of a topic carry-over for each of the data sets. Figure 4 shows that ψ differs greatly among the topics, typically ranging from 10% to 60%. Across all topics and data sets, the average topic carry-over probability is 45%, suggesting that the assumption of topic independency of the standard LDA model (equivalently: $\psi = 0$) is untenable.

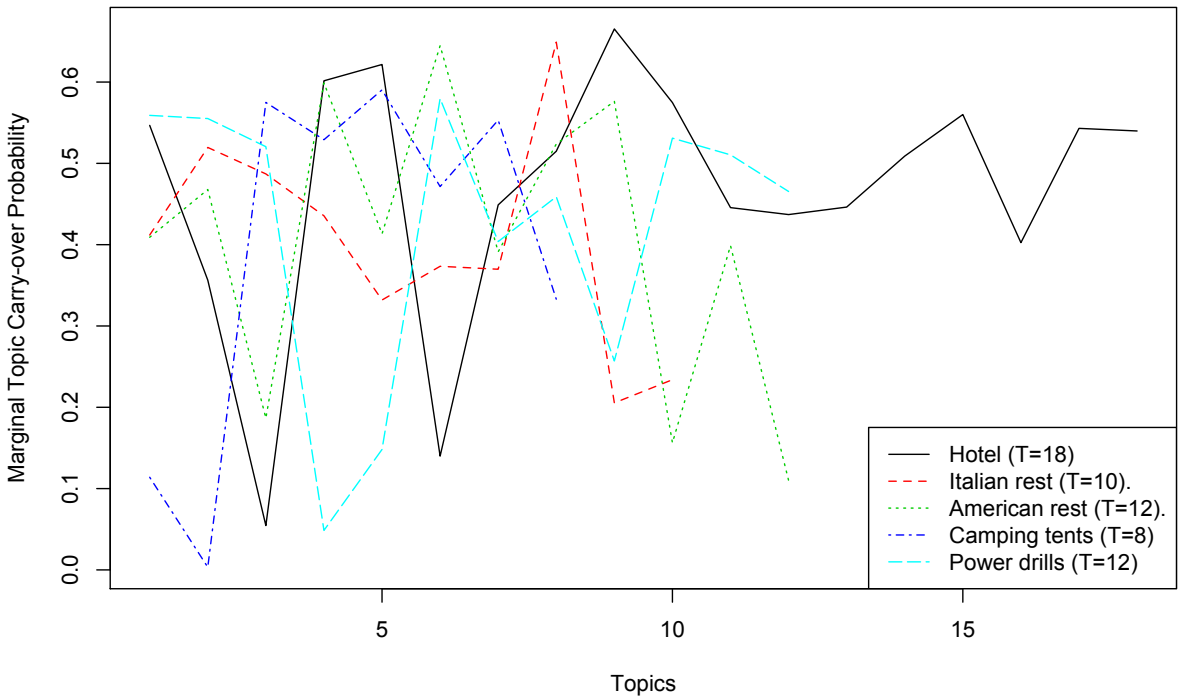
Our model allows for conjunctions and punctuation to change the probability of topic carry-over. Figure 5 and Figure 6 report how ψ changes as a result of the presence of selected covariates. As an illustrative example, we use results from the American restaurant data. Figure 5 reveals that full stops, exclamation marks and question marks have a significant influence on the probability of a topic carry-over. The presence of a

Figure 3: Topic carry-over from the AT-LDA model with covariates. Four selected reviews from American restaurant data set. Results from main-effects HAT-LDA with $T = 12$.



Legend: Examples shown are original reviews in which words greyed out indicate data eliminated in pre-processing. Words and signs underlined indicate covariates to following term. Words highlighted in color carry the topic of the preceding word (in pre-processed data). The full stop at the end of the last sentence in each reviews is ignored because it cannot influence downstream topic carry-over .

Figure 4: Marginal topic carry-over probabilities (ψ_t).



Legend: Reported are the posterior means of the topic carry-over probabilities, marginalized with respect to the presence of covariates. Results reported are from the AT-LDA with covariates and interaction effectss, given optimal T (Table 3).

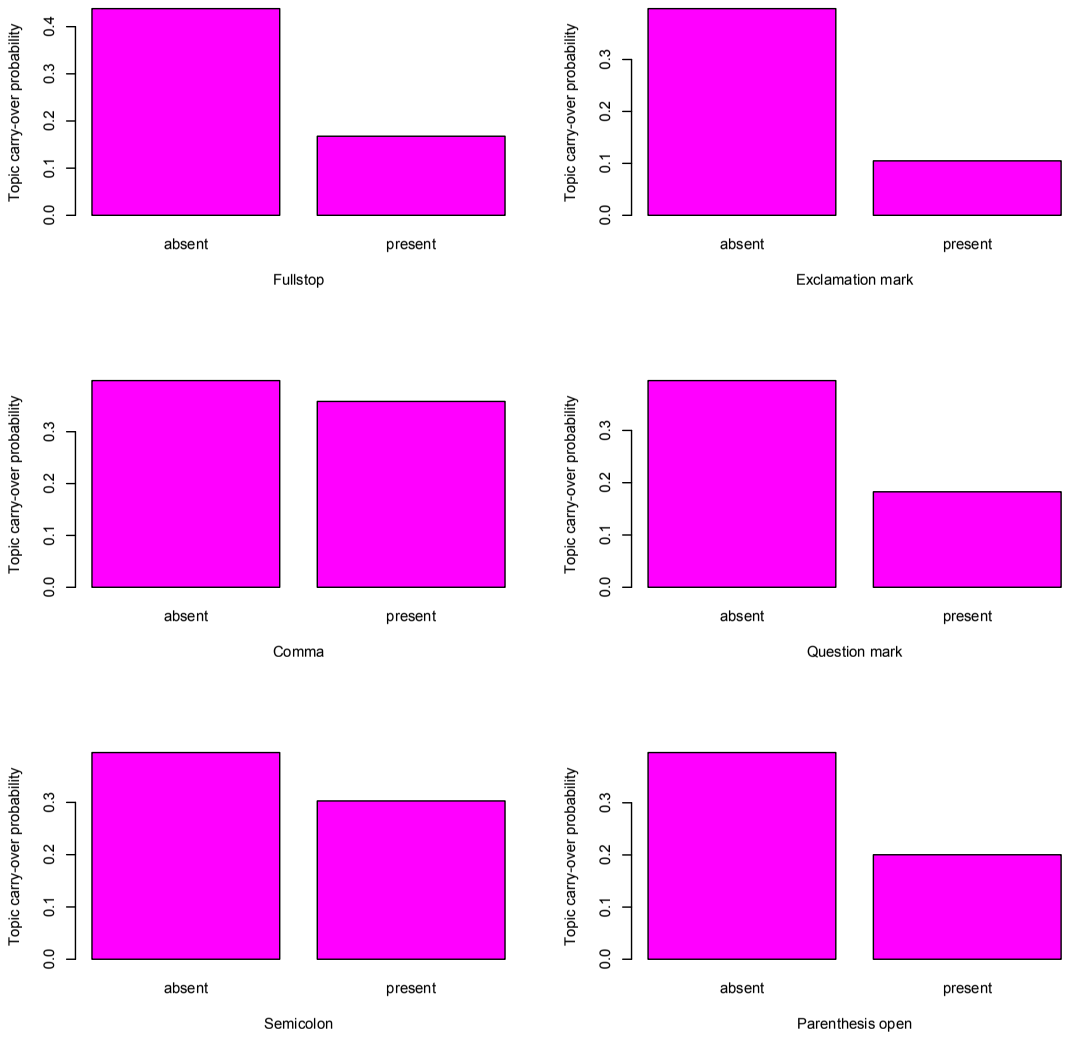
full stop in front of a word reduces the probability of a topic carry-over from the previous word from 44% to 17%. An exclamation mark reduces this probability from 40% to 10%, a 75% reduction in probability. A question mark cuts ψ in half (40% to 20%).

We obtain similar results from conjunctions. Figure 6 displays changes in the carryover probabilities. The conjunction "that" induces ψ to decrease from 40% to 21%. The conjunction "and," a coordinating joiner of parts of speech, increases the marginal probability of a topic carry-over from 39% to 43%. We observe the largest effect among conjunctions for the term "because" which reduces ψ from 39% to 9%. In general, we find that the marginal effect of structural elements in text to topic dependency is negative. That is, the presence of conjunctions or punctuation reduces autocorrelation in topics.

Table 5 presents results from the regression of ψ on the observed structural covariates for all data sets. To reduce clutter, we present results from the AT-LDA main effects model with common covariates. All coefficients listed in the table are "significant" in that 95% of their posterior mass is away from zero. Table 5 confirms that the probability of a topic carry-over is strongly influenced by the structure induced through conjunctions and punctuation. We find that only few of the structural covariates do not drive this probability as evidenced by coefficients not credibly different from zero. In general, we find that most (with respect to topics, marginal) coefficients are negative, indicating that the probability of carryover is decreased.

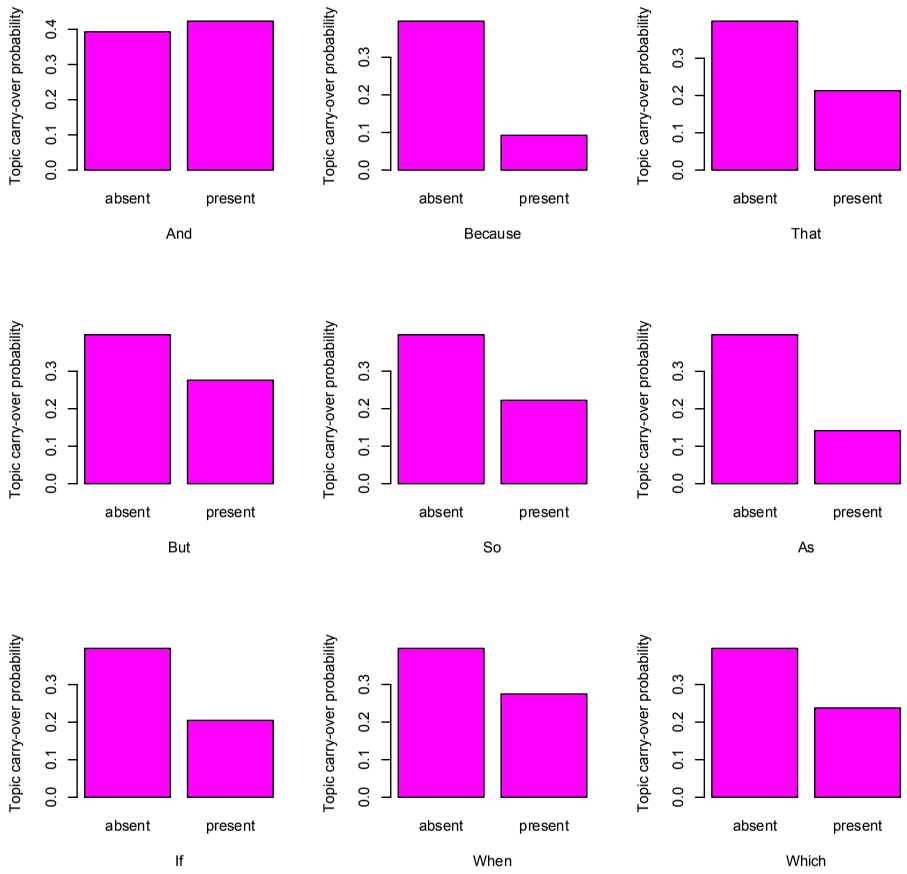
Since all covariates are dummy-coded, we can directly compare their influence by coefficient size. The largest influence on carry-over is exerted by a full stop ($\beta^{(.)}_{Tents} = -5.12$, $\beta^{(.)}_{Hotels} = -4.9$). The negative sign implies that a (preceding) full stop is associated with a lower probability of a carryover or, equivalently, a higher probability of an independent topic draw for the focal word from the prior. In a similar way, but less strongly, a comma reduces the probability of a topic carry-over to the focal word ($\beta^{(,)}_{Hotels} = -1.34$, $\beta^{(,)}_{Tents} = -1.49$). Interestingly, question marks and semicolons differ in their influence across the data sets. Question marks, in the luxury hotel data set, increase the probabil-

Figure 5: Marginal topic change probabilities given incidents of punctuation.



Legend: Probability of a topic change, given incidents of most frequent forms of punctuations (full stops, commas etc). Reported are probabilities marginalized with respect to the presence of all other covariates and topics. Results reported are from main-effects AT-LDA model, American restaurant data ($T = 12$).

Figure 6: Marginal topic change probabilities given incidents of conjunctions.



Legend: Probability of a topic change, given incidents of most frequent types of conjunctions. Reported are probabilities marginalized with respect to the presence of all other covariates and topics. Results reported are from main-effects AT-LDA model, American restaurant data ($T = 12$).

ity of carry-over ($\beta^{(?)}_{Hotels} = 0.12$), suggesting that the topic of the narrative continues. A possible explanation is that authors answer the question they raise. In the Italian restaurant data, the opposite is true ($\beta^{(?)}_{Ital} = -1.36$). A possible explanation is that questions raised in these reviews are of a rhetorical nature ("why come back?").

It is interesting to compare these results to Büschken and Allenby (2016) who consider carry-over across sentences but constrain topics to be homogeneous within sentences. In their analysis, sentence boundaries are indicated by full stops, commas, exclamation marks and question marks. The authors find that the probability of topics to carry-over is virtually zero. A reasonable explanation for the null result is that the model is concerned with the carry-over of topics across sentences, not words. We find that topic carry-over probabilities across words are mostly high, but typically reduced by sentence boundaries indicated by full stops, semicolons and colons. Exceptions are question marks in the hotel and the power drill data set which increase the likelihood of a topic carry-over.

With respect to the use of conjunctions in reviews, we find that conjunctions have a mixed influence on topic dynamics, suggesting that thematic context plays a role. Some conjunctions have a uniformly negative influence on carry-over. As an example, consider the term "because" which reduces ψ across all data sets as indicated by uniformly negative regression coefficients. Use of this term suggests that authors offer an explanation or justification for a preceding statement. Often, the cause of some event or situation is topically independent from the event as such ("Our table reservation had been canceled by the restaurant and the table given away because we arrived late"). Our results suggest that the topic of an explanation (or more specifically: the topic of the words that follow the term "because") has a higher probability to be generated independently from the topic of the statement. In a similar fashion, coordinating conjunctions such as "for" or "that" exhibit uniformly negative betas. In our data, they join parts of speech that exhibit low topical dependency.

Other conjunctions exhibit a more context-specific influence on topic dynamics. An

interesting example of context-specific influence is presented by the term "although". This term drives topic carry-over in different ways, given data sets. In the luxury hotel data set, its influence on carry-over is positive ($(\beta_{Hotels}^{(although)} = 0.69)$) whereas, in the Italian restaurant data set, its coefficient is negative ($(\beta_{Hotels}^{(altough)} = -1.64)$). The term "although" expresses surprise or something unusual or unexpected. It appears that when a hotel reviewer expresses surprise, she tends to do so in the context of the same topic. When a restaurant reviewer talks about something unexpected, this typically initiates a topic change. A possible explanation is that surprises in restaurants are rarely appreciated. The use of the term "although" then marks a change in the narrative explaining specific issues with the service.

In summary, we find that structural information present in the use of conjunctions and punctuation in text data presents important prior information to dynamics in latent topics. It is curious that this information, so easily available, is typically treated as noise in applications of topic models. This can be explained by the use of topic models as devices to find latent topics in a set of documents, not as devices to identify topic dynamics and its drivers within documents. The AT-LDA model is capable of combining these two perspectives.

4.4 Cut-point regression results

All models estimated in our analysis are supervised LDA models where customer ratings are considered part of the data likelihood. We relate the ratings to the latent topic proportions (θ_d) using an ordinal probit (cut-point) model (Rossi et al. 2001; Johnson and Albert 2006; Büschken et al. 2013). We compare topic and regression results of the standard LDA model to our proposed auto-correlated model (AT-LDA) using the two restaurant datasets for which our model produces superior results (see Table 2). Tables (6) and (7) display the most frequent terms for each topic from the LDA for these two data sets. At the top of each table, we provide a summary description of the most frequent

Table 5: Results from the hierarchical regression for AT-LDA main-effects covariate model (posterior means). Coefficients not credibly different from 0 are indicated as "n.s.". Coefficients for topic dummy variables are omitted for brevity.

Covariate	Manhattan Hotels	Italian Restaurants	American Restaurants	Camping Tents	Power Drills
Conjunctions					
for	-1.955	-1.605	-0.974	-0.843	-0.584
and	-0.613	-0.353	-0.277	-0.674	-0.747
nor	0.805	-0.943	0.385	0.835	-0.142
but	-1.704	-1.402	-1.191	-2.859	-0.717
or	0.461	-0.275	n.s.	-0.170	n.s.
yet	-1.131	-0.630	0.218	-0.931	0.227
so	-1.493	-1.558	-1.255	-3.015	-0.412
after	-0.938	-0.657	-0.371	-1.231	-0.491
although	0.685	-1.644	-0.413	0.543	0.424
as	-1.628	-1.308	-0.529	-2.059	-0.258
because	-0.995	-2.433	-0.513	-1.375	-0.481
before	-0.736	-0.623	-0.822	-0.921	0.276
even	0.689	-0.769	n.s.	n.s.	-0.437
if	-2.379	-1.672	-0.802	-2.244	-0.227
now	-0.955	n.s.	n.s.	-0.684	n.s.
once	-0.447	-0.287	0.361	-1.308	-0.244
provided	-0.341	-0.266	0.224	-0.126	0.734
since	-0.710	-0.252	-0.568	-0.688	n.s.
than	-1.097	-1.450	-0.909	-0.518	-0.948
that	-1.819	-2.571	-1.366	-1.769	-1.188
though	-0.753	-0.602	n.s.	-0.167	0.200
unless	0.298	-1.046	-0.556	-1.292	0.264
until	-0.737	-0.593	-0.181	n.s.	-0.469
when	-1.482	-1.409	-0.360	-1.654	-0.671
whenever	0.237	0.606	0.067	0.258	n.s.
where	-1.034	-0.784	-0.569	-0.475	0.451
whether	-1.138	-0.831	-0.154	n.s.	n.s.
which	-1.536	-1.728	-1.664	-1.365	-0.656
while	-1.416	-1.381	0.188	-1.077	-0.209
who	-0.676	-0.268	-0.686	-0.892	0.248
whoever	NA	-0.939	-0.198	NA	NA
why	-0.364	1.337	0.200	-0.674	-0.089
what	-1.312	-0.797	-0.836	-0.550	n.s.
whom	0.853	-0.801	-0.594	-1.606	NA
whose	-0.642	-0.763	0.084	NA	0.632
Punctuation					
,	-1.337	-0.670	-0.428	-1.488	-1.299
.	-4.904	-2.829	-2.276	-5.155	-3.321
;	-1.673	-1.254	-0.361	-1.374	n.s.
:	-1.583	0.326	0.088	-1.710	-0.681
!	NA	-1.896	-1.466	-3.026	-1.287
?	0.118	-0.431	n.s.	-1.358	0.060
&	0.392	1.734	0.519	0.194	n.s.
(	-0.744	-2.349	-0.892	-1.026	-0.623
)	-1.738	-1.504	-0.781	-1.700	-0.442

terms under each topic. For comparison, Tables (8) and (9) display the topics that emerge when applying the AT-LDA.

The LDA model tends to allocate food-related terms from the restaurant data sets evenly across the topics. For the American restaurant data (Table 6), 6 out of 9 topics contain words that describe food items (e.g., "chicken", "toast", "coffee"). Topic 3 in this data set contains words describing specific types of cuisine ("burgers", "fries"), and three other topics describe additional menu items (e.g., "salad", "wings" etc.). The remaining topics contain terms related to food in some way (i.e., "food", "menu", "taste", "flavor"). Similarly, for the Italian restaurant dataset (Table 7), all topics contain terms pertaining to food items (e.g. "mozzarella", "salad", "cheese") or food in general ("food", "kitchen", "menu").

Table (8) presents the most frequent terms from the American restaurant data for the AT-LDA model. In contrast to the LDA model, we find that items and ingredients to meals are concentrated to three out of 12 topics (Topics 6, 9 and 10). Two other topics (Topics 2 and 3) talk about positive vs. negative aspects of the dining experience without any reference to food items. Topic 11 from the AT-LDA collects terms that describe price or value ("price","prices", "portions") and Topic 7 describes various occasions for visiting a restaurant ("dinner","lunch", "breakfast"). The LDA does not identify similar topics describing occasions for visiting or value.

Similar results are found in the Italian restaurant data (Table 9). The AT-LDA concentrates menu items from this data set to two topics (Topic 5: "Pizza", Topic 8: "Ingredients"). Topics 4 and 9 from Italian restaurant reviews talk about wait times (for table, for service) and issues with the order, respectively, both of which describe negative service experiences. Again, from the LDA, no such topics are discernible. Topic 10 from the AT-LDA expresses a situation of conflict between patrons and service employees that escalates to the manager. This topic, too, is not identified by the LDA although its relevance to the overall evaluation of a dining experience is apparent. In summary, applying the LDA

model to the restaurant data sets results in topics that exhibit significant overlap with respect to food and menu items. The AT-LDA model, in comparison, finds a larger and more differentiated set of topics and words more exclusive to topics (Airoldi and Bischof (2016)). The AT-LDA identifies topics clearly unrelated to food or menu items and seems more powerful in finding topics that describe negative aspects of the dining experience or the value of a restaurant visit. Such topics seem to more relevant to identifying potential drivers to the customer satisfaction rating.

Tables (10) and (11) summarize results from regressing the customer rating on topic shares. Note that the regression model is not a priori identified because the topic shares from an LDA-type model, by definition, sum to 1. We post-process the coefficients by setting the coefficient of one arbitrarily chosen topic to zero. From applying the LDA to American restaurant data (Table 10), we find that only the topic that talks about the waitress drives the rating down, whereas all topics pertaining to menu items emerge as positive drivers of the overall rating. In comparison, results from the AT-LDA suggest that menu items have less of an influence and that an increasing share of the topic related to a bad experience has a strong negative influence on the rating. This suggests that the LDA inflates the role of item-related topics and does not adequately capture the role of a negative dining experience with respect to the satisfaction rating. An even more interesting result emerges from comparing cut-point regression results from the Italian restaurant data (Table 11). The supervised LDA model leads to all topics, except for Topic 1 ("Atmosphere"), being credibly negative drivers of the overall rating. In the case of topics such as "Good food", "Pizza", or "Menu", this result has little face validity as these topics have either positive or neutral valence. The AT-LDA, in comparison, identifies only 3 out of 9 topics as significant drivers. The topic "escalated conflict" emerges as the only negative driver of the rating, whereas the topics "Best Restaurant" and "Great Experience" influence the rating positively. This suggests that the LDA may identify topics as significant drivers that are difficult to interpret using a small number of

Table 6: American restaurant data. Most frequent terms (top 20), given ϕ, from LDA Model, T=9.

Rank	Topic 1 Function words	Topic 2 Good restaurant	Topic 3 Burger & fries	Topic 4 Items 2	Topic 5 Waitress	Topic 6 Items 3	Topic 7 Items 4	Topic 8 Good place	Topic 9 Great service
1	just	good	burger	really	us	wings	salad	bar	food
2	got	breakfast	fries	sandwich	food	food	ordered	place	great
3	one	restaurant	food	fries	waitress	flavor	menu	good	service
4	chicken	food	burgers	came	table	cheese	restaurant	food	restaurant
5	get	buffet	cheese	little	service	fresh	wine	great	staff
6	like	coffee	just	good	restaurant	chicken	served	nice	place
7	back	one	dog	like	back	sauce	steak	go	atmosphere
8	time	day	like	lot	will	taste	dinner	night	best
9	much	well	hot	got	time	sweet	dessert	drinks	always
10	two	eggs	good	us	never	bacon	good	area	friendly
11	going	place	place	didnt	minutes	one	two	atmosphere	will
12	place	dinner	best	cheese	asked	made	try	room	family
13	little	lunch	places	chicken	one	can	chocolate	service	excellent
14	diner	eat	want	just	order	love	appetizer	just	experience
15	didnt	home	get	burger	manager	hot	us	also	wonderful
16	order	better	go	menu	said	perfect	time	pizza	new
17	can	local	can	bar	go	beer	bread	quite	well
18	wasnt	toast	also	pretty	got	spot	dining	table	time
19	food	sunday	make	back	another	delicious	cake	around	can
20	couple	family	little	ordered	take	tasty	small	eat	recommend

Table 7: Italian restaurant data. Most frequent terms (top 20), given ϕ, from LDA model, T=8.

Rank	Topic 1 Atmosphere	Topic 2 Function words	Topic 3 Pizza	Topic 4 Items	Topic 5 Menu	Topic 6 Bar & wait	Topic 7 Order	Topic 8 Good food
1	food	got	pizza	sauce	italian	can	food	good
2	great	really	crust	fresh	sandwich	bar	ordered	salad
3	restaurant	two	pizzas	italian	menu	wait	restaurant	food
4	service	just	like	dish	cheese	area	us	one
5	place	one	cheese	dinner	beef	staff	back	bread
6	best	came	good	wine	also	dining	service	like
7	italian	didnt	chicago	delicious	area	like	will	place
8	family	back	really	made	fries	well	never	pasta
9	atmosphere	little	thin	fish	food	enough	minutes	dont
10	excellent	cheese	one	dessert	lunch	take	asked	much
11	go	get	place	house	order	business	waiter	served
12	always	good	get	us	lettuce	customers	arrived	better
13	will	us	slice	visit	fresh	dont	told	meal
14	one	went	just	enjoyed	small	course	order	sauce
15	good	said	best	perfect	mozzarella	sure	owner	try
16	recommend	bar	order	cream	decor	get	time	italian
17	definitely	waitress	style	shrimp	home	makes	experience	restaurant
18	friendly	room	little	large	mayo	experience	manager	also
19	wonderful	took	new	fried	hot	kitchen	went	menu
20	many	pretty	lot	everything	bun	view	dinner	dressing

Table 8: American restaurant data. Most frequent terms (top 20), given ϕ, from AT-LDA Model, T=12.

	Topic 1 Go back	Topic 2 Good experience	Topic 3 Bad experience	Topic 4 Layout	Topic 5 Best restaurant	Topic 6 Items 1	Topic 7 Occasions	Topic 8 Waitress	Topic 9 Items 2	Topic 10 Best food	Topic 11 Price/ value	Topics 12 Function words
1	go	food	food	bar	restaurant	fries	dinner	us	salad	hot	food	really
2	will	great	service	room	best	cheese	lunch	waitress	fresh	food	good	just
3	back	service	restaurant	dining	new	sandwich	night	came	ordered	wings	menu	like
4	can	staff	husband	tables	years	chicken	breakfast	minutes	cream	dog	great	got
5	place	good	manager	area	times	burger	day	order	chocolate	burgers	bar	good
6	get	atmosphere	one	street	time	came	time	table	chicken	taste	restaurant	little
7	like	place	ordered	back	ever	bacon	meal	took	cheese	flavor	price	didnt
8	eat	friendly	will	right	many	sauce	hour	back	steak	best	prices	wasnt
9	didnt	nice	well	restaurant	one	meat	eggs	brought	cake	burger	small	pretty
10	make	experience	told	table	place	onion	sunday	got	crab	well	place	one
11	going	restaurant	never	parking	area	barbecue	went	drinks	rib	spot	pretty	ordered
12	try	always	wife	near	visit	bun	two	asked	dessert	chicago	items	went
13	want	excellent	however	lot	family	bread	buffet	food	ice	great	sandwiches	burger
14	time	well	asked	building	eat	served	half	time	delicious	made	reasonable	still
15	never	wait	experience	front	town	fresh	special	drink	sweet	love	also	menu
16	find	wonderful	like	door	places	little	one	get	potatoes	location	large	much
17	just	recommend	bad	inside	eaten	lettuce	saturday	told	served	sauce	fast	said
18	say	pleasant	just	side	friends	fried	friday	another	sauce	dogs	average	meal
19	know	dining	server	space	restaurants	cut	place	waiter	potato	one	portions	get
20	see	family	2	wall	last	side	coffee	check	cooked	place	quality	also

Table 9: Italian restaurant data. Most frequent terms (top 20), given ϕ, from AT-LDA Model, T=10.

Rank	Topic 1 Best Restaurant	Topic 2 Layout	Topic 3 Great Experience	Topic 4 Wait	Topic 5 Pizza	Topic 6 Go back	Topic 7 Menu	Topic 8 Ingredients	Topic 9 Order Issues	Topic 10 Escalated Conflict
1	italian	bar	food	minutes	pizza	go	menu	sauce	got	food
2	best	room	great	time	crust	can	good	cheese	really	us
3	pizza	dining	service	table	good	will	wine	bread	just	restaurant
4	restaurant	area	restaurant	night	cheese	place	salad	fresh	one	service
5	new	right	place	dinner	pizzas	dont	pasta	garlic	came	never
6	one	small	good	us	like	back	also	salad	two	ordered
7	chicago	tables	atmosphere	wait	thin	get	dinner	tomato	took	asked
8	style	restaurant	staff	order	pretty	like	meal	pasta	went	owner
9	food	lot	friendly	years	little	know	ordered	mozzarella	get	waiter
10	restaurants	back	experience	first	slice	didnt	dish	parmesan	us	said
11	many	parking	recommend	last	place	try	large	served	good	told
12	ever	kitchen	always	party	order	make	dessert	dressing	sandwich	came
13	beef	located	excellent	day	slices	eat	served	sausage	waitress	experience
14	york	street	family	next	really	say	portions	spaghetti	said	just
15	great	building	nice	every	much	time	salads	flavor	didnt	manager
16	favorite	just	prices	take	two	going	lunch	ravioli	wanted	server
17	places	front	will	long	large	think	small	red	ordered	bill
18	eaten	counter	well	15	pepperoni	just	dishes	baked	menu	bad
19	area	side	wonderful	people	pie	want	selection	italian	like	waitress
20	dish	around	also	later	just	better	portion	cream	meal	however

words with high probabilities as reported in Tables (6) and (7).

Table 10: American restaurant data: Results from topic regression. Results from best-fitting model with respect to number of topics.

Parameter	Topic	LDA Posterior Mean	Topic	AT-LDA Posterior Mean
Regression coefficients				
β_0	Intercept	**-1.276**	Intercept	-0.554
β_1	Function words	0.843	Go back	0.001
β_2	Good restaurant	0.415	Good experience	**4.965**
β_3	Burgers & fries	**1.051**	Bad experience	**-4.425**
β_4	Items 2	**0.886**	Layout	0.00*
β_5	Waitress	**-2.715**	Best restaurant	**2.194**
β_6	Items 3	**2.366**	Items 1	**1.159**
β_7	Items 4	**1.446**	Occasions	0.630
β_8	Good place	0.00*	Waitress	**-1.452**
β_9	Great service	**5.387**	Items 2	**2.761**
β_{10}	-	-	Best food	0.819
β_{11}	-	-	Price/value	-0.639
β_{12}	-	-	Functions words	-0.127
Cut-points				
c_4		0.133*		0.133*
c_3		**-0.559**		**-0.562**
c_2		**-1.180**		**-1.200**
c_1		-1.578*		-1.578*
R^2		0.737		0.829

Legend: *: Indicates parameter fixed for identification. Parameters credibly different from zero in boldface.

Table 11: Italian restaurant data: Results from topic regression. Results from best-fitting model with respect to number of topics.

Parameter	Topic	LDA Posterior Mean	Topic	AT-LDA Posterior Mean
Regression coefficients				
β_0	Intercept	**2.168**	Intercept	0.009
β_1	Atmosphere	**2.054**	Best restaurant	**2.575**
β_2	Function words	**-2.975**	Layout	0.00*
β_3	Pizza	**-2.319**	Great experience	**4.783**
β_4	Items	0.00*	Wait	0.392
β_5	Menu	**-2.277**	Pizza	-0.692
β_6	Bar & wait	**-3.011**	Go back	-0.501
β_7	Order	**-5.952**	Menu	0.395
β_8	Good food	**-4.805**	Ingredients	-0.128
β_9	-	-	Order issues	-0.743
β_{10}	-	-	Escalated conflict	**-6.877**
Cut-points				
c_4		0.128*		0.128*
c_3		**-0.509**		**-0.476**
c_2		**-1.106**		**-1.071**
c_1		-1.643*		-1.643*
R^2		0.786		0.837

Legend: *: Indicates parameter fixed for identification. Parameters credibly different from zero in boldface.

5 Concluding Remarks

In this paper we examine the use of an autocorrelated topic model for analyzing text data. Topics are autocorrelated when they can carry-over from word to word in speech, and in our empirical analysis we find that it outperforms standard topic models across a variety of data sets. The reason for this result is that topic carry-over is a regular feature of customer review text data that standard topic models cannot account for. Although the IID assumption of topic assignment in LDA models has been criticized in the literature as unrealistic before, we provide model-based evidence for violation of this assumption and a way to solve this problem.

In our application of the model to different datasets, we examine the role played by conjunctions and punctuation in signaling topic change. The difference between these two categories of covariates is that conjunctions are joiners of speech and incidents of punctuation present natural separators of speech. Because we incorporate this information as covariates in the model, we can use it without compromising inference with respect to the topics themselves. In our empirical analysis, we find these syntactic covariates to be highly predictive of topic carry-over. Typically, conjunctions and punctuation are removed prior to the model-based analysis of text data. The primary motive for this "pre-processing" is that such data are not diagnostic with respect to topics. While this is true, our results suggests that syntactic covariates are highly diagnostic of topic changes and, through this mechanism, are useful in analyzing the latent structure of text. In short, our results present a strong case to not discard this data.

From a practical perspective, we find that a model with autocorrelated topics improves driver analysis of satisfaction if reviews are more complex (more words, larger vocabulary). This result can guide managers in moving away from simpler models when these do not suffice. Compared to results obtained via the standard approach to topic analysis (LDA), our model with autocorrelated topics generally results in a larger and more diverse set

of topics when applied to datasets comprised of larger sentences and more words. In our analysis of the restaurant review data, we find that the AT-LDA identifies topics more focused on specific themes (e.g. dining occasions, value-for-money, negative service experience) that the LDA does not identify with similar clarity. This is because the LDA has a tendency to allocate ubiquitous terms (food, menu items) more uniformly across the topics. This tendency also reduces its ability to explain customer satisfaction ratings.

To conclude, this research suggests that the analysis of serial topic dependency presents a fruitful area of future research to advance the use of topic models for the rapidly increasing amount of text data in marketing. The model proposed here is relatively simple in that it considered first-order topic dependency across an observed sequence of words only. Yet, it outperforms a model with IID topic draws across all five data sets analyzed (see Table 3). A casual inspection of results from our model (Figure 3) suggests that topic dependency may extend across longer sequences of words, suggesting the need for a more complex model of serial topic dependency. In particular, it would be interesting to investigate models that would allow us to identify at which point in the data (e.g., which word in a sentence or paragraph) the latent topic changes. In marketing and econometrics, change-point models for various types of latent discrete states have been developed (Chib 1998; DeSarbo et al. 2004; Fader et al. 2004; Netzer et al. 2008; Gopalakrishnan et al. 2016). Blending this type of dynamic modeling of latent structure with topic analysis presents a natural starting point for such an investigation.

References

Airoldi, Edoardo M., Jonathan M. Bischof. 2016. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association* **111**(516) 1381–1403.

Blei, David M, John D Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 17–35.

Blei, D.M., Yew-Kwang Ng, M.I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3** 993–1022.

Büschken, J., T. Otter, G.M. Allenby. 2013. The dimensionality of customer satisfaction survey responses and implications for driver analysis. *Marketing Science* **32**(4) 533–553.

Büschken, Joachim, Greg M Allenby. 2016. Sentence-based text analysis for customer reviews. *Marketing Science* **35**(6) 953–975.

Chib, Siddhartha. 1998. Estimation and comparison of multiple change-point models. *Journal of econometrics* **86**(2) 221–241.

DeSarbo, Wayne S, Donald R Lehmann, Frances Galliano Hollman. 2004. Modeling dynamic effects in repeated-measures experiments involving preference/choice: An illustration involving stated preference analysis. *Applied Psychological Measurement* **28**(3) 186–209.

Fader, Peter S, Bruce GS Hardie, Chun-Yao Huang. 2004. A dynamic changepoint model for new product sales forecasting. *Marketing Science* **23**(1) 50–65.

Gopalakrishnan, Arun, Eric T. Bradlow, Peter S. Fader. 2016. A cross-cohort changepoint model for customer-base analysis. *Marketing Science* **-**(Articles in advance) –.

Griffiths, Thomas L, Mark Steyvers, David M Blei, Joshua B Tenenbaum. 2004. Integrating topics and syntax. *Advances in neural information processing systems*. 537–544.

Johnson, Valen E, James H Albert. 2006. *Ordinal data modeling*. Springer Science and Business Media.

Meyer, Charles F. 1987. *A linguistic study of American punctuation*. P. Lang.

Nallapati, Ramesh, James Allan. 2002. Capturing term dependencies using a language model based on sentence trees. *Proceedings of the eleventh international conference on Information and knowledge management*. 383–390.

Netzer, Oded, James M Lattin, V Srinivasan. 2008. A hidden markov model of customer relationship dynamics. *Marketing Science* **27**(2) 185–204.

Rossi, Peter E., Zvi Gilula, Greg M. Allenby. 2001. Overcoming Scale Usage Heterogeneity. *Journal of the American Statistical Association* **96**(453) 20–31.

Say, Bilge, Varol Akman. 1996. Current approaches to punctuation in computational linguistics. *Computers and the Humanities* **30**(6) 457–469.

Tirunillai, Seshadri, Gerard J Tellis. 2014. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research* **51**(4) 463–479.

Trusov, Michael, Liye Ma, Zainab. Jamal. 2016. A cross-cohort changepoint model for customer-base analysis. *Marketing Science* **35**(3) 405–426.

Wallach, Hanna M. 2006. Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning*. ACM, 977–984.

A Appendix

A.1 Simulation Study: Empirical Identification of the AT–LDA topic model with covariates

In the following, we demonstrate statistical identification of the AT-LDA model with a hierarchical model for topic carry-over, using a simulation study. The simulation is based on a vocabulary of $V = 1,000$ unique terms and four topics ($T = 4$). We generate $D = 1,000$ documents with 40 words each for a corpus of 40,000 words. This set-up is similar to the data sets used in the empirical analysis. We generate the true word-topic probabilities (ϕ_t) and the true document-topic probabilities (θ_d) from symmetric Dirichlet distributions with priors $\alpha = 1/T$ and $\beta = 50/V$ for the document topic probabilities and topic-term probabilities, respectively. For the hierarchical regression, we generate six binary covariates per word (except for the first word in each document) from a binomial distribution with $p_{cov} = 1/3$ for each covariate and also use dummy variables for the topic of the previous word as additional covariates. p_{cov} drives the frequency at which covariates to words appear in the data. Topic dummies in the regression account for a non-zero probability of a topic carry-over when (all) covariates are absent. All true coefficients for the hierarchical regression of ζ on covariates are generated randomly from a uniform distribution with boundaries $[-2, +2]$. Across words and topics, this leads to an average probability of a topic carry-over of about 25%. For estimation, we generate start values for topic assignments randomly and compute initial values for all parameters dependent on topic assignments based on that (ϕ, θ). The latent carry over indicators ζ_w are all started at 0. After running the MCMC, we switch the topic labels (if necessary) once post-hoc to compare true vs. estimated parameter values. In Table (12), we report the posterior estimates of the hierarchical regression and the hit rates for z and ζ after the label switch. Figure (7) shows the recovery of ϕ and θ from our MCMC after subjecting these parameters to the same label switch.

Table 12: Results from simulation study

Parameter	True Value	Posterior Mean	Posterior SD
Recovery of hierarchical regression coefficients			
δ_0	-2.024	-1.998	0.005
$\delta_{T=2}$	1.692	1.549	0.062
$\delta_{T=3}$	2.869	2.800	0.098
$\delta_{T=4}$	1.330	1.382	0.068
δ_1	-1.786	-1.818	0.122
δ_2	-1.072	-0.844	0.075
δ_3	-0.575	-0.657	0.055
δ_4	-2.966	-3.043	0.076
δ_5	-0.496	-0.432	0.077
δ_6	1.765	1.647	0.061
ζ (carry-over indicator: hit rate)	1	0.817	0.002
z (topic assigments: hit rate)	1	0.947	0.001

Legend: True and estimated regression coefficients are obtained via $delta^{(\zeta=1)} - delta^{(\zeta=0)}$. That is, IID topic assigment ($\zeta = 0$) is the contrast state.

Figure 7: Recovery of ϕ_t and θ_d across all topics and documents. Shown are posterior means from the MCMC against true values.

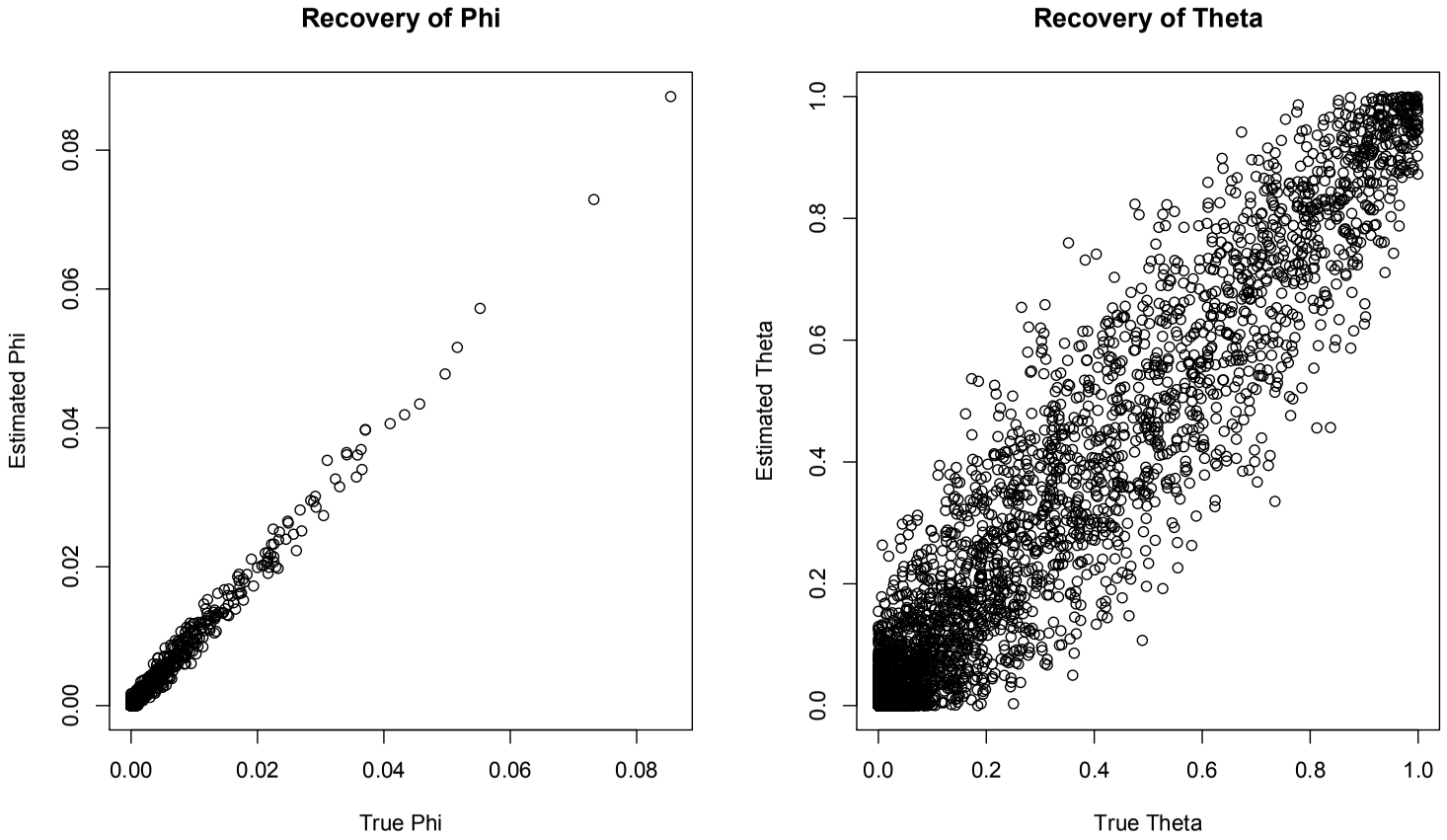
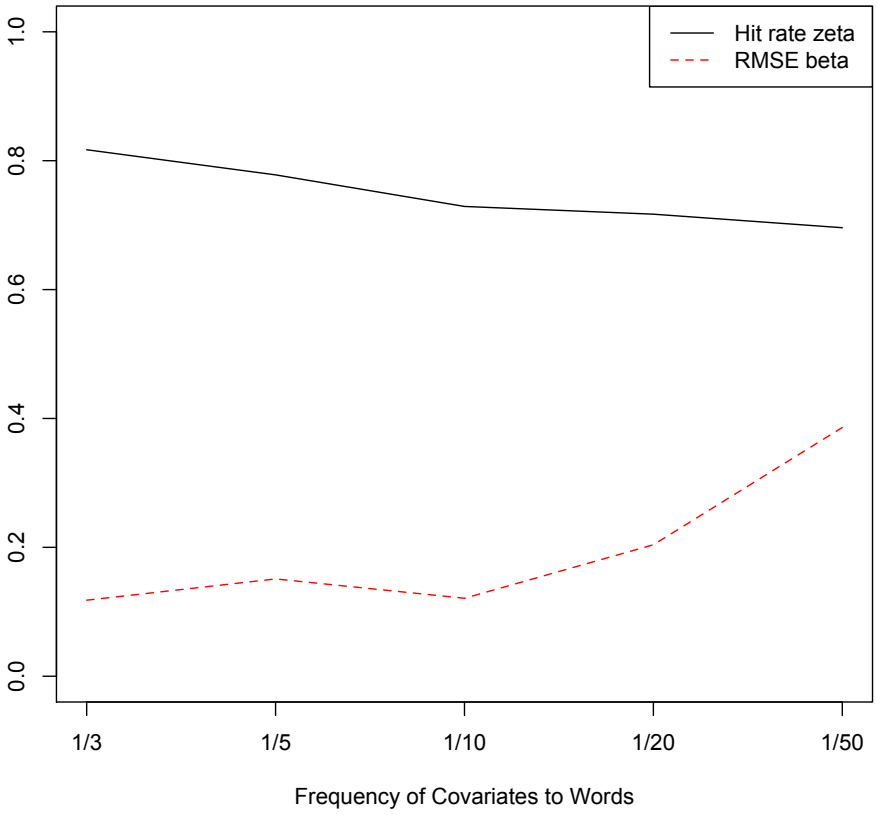


Table (12) shows that the MCMC recovers the parameters of the hierarchical regression of ζ on covariates at very high accuracy. The true values of all regression coefficients are well within the 95% credibility range of the posterior means of the estimated parameters as indicated by their posterior standard deviation. We also report the hit rate of the latent topic assignments and the latent topic carry-over indicator. The hit rate of the z is 95%, the hit rate of the ζ is 82%. While both hit rates are satisfactory, we note that the hit rate of ζ can be increased in simulation by introducing additional covariates or increasing the effect of existing covariates on the probability of a topic carry-over. Figure (7) reveals that the word-topic probabilities are recovered well. This is because, for the estimation of ϕ_t, information is pooled across all the documents. Thus, recovery of ϕ_t is not affected by the relatively small number of words per document. We note that there is no such pooling for estimation of θ_d. This is the reason for larger deviation of estimated and true document topic probabilities. This deviation can, of course, be trivially reduced by allowing for longer documents.

An issue relevant to recovery of the AT-LDA model is the frequency at which covariates occur in the data. As conjunctions and punctuation appear more rarely in the data (e.g, Twitter data due to size restrictions relatively rarely contain complete sentences and, hence, punctuation or conjunctions), recovery of the hierarchical regression to topic carryover becomes more difficult. To explore this issue, we changed the frequency at which covariates appear in the data, holding all other parameters of the simulation constant. Fig. (8) shows the hit rate for ζ and the RMSE of the regression coefficients pertaining to the covariates, given rates of occurence ranging from 33% to 2%. Fig. (8) reveals that the hierarchical structure of the model can be well identified even when covariates appear rarely in the data.

Figure 8: Sensitivity of recovery of hierarchical regression in AT-LDA with covariates with respect to rate of occurence of covariates.



Legend: Graph shows hit rate of ζ and the root mean squared error of regression coefficients pertaining to (simulated) structural covariates. X-axis denotes frequency at which covariates to words appear. For example, 1/3 indicates that 33% of the words are associated with a covariate, 1/10 indicates that 10% of the words are associated with a covariate etc.

A.2 MCMC

For MCMC estimation of the AT-LDA, we rely on the procedure outlined in appendix (A.5) to Büschken and Allenby (2016) which we apply to the word level of our data (see also the Web Appendix to this paper in the following Section A.3). Because we use covariates to the probability of a topic change $p(\zeta|\psi)$ and instead , we depart from their approach for the estimation of the hierarchical regression in our model. The update of δ is accomplished as follows:

$$p(\delta|else) \propto \prod_{d=1}^{D}\prod_{n=2}^{N_d} p(\zeta_{d,n}|z_{d,n-1}=t, x_n, \delta) \times p(\delta) = \prod_{d=1}^{D}\prod_{n=2}^{N_d} \frac{\exp[\delta_{0,z_{n-1}} + \tilde{x}_n'\delta]}{1+\exp[\delta_{0,z_{n-1}} + \tilde{x}_n'\delta]} \times p(\delta) \tag{A.1}$$

For $p(\delta)$, we assume a standard weakly informative multivariate normal distribution. Because of the non-conjugacy of (A.1), we draw δ by way of a RW-Metropolis step. Note that the first word in each of the documents does not inform δ because $\zeta_1 = 0$ by definition.

B Web appendix

Applied to the word level, the MCMC sampling procedure in Büschken and Allenby (2016) for the AT-LDA model is as follows. The factorization of the joint posterior distribution of the knowns and unknowns in Eq. (4) suggests the following sampling steps:

1. $p\left(\phi_t|w,z,\beta\right) \propto \prod_{d=1}^{d} \prod_{n=1}^{N_d} p\left(w_n|\phi_t, z_n = t\right) \times p\left(\phi_t|\beta\right)\ \forall t$

2. $p\left(\delta|else\right) \propto \prod_{d=1}^{d} \prod_{n=1}^{N_d} p\left(\zeta_n|x_n, \delta, z_{n-1}\right) \times p\left(\delta\right)$

3. On the document level (omitting subscript d for z and w to improve readability):

   (a) $p\left(z,\zeta|else\right) \propto p\left(w_1|\phi, z_1\right) \times p\left(z_1|\theta_d\right) \times$
   $\prod_{n=2}^{N_d} p\left(w_n|\phi, z_n, z_{n-1}, \zeta_n\right) \times p\left(z_n|z_{n-1}, \theta_d, \zeta_n\right) \times p\left(\zeta_n|z_{n-1}, \psi\right)$

   (b) $p\left(\theta_d|else\right) \propto p\left(z_1|\theta_d\right) \times \prod_{n=2}^{N_d} p\left(z_n|z_{n-1}, \theta_d, \zeta_n\right) \times p\left(\theta_d|\alpha\right)$

Note that, conditional on ζ, the hierarchical regression of topic carry-over indicators on observed covariates is a standard binary logit model. Because of the first order carry-over effect of the topics, we write down the joint probability of all quantities with respect to two subsequent words given everything else:

$$\begin{aligned} p\left(w_n, w_{n+1}, z_n, z_{n+1}, \zeta_n, \zeta_{n+1}|\phi, \psi, \theta_d, \alpha, \beta\right) \quad &= \\ p\left(w_n, w_{n+1}|\phi, z_n, z_{n-1}, \zeta_n, \zeta_{n+1}\right) \times & \\ p\left(z_n|z_{n-1}, \theta_d, \zeta_n\right) \times p\left(z_{n+1}|z_n, \theta_d, \zeta_{n+1}\right) \times & \\ p\left(\zeta_n|z_{n-1}, \psi\right) \times p\left(\zeta_{n+1}|z_n, \psi\right) & \end{aligned}$$

Note that in the above expression $p\left(z_{n+1}|z_n, \theta_d, \zeta_{n+1}\right)$ is a constant with respect to z_n. This is because, as shown above, it is either 1 (if $\zeta_{n+1} = 1$) or independent of z_n (if $\zeta_{n+1} = 0$). The joint distribution of w_n, w_{n+1} factorizes as follows:

- $p\left(w_n, w_{n+1}|\phi, z_n, z_{n-1}, \zeta_n = 0, \zeta_{n+1} = 0\right) = p\left(w_n|\phi, z_n\right) \times p\left(w_{n+1}|\phi, z_{n+1}\right)$

- $p\left(w_n, w_{n+1}|\phi, z_n, z_{n-1}, \zeta_n = 0, \zeta_{n+1} = 1\right) = p\left(w_n|\phi, z_n\right) \times p\left(w_{n+1}|\phi, z_n\right)$
- $p\left(w_n, w_{n+1}|\phi, z_n, z_{n-1}, \zeta_n = 1, \zeta_{n+1} = 0\right) = p\left(w_n|\phi, z_{n-1}\right) \times p\left(w_{n+1}|\phi, z_{n+1}\right)$
- $p\left(w_n, w_{n+1}|\phi, z_n, z_{n-1}, \zeta_n = 1, \zeta_{n+1} = 1\right) = p\left(w_n|\phi, z_{n-1}\right) \times p\left(w_{n+1}|\phi, z_{n-1}\right)$

where the first expression is the LDA model (no topic carry-over) and the last expression presents the case of a repeated topic carry over. In each case, the probabilities of the words factorize, given z, ζ, ϕ.

B.1 Draw of z_n and ζ_n

Analogous to Gibbs sampling for the HMM (Frühwirth-Schnatter 2006), we consider a joint "single-move" Gibbs sampler of the topic and the stickiness indicator. The joint posterior of z_n, ζ_n is obtained by dropping all elements independent of z_n and ζ_n from Eq. (4) and treating the latent variables z_{n-1}, ζ_{n-1}, z_{n+1} and ζ_{n+1} as observed:

$$\begin{aligned} &p\left(z_n = t, \zeta_n|else\right) \propto \\ &\qquad p\left(w_n, w_{n+1}|\phi, z_n = t, z_{n-1}, \zeta_n, \zeta_{n+1}\right) \times \\ &p\left(z_n = t|z_{n-1}, \theta_d, \zeta_n\right) \times p\left(\zeta_n|z_{n-1}, \psi\right) \times p\left(\zeta_{n+1}|z_n = t, \psi\right) \end{aligned} \tag{B.1}$$

Using results from the above:

$$\begin{aligned} &p\left(z_n = t, \zeta_n = 1|\zeta_{n+1} = 0, else\right) \propto \\ &\quad p\left(w_n|\phi, z_{n-1}\right) \times p\left(w_{n+1}|\phi, z_{n+1}\right) \times p\left(z_n = t|z_{n-1}, \theta_d, \zeta_n = 1\right) \times \\ &p\left(\zeta_n = 1|z_{n-1}, \psi\right) \times p\left(\zeta_{n+1}|z_{n-1}, \psi\right) \propto \phi_{z_{n-1}}^{(w_n)} \times \psi_{z_{n-1}} \times \left(1 - \psi_{z_{n-1}}\right) \end{aligned}$$

$$p\left(z_n = t, \zeta_n = 1 | \zeta_{n+1} = 1, else)\right) \propto$$
$$p\left(w_n | \phi, z_{n-1}\right) \times p\left(w_{n+1} | \phi, z_{n-1}\right) \times p\left(z_n = t | z_{n-1}, \theta_d, \zeta_n = 1\right) \times$$
$$p\left(\zeta_n = 1 | z_{n-1}, \psi\right) \times p\left(\zeta_{n+1} | z_{n-1}, \psi\right) \propto \phi_{z_{n-1}}^{(w_n)} \times \phi_{z_{n-1}}^{(w_{n+1})} \times \psi_{z_{n-1}} \times \psi_{z_{n-1}}$$

$$p\left(z_n = t, \zeta_n = 0 | \zeta_{n+1} = 0, else\right) \propto$$
$$p\left(w_n | \phi, z_n = t\right) \times p\left(w_{n+1} | \phi, z_{n+1}\right) \times p\left(z_n = t | z_{n-1}, \theta_d, \zeta_n = 0\right) \times$$
$$p\left(\zeta_n = 0 | z_{n-1}, \psi\right) \times p\left(\zeta_{n+1} | z_n = t, \psi\right) \propto \phi_t^{(w_n)} \times \theta_{d,t} \times \left(1 - \psi_{z_{n-1}}\right) \times \left(1 - \psi_t\right)$$

$$p\left(z_n = t, \zeta_n = 0 | \zeta_{n+1} = 1, else\right) \propto$$
$$p\left(w_n | \phi, z_n = t\right) \times p\left(w_{n+1} | \phi, z_n = t\right) \times p\left(z_n = t | z_{n-1}, \theta_d, \zeta_n = 0\right) \times$$
$$p\left(\zeta_n = 0 | z_{n-1}, \psi\right) \times p\left(\zeta_{n+1} | z_n = t, \psi\right) \propto \phi_t^{(w_n)} \times \phi_t^{(w_{n+1})} \times \theta_{d,t} \times \left(1 - \psi_{z_{n-1}}\right) \times \psi_t$$

In the case of $n = N_d$, $p\left(w_{n+1} | \cdot\right)$ and $p\left(\zeta_{n+1} | \cdot\right)$ can be dropped because these distributions do not exist. Note that, in the case of a topic carry-over from word n to $n+1$ and $\zeta_n = 0$, the downstream likelihood of z_n consists of two words. If, however, $\zeta_n = 1$ the posterior does not depend on z_n because the topic is already determined. Essentially, the above expressions deal with the question whether to choose the "observed" previous topic assignment z_{n-1} for the current word w_n or to consider the case that z_n originates from θ_d. The above expressions give rise to $T+1$ multinomial probabilities from which we can jointly draw z_n, ζ_n.

B.2 Draw of θ_d

In MCMC sampling for the standard LDA, the full conditional draw of θ_d is based on using the multinomial topic assignment of all words (or sentences) in a document as likelihood information. The multinomial likelihood of the topic assignments is combined with the

Dirichlet prior $p(\theta|\alpha)$ for a conjugate update via a Dirichlet posterior in which the topic assigments are simple counts (see Eq.(**??**)). For the sticky LDA, we have to keep track of the topic assignments which are downstream of θ_d and disregard topic assignments due to $\zeta = 1$:

$$p\left(\theta_d|else\right) \propto p\left(z_1|\theta_d\right) \times \prod_{n=2}^{N_d} p\left(z_n|z_{n-1}, \theta_d, \zeta_n\right) \times p\left(\theta_d|\alpha\right) \quad =$$

$$\prod_{n:\zeta_n=0} p\left(z_n|\theta_d\right) \times p\left(\theta_d|\alpha\right) \times \prod_{n:\zeta_n=1} 1 \quad =$$

$$\prod_{n:\zeta_n=0} p\left(z_n|\theta_d\right) \times p\left(\theta_d|\alpha\right)$$

We use the count matrix C^{TD} to collect topic assignments conditional on $\zeta_n = 0$ and then proceed as in the standard LDA.

B.3 Draw of ϕ

The draw of ϕ_t is not affected because we can treat the z as observed. This update can be conducted in the usual way.