7-11-2024 9:30 AM

mixSTM: Adapting the Structural Topic Model for a quantitative analysis of focus group data

Pascale A. Nevins, *University of Western Ontario*

Supervisor: Lizotte, Daniel J., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biostatistics

Abstract

The Structural Topic Model (STM) incorporates external information about expected document-topic proportions to enhance the model. Motivated by focus groups, whose transcripts represent text data inherently grouped by session, we propose three extensions to the STM: 1) mean document-topic proportion estimation using a regression with random effects; 2) partitioned estimation of group-specific topic covariance matrices; and 3) a post hoc mixed effects regression on topic prevalence which incorporates latent variable uncertainty into the coefficient estimates. We explore the utility of these modifications through simulated examples and apply them to focus group transcripts from a pan-Canadian study on homelessness. The new methods, collectively the "mixSTM", improved topic model fit when there was complex group-related variation in topic prevalence and provided new avenues for interpretation. These methods may better represent analyst beliefs about qualities of grouped text data, although there is a risk of over-complicating the estimation given small, qualitative data sources.

Keywords

Topic Model, Structural Topic Model, Variational Inference, Focus Groups, Clustered Data.

Summary for Lay Audience

The output of health or social research can be text data, such as when conducting focus groups. Topic modelling is a quantitative method capable of extracting information rapidly from large collections of text. To do so, topic models propose a theoretical generation mechanism for text which assumes the existence of "topics": groups of words that tend to co-occur and thus share elements of meaning. Each document (text segment) in a collection is assumed to be composed of multiple topics in different amounts; the goal of topic modelling is to find both the topics and the amounts. The Structural Topic Model (STM) of Roberts et al. (2014) lets external information about each document (e.g., the author/speaker) influence how much the topics are expected to be used in each document. This thesis proposes modifications to the STM when documents are grouped. One example of grouped documents could be focus group transcripts: a transcript is composed of text segments that are more related to each other than to segments of other transcripts, so it forms an inherent grouping. We provide two extensions to the generative model of the STM and one for exploring the results of the topic model. We showed with simulated data that our changes to the generative model could obtain closer estimates of topic proportions to the truth than the STM, particularly if there were many external quantities whose relationship with topics varied between groups. We then applied our methods to focus group transcripts from a pan-Canadian study on homelessness, and showed that our results align with previous analyses of the data and revealed some additional word associations. Thus, these methods have the same advantages as the original STM, with potential benefits related to the incorporation of the grouping structure into estimation and the ability to interpret the output in light of the groups. These extensions to the STM are applicable to many grouped-document settings in that they may better represent the beliefs that analysts have about how topics are distributed across documents.

Acknowledgements

I would like to acknowledge the many people without whom this thesis would not exist. First and foremost, a heartfelt thank you to my supervisor, Dr. Dan Lizotte, for your guidance and support throughout all stages of the project. Thank you for asking the right questions and encouraging me: I always left a meeting with you feeling more capable than when I walked in. I'm also immensely grateful to the whole Homelessness Counts team, especially Sara Husni and Dr. Cheryl Forchuk. Working with data collected as a part of the project has been incredibly motivating. Thank you to my thesis advisory committee, my professors, and my biostatistics cohort (Chenyang and Jiaqi), who have all had a part in my career at Western. It's been a pleasure to be part of this community. A final thank you to Dr. Monica Taljaard, who opened my eyes to biostatistics and whose mentorship has been instrumental to my development as a researcher.

I would also like to acknowledge the financial support I received to continue my studies in biostatistics at Western: from a Canada Graduate Scholarship in my first year, Dr. Dan Lizotte, and the Western Graduate Research Scholarship.

Land Acknowledgement

I have lived and worked on this thesis in a city which resides in the traditional territories of the Anishinaabek, Haudenosaunee, Lūnaapéewak, and Chonnonton Nations. These lands are connected with the London Township and Sombra Treaties of 1796 and the Dish with One Spoon Covenant Wampum. I respect the longstanding relationships that Indigenous Nations have to this land and acknowledge historical and ongoing injustices that Indigenous Peoples (First Nations, Métis and Inuit) endure in Canada.

The work contained herein has been developed and written within a euro-centric conceptualization of knowledge and is supported by settler institutions. I acknowledge that there are many Ways of Knowing which are not captured within this lens. In particular, this thesis focuses on text as a source of information– but not all knowledge is or need be recorded in written form. Many voices and sources of knowledge, especially those of Indigenous Peoples, have been systematically undervalued throughout history and continue to be overlooked to this day.

Table of Contents

Abstract ii

Summary for Lay Audience iii

Acknowledgements iv

Land Acknowledgement v

Table of Contents vi

List of Tables xiii

List of Figures xv

List of Appendices xvi

1 Introduction 1

1.1 Topic modelling text . . . 2

1.1.1 Introduction to topic models . . . 2

1.1.2 Extensions to LDA . . . 4

1.2 Contributions . . . 5

2 Pairing topic modelling and focus group analysis 8

2.1 Extracting meaning from topic models . . . 8

2.1.1 Comparing and validating topic models . . . 8

2.1.2 Choosing the number of topics . . . 11

2.2 Topic models and qualitative research, in brief . . . 13

2.2.1 An overview of qualitative research . . . 13

2.2.2 Comparing topic models to qualitative approaches . . . 14

2.3 Modelling the data output of focus groups . . . 15

2.3.1 What are focus groups? . . . 15
2.3.2 The potential of topic models in the analysis of focus groups . . . 17
2.3.2.1 Defining documents in focus groups . . . 20
I. Length and partitioning of transcripts. . . . 21
II. Document partitioning in the context of interpretation. . . . 22
III. For the analyst. . . . 23
2.3.2.2 Facilitator speech . . . 24
I. As part of the documents. . . . 24
II. As a covariate. . . . 25
2.3.2.3 Other considerations for focus group transcript processing 26
2.3.3 Motivating the mixSTM through focus groups . . . 27

3 Variational Expectation-Maximization in the STM 29

3.1 Statistical background . . . 29
3.1.1 Principles of Bayesian modelling . . . 29
3.1.1.1 Priors . . . 30
3.1.1.2 Posterior quantities and inference . . . 30
3.1.2 Variational inference . . . 32
3.1.2.1 Limitations of variational inference . . . 35
3.2 The Correlated Topic Model . . . 35
3.2.1 Introduction to the CTM . . . 36
3.2.2 Variational Expectation-Maximization for the CTM . . . 37
3.3 The Structural Topic Model . . . 38
3.3.1 Introduction to the STM . . . 39
3.3.2 Obtaining the STM's objective function . . . 41
3.3.2.1 Finding the $q^*(\cdot)$ distributions . . . 41
3.3.3 E-step: Maximizing the variational distributions . . . 44
3.3.4 M-step: Update model parameters (μ and Σ) . . . 45

3.3.4.1 Incorporating covariates in the updates to μ 45
3.3.4.2 Maximum likelihood estimation for Σ 46
3.3.5 Bringing it all together: iteration 47
3.3.6 Benefits of the STM 48

4 Incorporating grouping structure into the estimation of the mean topic weights 49
4.1 The STM's incorporation of covariates 49
4.1.1 Introduction to regression updates to μ 49
4.1.2 The STM's variational linear regression 51
4.1.2.1 Regression details 51
4.1.2.2 Regularization 53
4.2 Clustered document models for μ 55
4.2.1 Motivating a change to the estimation of μ_k 55
4.2.2 Potential regression models 57
4.3 Penalized mixed models fit using streamlined variational Bayes 59
4.3.1 Global-local priors 59
4.3.2 Mean-field variational updates for γ and u 60
4.3.3 Comparison to the STM's variational regression 62
4.3.4 The mixSTM's generative model with a novel mean update 64
4.4 Simulation studies for the mixSTM's mean update 65
4.4.1 Simulation methods 65
4.4.1.1 Simulation comparisons 67
4.4.2 Simulation results 69
4.4.2.1 Retrieval of true topic proportions θ 69
4.4.2.2 Held-out likelihood 74
4.4.2.3 Topic quality 75
4.4.3 Discussion 77

5 Incorporating groups into the estimation of topic covariances 79

5.1 The STM's global topic covariance matrix . . . 79

5.1.1 Sampling from the variational distribution of θ . . . 79

5.2 Two-level nested groupings provide a new opportunity for Σ . . . 80

5.2.1 Justification for partitioning Σ . . . 81

5.2.2 Σ_g in the variational EM algorithm . . . 82

5.2.3 Generative model for the mixSTM . . . 84

5.3 Simulation studies for the mixSTM's covariance update . . . 84

5.3.1 Simulation methods . . . 85

5.3.1.1 Simulation comparisons . . . 86

5.3.2 Simulation results . . . 87

5.3.2.1 Σ_g in the mixSTM . . . 87

5.3.2.2 Σ_g in the CTM . . . 88

5.3.2.3 Held-out likelihood . . . 89

5.3.2.4 Variance estimates using ν re-estimation . . . 90

5.3.3 Discussion . . . 91

6 Estimating the effect of covariates on topic proportions in a setting with grouped documents 95

6.1 The STM's strategy for a regression on topic prevalence . . . 95

6.1.1 The Method of composition . . . 95

6.1.2 Accessing θ, our dependent variable . . . 97

6.1.3 Inference on θ . . . 98

6.1.3.1 Choosing a regression family and link function . . . 99

6.1.3.2 Choosing priors for the coefficients . . . 100

6.2 Implementation in a setting with grouped documents . . . 100

6.2.1 Motivating the addition of random effects . . . 101

6.2.2 The Method of composition for mixed effect models . . . 103

6.2.2.1 Approaches to a joint posterior and their assumptions . . . 104
6.2.3 Implementation . . . 105
6.2.4 Example . . . 106

7 Case study applying the mixSTM to focus group transcripts from the Homelessness Counts study 110

7.1 Background . . . 110
7.1.1 Research questions . . . 111
7.2 Quantitative methods . . . 112
7.2.1 Preprocessing for focus group transcripts . . . 112
7.2.1.1 Metadata of interest . . . 112
Facilitator questions . . . 113
7.2.1.2 Document segmentation . . . 114
7.2.1.3 Text cleaning . . . 116
7.2.2 Applying the mixSTM to focus group data . . . 118
7.2.2.1 Choosing K . . . 118
7.2.2.2 Running the mixSTM . . . 119
7.2.2.3 Estimating the covariate effects . . . 120
7.3 Results . . . 121
7.3.1 Describing the corpus . . . 121
7.3.2 Selecting K and running the model . . . 122
7.3.3 Topic interpretation . . . 124
7.3.3.1 High-quality topics . . . 124
Topic 2: Youth homelessness and intersections with school systems. . . . 125
Topic 4: Sharing information among service providers. . . . 125
Topic 5: Community locations where homelessness is observed. . . . 126

Topic 6: Comfort and safety of people with lived experience of homelessness. 127
Topic 12: Using databases to track and report on homelessness. 128
Topic 13: By-name lists and coordination around these lists. 129
Topic 14: Addictions, health, and mental health issues contribute to houselessness. 129
Topic 16: Vulnerable populations' intersections with the shelter system. 130
7.3.3.2 Lower quality topics 131
Topic 1: Collecting data and current systems. 131
Topic 3: Evictions and tenant-landlord contacts; data issues when it comes to Indigenous peoples. 131
Topic 7: Case management for people experiencing homelessness. 132
Topic 8: Making new contacts with people and organizations. 132
Topic 9: Observations about service utilization 133
Topic 10: Connections with invisible homelessness and access to services 133
Topic 11: Point-in-time counts and questions/answers from PEH. 133
Topic 15: Renting and system timings. 134
Topic 17: The "soaker" 134
7.3.4 Estimate Effect 135
Table 7.1: Covariate associations with subpopulations and causes of homelessness. 135

Table 7.2: Covariate associations with question-attributable topics. . . . 137
Table 7.3: Covariate associations with lower quality topics. . . 138
7.4 Discussion . . . 140
7.4.1 Comparison to previous work . . . 140
7.4.1.1 Subpopulations and causes of homelessness . . . 140
7.4.1.2 Data collection and use . . . 143
7.4.1.3 Locations . . . 145
7.4.2 Strengths and limitations of our topic model analysis . . . 146
7.4.3 Implications . . . 148
7.4.4 Conclusions . . . 150

8 Conclusion **151**
8.1 Discussion of contributions and future work . . . 151
8.1.1 Random effects in the mean topic prevalence model . . . 151
8.1.2 Group-specific topic covariance matrices . . . 153
8.1.3 Estimating the effect of covariates on topic prevalence . . . 154
8.1.4 Future work on the STM and mixSTM for grouped documents . . . 156
8.2 The mixSTM in a broader context . . . 160
8.2.1 Alternative models of grouped documents . . . 160
8.2.2 Applying the mixSTM to other corpora in health research . . . 162
8.2.3 Focus group analyses through a quantitative lens . . . 164
8.3 Conclusion . . . 165

Bibliography **167**

Appendices **185**

List of Tables

4.1 Comparison of mixSTM and STM using the difference of L1 distances and KL divergences to true topic proportions. 70

4.2 Mean (SD) per-document held-out likelihood for the mixSTM (mSTM), STM, and CTM. 74

4.3 Per-topic and overall sum of semantic coherence scores and frequency-weighted exclusivity scores for two scenarios. 76

5.1 Comparison of the mixSTM estimated with separate covariances per group (mSTM_G) to the mixSTM estimated with only one global covariance (mSTM_{NG}) 87

5.2 Comparison of a CTM estimated with separate covariances per group (CTM_G) to a CTM estimated with only one global covariance (CTM_{NG}) 89

5.3 Mean (SD) per-document held-out likelihood for a mixSTM with and without a groupwise covariance matrix. 90

5.4 Comparison of distribution of posterior variational variance estimates ν_d estimated under local, global-grouped, and global methods for simulation scenario 2. 90

5.5 Median (Q1, Q3) variance of 100 samples drawn from the estimated variational posterior of $\theta_{d,k}$, using the local, global-grouped, and global approximations to ν_d in the mixSTM. 91

6.1 Comparison of coefficients estimated using a mixed effect linear regression with "Global" and "None" uncertainty incorporation. 107

6.2 Comparison of coefficients estimated using original fixed effects regression with "Global" and "None" uncertainty incorporation. 107

7.1 Regression results on talk turn-topic proportions for Topic 2 (Youth/family), Topic 3 (Evictions and other issues), and Topic 6 (Safety and comfort of PEH). 136

7.2 Regression results on talk turn-topic proportions for Topic 4 (Sharing information), Topic 12 (Using databases), and Topic 13 (By-name lists and coordinated access). 137

7.3 Regression results on talk turn-topic proportions for Topic 1 (Collecting data and current systems), Topic 7 (Case management for PEH), and Topic 8 (Making contact with PEH and organizations). 139

A.1 Summary of topic model output B2.70: A topic model with 5 topics, 1000 documents and a 599 word dictionary. 191

List of Figures

3.1 Plate diagram representing the Structural Topic Model (STM) 39

4.1 Sorted L1 distances between estimated and true topic proportions for 100 simulations of scenarios A1, A2, B1, and B2. 71

4.2 (A): Sorted L1 distances between MAP estimates and true document-topic proportions versus the difference in L1 distances (mixSTM-STM) for 100 simulations of scenario B2. (B): Histogram of L1 distances between MAP estimates and true document-topic proportions for the first 20 simulations of scenario B2. 73

4.3 (A): Histogram of difference in L1 distance (mixSTM-STM) between MAP estimates and true document-topic proportions for scenario B2. (B): L1 distance between the MAP estimates and the true document-topic proportions for the mixSTM plotted against the same for the STM, with axes on the $\log_{10}$ scale for scenario B2. 74

5.1 Distributions of variational posterior variances ν under "global" (NG), "global-grouped" (G), and "local" (L) estimation methods for scenario 2 in the mixSTM. 90

7.1 Exclusivity, Semantic Coherence, Partial Held-Out Likelihood, and Residual Dispersion for 6-30 topics. 122

A.1 Scenario A1, L1 distance. 187

A.2 Scenario A2, L1 distance. 188

A.3 Scenario B1, L1 distance. 188

A.4 Scenario B2, L1 distance. 189

A.5 Scenario B2, KL Divergence. 190

List of Appendices

Appendix A Content Appendices 185

Appendix B Ethics Approval 196

Chapter 1

1 Introduction

Many forms of text data are generated over the course of health and social research. Text can be the intended data source of prospective research (e.g., surveys with open-ended questions, interviews, or focus groups) or obtained for retrospective studies (e.g., meta-research on publications or analyses of social media posts). Traditionally in the health and social sciences, data that take the form of text have been analyzed qualitatively, taking advantage of qualitative research's strengths of close reading, ability to interpret tone and feelings, narrowed scope which extracts insights at the level of the discourse or person rather than the population, and reflection on how the researcher's position with respect to the question, data, and research subjects shape the results. The insights obtained from text data play a fundamental role in a well-rounded understanding of health on the individual and population levels.

With the increasing accessibility of machine learning, text data have become more and more analyzable in quantitative ways. Natural language processing techniques are a class of methods which can rapidly process large amounts of text to extract elements of interest. These computational approaches, such as the widely-used topic modelling, are still vulnerable to many of the same biases that plague qualitative research. Although attractive to analysts for the promise of reproducible results, the robustness of these methods relies on deliberate design choices, principled analyses, and validity checks. The output of a quantitative approach to text data can be a combination of numerical or statistical and textual information, requiring an interpretation under both quantitative and qualitative lenses.

In what follows, we will develop and illustrate a grouped structural topic model for the analysis of text data, using the motivating case of qualitative focus groups.

1.1 Topic modelling text

1.1.1 Introduction to topic models

Topic models are un- or semi-supervised machine learning methods applicable to collections of sparse, discrete data. Applied to text, they are motivated by the desire to describe a collection of documents using a simpler but still meaningful representation. As the name suggests, this lower dimensional representation takes the form of "topics", which are latent (unobserved) distributions over discrete data, often words. A topic model is presented in terms of a data generation mechanism for the text, which makes distributional assumptions about the topic assignments of words. To estimate the topics, the probability of a word in a document belonging to a given topic is iteratively maximized across all words. The topics are then represented in terms of their highest probability words, which can be used to attribute meaning to the topic.

To elaborate on specifics of topic modelling, we must define some notation:

- A token or word (w): The unit of discrete data. In many cases, this is the basic semantic component of language, a unigram or single word, but bigrams (two words) or any other discrete data source can also form the tokens. Tokens are represented numerically as a vector of zeroes with a singular one-entry which indicates the presence of the word.
- A document (d): A collection of N_d words. Often the "document" designation arises naturally from the way the text is generated or collected (such as in online reviews or social media posts). Topic models are applied to collections of documents.
- The vocabulary: All V unique words across all documents, where each word is used in total c_v times.
- A corpus (plural: corpora): All D documents; the whole text collection of interest.

There are $\mathcal{N} = \sum_D N_d = \sum_V c_v$ words in the whole corpus.

- The topics (k): There are assumed to be K latent topics in the corpus. The goal of topic modelling is to uncover the distribution of words assigned to each topic and how these topics are distributed across documents.

One of the most widely used topic models is Latent Dirichlet Allocation (LDA), pioneered for text data by Blei, Ng, and Jordan in 2003 [17]. In LDA, the assumed generative process for a corpus is the following:

$$\begin{aligned} \theta_d &\sim \text{Dirichlet}(\alpha) \qquad \text{for each } d = 1, ..., D \\ z_{d,n} &\sim \text{Multinomial}(1, \theta_d) \qquad \text{for each } n = 1, ..., N_d \\ w_{d,n} &\sim \text{Multinomial}(1, \beta_{z_{d,n}}) \qquad \text{for each } n = 1, ..., N_d \end{aligned} \tag{1.1}$$

That is, for each document d, θ_d (a vector of document-topic proportions) is sampled from a Dirichlet distribution with some parameter α. For each word index, $z_{d,n}$, in the document, a topic is sampled according to a multinomial distribution with a probability defined by θ_d. Then, the word label, $w_{d,n}$, is sampled from a multinomial distribution taking into account the probability of a word being used and the sampled topic (the topic-word proportions, β) [17].

This process does not represent a true mechanism that humans follow when speaking or writing, but rather a model that permits us to infer the hidden representation (topics) knowing the number and frequency of words and the number of topics. It must be noted that while the first two are inherently known when one is in possession of text, the "true" number of topics is unknowable except in simulation. For real text data, quantitative and subject-specific knowledge-based approaches must be used to choose the number of topics (a selection of which will be elaborated on in section 2.1.2).

Importantly, a topic has no meaning unless attributed one by a person. For some choices

of α and β, topics often emerge as apparently coherent sets of words because in most text, words that tend to co-occur are related in meaning. The utility of topics as qualitative descriptions of text relies on this assumption about word co-occurrence.

Another key assumption made by LDA is that the relative distance of words from one another within the document is unneeded in order to obtain a meaningful extraction of topics, often called the "bag-of-words assumption"; this exchangeability (irrelevant order) assumption is also assumed about documents in LDA [17, 76]. Finally, LDA assumes that all documents can be represented as mixtures of different proportions of topics (unlike single-membership models, which assume each document discusses only one topic) and that the topics are independent [17, 30].

1.1.2 Extensions to LDA

Other topic models have been developed in the years since LDA's inception which change the modelling assumptions and data generation mechanism but keep the underlying idea of topics as latent quantities.

A prevalent example in the literature is the Correlated Topic Model (CTM, [14, 16]), which allows for correlation between topics. The CTM models the topic proportions in a document as coming from independent logistic normal distributions with a globally shared covariance matrix across all documents, rather than the independent Dirichlet distributions in (1.1). Modelling a covariance between topics allows the model to capture additional structure in the form of which topic pairs tend to appear together in documents. In many corpora, relaxing LDA's assumption of independent topics may be considered a more realistic generative model of text [16].

Of interest to the present thesis is a topic model proposed by Roberts et al. in 2013: the Structural Topic Model (STM) [96]. The STM uses the framework of the CTM and takes inspiration from topic models which utilize external information about documents (such as

the Dynamic Topic Model [15] and the Expressed Agenda Model [51]) with the goal of incorporating document metadata or "covariates" into the topic estimation and interpretation. Covariates are used in the STM at three points: one, they are allowed to provide additional information for the estimation of document-topic proportions, θ_d (called the topic prevalence model); two, they can provide information for the topic-word probabilities β_k (the topic content model); and finally, they can be used as covariates on a post hoc regression on topic prevalence or content [97, 99]. The STM posits that adding covariates to the estimation of topic content or prevalence strengthens the inference through partial pooling: by providing relevant prior information to the model, documents of the same level of a given variable can share information about expected prevalence and content [97]. The ability to estimate associations of the covariates with the topic prevalence or content also provides rich information and more relevance for these latent concepts, although any inference should be treated as exploratory.

The STM has seen applications in many contexts for the analysis of text data, including hospitality [58], transportation [77], history and economics [50, 72], and politics and sociology [100, 117], where it has been applied to user reviews [58, 46], news and scientific articles [130, 117], and social media and forum posts [88], among plenty of others.

1.2 Contributions

The objective of this thesis is to develop a method of incorporating document grouping structure into the topic prevalence model of the Structural Topic Model and apply it to the motivating case of focus group transcripts. Borrowing from the mixed model literature, we see a "cluster" (henceforth, "group") of documents as a collection of documents which are nested within some inherent hierarchical structure. Although the STM is a powerful tool that can accommodate groups using dummy-coded variables in its estimation and post hoc inference, explicitly incorporating grouped structure provides new avenues for interpreta-

tion and fit.

The proposed modifications to the STM occur in three parts:

1. Mean document-topic proportions: We propose using a mixed effects regression to estimate the mean document-topic proportions in place of the fixed effects regression in the original STM. This will explicitly allow for correlation between group-level effects, require fewer degrees of freedom for estimation, and permit different penalization on grouped and non-grouped coefficient estimates.

2. Topic covariance: We propose estimating groupwise covariance matrices to model topic correlation, rather than a single shared covariance matrix across all documents. Estimating a covariance matrix shared only between documents within a group provides an opportunity to explore the relationships between topics more granularly and may improve the precision available to posterior inference under some estimation methods.

3. Inference on topic prevalence: Inference on the topic prevalence is currently only implemented for fixed effects regressions, because of the complexities involved accounting for multiple sources of uncertainty. We propose a method to extend the current approach to random or mixed effects models, which will allow for interpretation in terms of the amount of variance in topic prevalence explained by the grouping structure.

Collectively, these modifications are henceforth referred to as the "mixSTM" or Mixed Structural Topic Model.

In Chapter 2, we will elaborate on topic model interpretation, validation, and application to focus groups. In Chapter 3, we will describe details of variational inference for the STM. In Chapters 4 to 6, the three modifications to the STM that compose the mixSTM will be motivated, discussed, and illustrated through simulation. In Chapter 7, we will apply the

mixSTM to focus group data from the Homelessness Counts study. In Chapter 8, we will discuss future avenues for the mixSTM and alternative models.

Chapter 2

2 Pairing topic modelling and focus group analysis

2.1 Extracting meaning from topic models

In the case of text data, topics are collections of words. However, there is no inherent number of or label for the topics: these must be chosen and assigned by the analyst [52].

2.1.1 Comparing and validating topic models

Without validation steps, a topic model's results are difficult to interpret in a justifiable way and can lead to misleading conclusions, for example if only the top few words for each topic are considered during labelling [52, 63, 133]. Instead, topic models should be scrutinized on as many dimensions as possible, which may include checking the validity of statistical assumptions; reading representative documents and comparing the topics' words; and comparing the topics to other models of the text or external event markers which delimit expected changes in the results [36, 63, 133].

In terms of statistical validation, one option is to look at (partial) held-out likelihood. Held-out likelihood evaluates how well a topic model fit on a subset of the data (in documents or words) is able to correctly attribute topic assignments to unseen word indices; a larger held-out likelihood implies the topic model is a better representation of the text [16, 63]. However, when documents are particularly small or few in number, the choice of held-out content can dramatically affect the results [99]. As such, using a cross-validated held-out likelihood (where subsets of the data are held out in turn and the results are averaged across all held-out samples) may be a more representative measure. A similar option for statistical validation is to use the instantaneous mutual information [82]. Once conditioned on the topic, knowing the document index should provide no information about the topic assign-

ment of the words in most topic models; the instantaneous mutual information provides evidence for whether this independence is a realistic assumption [82, 100]. Relying only on statistical measures like the above, however, does not guarantee resulting topics will be useful: Chang et al. [28] showed that maximizing held-out likelihood was negatively correlated with maximizing a measure of the topics' semantic interpretability.

As such, a crucial step for *all* topic models is that of validating the semantic interpretation of the topics by looking at words and documents. Domain knowledge and an understanding of the corpus become essential in this kind of evaluation.

When looking at the high-probability words in a topic, a key criterion is whether the words are exclusive to the topic, i.e., whether words in separate topics tend to be distinct. Repeated words across topics are not strictly a concern– a desirable property of topic models in relation to semantics is "relationality": the fact that words can be interpreted in context rather than in isolation [36], and it is difficult to have completely exclusive topics in any corpus. However, lots of overlap between topics in high probability words, particularly of words without multiple contextually relevant meanings, can be a sign of redundancy: too large a number of topics are being estimated relative to the number of distinct concepts [36]. Exclusivity is often assessed qualitatively (looking at the output of the topics) but can also be quantified, for example using the FREX metric of Bischof and Airoldi [10] which weights exclusivity by the frequency of the words.

Another criterion is that of "semantic coherence": whether high probability words for a topic tend to co-occur in documents and form a related set. Mimno et al. [82] provide one commonly used quantitative method of evaluating semantic coherence, although others exist in the literature [101]. Another way to assess whether the high probability words associated with a topic form a semantically related set involves coding word intrusion [28]. To do this, a word with low probability is added to a list of high probability words associated with a certain topic and people familiar with the corpus are invited to identify the

word which does not belong: the more often the true intruder word is correctly identified, the better quality the topics are assumed to be [28]. Methods to automate the coding of word intrusion have been proposed (e.g., [73]).

One should supplement the interpretation of the topic obtained from high probability words by looking at the documents whose topic proportions favour one topic [99]. Documents can provide additional context and nuance to interpretations. In multi-membership models like the topic models discussed here, documents are assumed to be composed of multiple topics. As such, looking at documents which have an intermediate probability of several topics can provide further evidence for the interpretation of one topic in contrast to the others, particularly when said topics share some high probability words [36]. Similarly to the concept of "word intrusion", one could measure "topic intrusion" by providing people with a set of documents where all but one have a high proportion of the topic at hand and asking them to identify the outlier [28]. An ideal approach to semantic validation is often considered to be hand-coding a set of documents and checking whether the codes match up with the high probability topics for that document [36, 52]. Rigorous checks would require the coder to be blinded to the topic model results, but have in-depth subject knowledge of the corpus' content. This kind of validation by human coding can thus become resource intensive.

Finally, topic model validation can occur by seeking external information which corroborates the topic interpretation. A potential implementation of this could be comparing the estimated topics to themes or codes identified during an independent qualitative analysis of the corpus or a subset of the corpus (e.g., [7, 83, 84]). This can validate whether the information captured by the topic model is meaningful and related to what is or would be identified as important by humans. However, this external validation strategy relies on congruent research questions from the two approaches and, like any qualitative approach, requires significant time to be done thoroughly. Alternately, DiMaggio et al. suggest an ex-

ternal validation practice where surrounding information regarding the documents is used to form and test hypotheses about how the topics are expected to vary [36]. DiMaggio et al. use the example of demarcating events in politics affecting word use in the documents [36]; however, not all corpora can be expected to have external information which obviously relates to changes in the topics. The integration of this final external validation approach with the STM is straightforward, as the STM provides a method to quantify the relationship between covariates and changes to topic prevalence [99].

2.1.2 Choosing the number of topics

One key decision when implementing any topic model, including the STM, is choosing the number of topics. There is no sure way of determining the "correct" number of topics in a corpus since topics are inherently unobserved. However, comparing models fit using a varying number of topics using a combination of numerical and semantic considerations can inform this choice [99]. The `stm` package in R provides a `searchK` function which computes exclusivity, semantic coherence, held-out likelihood, and residual dispersion for any number of topics in a specified range [99].

To quantify exclusivity, `stm` uses a simplified version of the FREX metric of Bischof and Airoldi [10], which is a weighted harmonic mean of a word's frequency and exclusivity ranking among the M most probable words [99]. For a word label v in a vocabulary of size V within topic k, with a weight $\omega \in [0, 1]$, and where $r(\cdot)$ denotes the rank, the exclusivity is [99]:

$$E_{k,v} = \left(\frac{\omega}{r(\beta_{k,v} / \sum^{K} \beta_{k,v})/V} + \frac{1-\omega}{r(\beta_{k,v})/V} \right)^{-1} \tag{2.1}$$

The exclusivity of a given topic is then $E_k = \sum^{M} E_{k,v}$. One can also look at the words with the highest FREX scores within a topic: these will represent the words which are the most unique to the topic but that also tend to occur more frequently in the corpus. This limits the presence of single-use words in the list, since these are exclusive to a topic by definition

but not typically representative of the documents.

To quantify semantic coherence, `stm` uses the metric of Mimno et al. [82]. For two words m_i and m_j among the M most probable words for topic k, $D(m_i, m_j)$ the number of documents in which m_i and m_j co-occur at least once, and $D(m_j)$ the number of documents in which m_j occurs at least once, the semantic coherence is [99, 82]:

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log\left(\frac{D(m_i, m_j) + 1}{D(m_j)}\right) \tag{2.2}$$

A higher semantic coherence quantifies the idea that high probability words in a topic tend to co-occur. However, as noted by Roberts et al., this criteria can be easy to maximize when the high probability words are very common across the corpus, so it should be used in conjunction with other metrics [100, 99].

Residual dispersion, per Taddy [113], is an approximate measure of overdispersion of the residuals in the topic model. Under the assumed model, the residual dispersion should be 1; a value of more than 1 suggests more topics may be required to account for the existing variability [113, 99]. For full details of the calculation, readers are directed to section 4.2 of Taddy (2012) [113].

These quantitative comparisons can help shorten the list of candidate models, so that qualitative comparisons (which look more closely at the words and documents attributed to a topic) can subsequently be used to choose the final number.

The choice of K is often bemoaned as a highly subjective step in an otherwise straightforward quantitative analysis [83]. However, when representing something with as much meaning to humans as text, one could argue that incorporating a subjective step like topic number choice is more coherent with the overall premise. After all, choosing the number of topics defines the content of the topics in terms of their granularity (ability to extract nuance versus broad themes), and a choice one way or the other may better serve the goals

of the inference for a given corpus [64]. Finally, as DiMaggio et al. point out [36], topics should be considered as a whole: a topic model will often assign words to one or more "noisy" topics to reduce the variance in the others.

2.2 Topic models and qualitative research, in brief

Having presented the goals, premise, and subjectivity involved in topic modelling, we now turn our eyes briefly to qualitative research and its relationship with topic models as an approach to text.

2.2.1 An overview of qualitative research

Qualitative research, like quantitative research, starts at the time of data collection. Participants to the research must be sampled, but the goal of sampling in qualitative research is not necessarily achieving a specific sized group of participants representative of the population, but instead about identifying a few participants who will provide rich perspectives, opinions, or actions relevant to the research question [75, 87]. Data collection can proceed through a wide variety of methods, such as conversations (structured or unstructured), field notes and observations, or the collection of visual or other media [75]. Crucially, the data collected in qualitative research is dense, allowing for open-ended inter- and extrapolation and requiring a degree of subjective evaluation– rather than the standardized closed-ended questions and measurement scales of quantitative research. The researcher's positionality [102] with respect to the participants and research setting is key in both the data collection and interpretation in qualitative research: the researcher is an unignorable part of the research process [111]. Qualitative approaches to analysis require immersion into the collected data through close readings, and take time and effort in an iterative process as the researchers refine their ideas and strengthen their conclusions [75, 89].

2.2.2 Comparing topic models to qualitative approaches

Topic models have been compared to more traditional qualitative coding and analysis methods in several contexts. For example, Baumer et al. compared a topic model to grounded theory, in terms of assumptions and output, and found that the topic model's topics shared commonalities with the themes obtained from the grounded theory approach but tended to be less abstract [7]. Similarly, Miner et al. compared topics generated using LDA to a qualitative approach, and agreed that although there was some degree of alignment between qualitatively and quantitatively generated themes, the topics lacked nuance, and the representative documents for the themes in common between the two methods differed [83]. The theoretical congruence of topic models and discourse analysis [64] and many other qualitative analysis approaches [63] have been discussed at length.

Although directly comparing results can provide evidence of the topic model's relevance to the qualitative context, more often topic models are seen not as a replacement for qualitative analysis, but as a supplementary method to enhance qualitative conclusions (e.g., [25]). Isoaho et al. review a number of applications of topic models mixed with qualitative methods and comment on the suitability of each [63]. Cahill suggests an iterative procedure for topic model analysis, whereby the analyst first familiarizes themselves with the data through a close reading, performs the topic model, and interprets the results with a qualitative lens [23]. In contrast, to minimize some of the inherent biases of qualitative research, Debnath et al. propose a mixed grounded theory and topic modelling analysis where topic modelling is used blindly on the text, and the high probability words and associated documents are used to shape the initial directions of the narrative analysis [33]. Nelson develops a content analysis approach whereby the text is first explored quantitatively, uses a mix of quantitative and qualitative approaches to refine ideas and interpretations, and then returns to quantitative methods of validating observed patterns [89]. Many other potential integrations of topic models and qualitative research exist.

Strengths of topic modelling for representing text are its relative rapidity when applied to large datasets, its reproducibility, its ability to make additional quantitative inferences on the extracted topics, and that the classifications of the text into themes proceeds independently of the researcher [89, 25]. The perceived objectivity of the topic model should not be overstated, however: many subjective decisions, including the number of topics, textual components to include and exclude, and text segmentation, must be made before the analysis can proceed, and once the topic model is fit, the resulting themes must be interpreted within the bounds of the corpus' subject [89]. That these choices and interpretations are context-dependent does not diminish the possibility of a rigorous topic model analysis, but it necessitates a good understanding of the limits of topic modelling, transparent discussion of modelling choices, and validation [30].

2.3 Modelling the data output of focus groups

The present thesis will not deliberate the congruence of topic models with qualitative approaches in the theoretical sense. Instead, we focus on the applicability of topic models to a specific qualitative data source: focus groups.

2.3.1 What are focus groups?

Focus groups are a data collection tool that take the form of a facilitator- (or "moderator")-led discussion with multiple research participants simultaneously. They have their origins in marketing research, where they are often chosen in order to collect many opinions on or experiences regarding a given subject in a short period of time [87].

Focus groups are often described as multi-person interviews. Relative to interviews, each individual participant in the focus group will typically say less, but obtaining multiple perspectives can be done more quickly [87]. One of the strengths of focus groups relative to individual interviews is that they allow for discussion. Participants will provide multiple

contrasting or confirmatory points of view to a narrative or description. Further, the discussion in focus groups may give voice to underlying beliefs that people may not typically articulate [18]. Also, using a focus group can also help limit the impact of the researcher's presence on the conversation: unlike in an interview, the participants are describing their experiences in a group composed mostly of their peers [87, 54].

Participant sampling is a fundamental aspect of all data collection. In a focus group, participants are often sampled such that they share an attribute in common (such as location, personal characteristics, or interests), to foster a sense of belonging and encourage discussion [87, 111]. The number of groups and the size of the focus group can vary: the former often depends on the diversity of the perspectives of the population under study; the latter is a question of resources and dynamics: a small focus group virtually guarantees everyone's perspective will be shared in depth while a larger one might only obtain a few answers from each person in the interest of time [87].

Most but not all focus groups rely on a "guide": a list of questions that the facilitator aims to ask during the session [111]. The guiding prompts made by the facilitator should generally be open-ended, leaving room for the participants to develop their answer. The extent to which the facilitator attempts to adhere to the phrasing and order of the questions in the guide, and how much or what style of leadership the facilitator is responsible for, should also established beforehand [111]. A "semi-structured" approach is often used, in which the questions are asked in whatever form feels most natural for a conversation, but "self-managed" focus groups, where the guide has low to no importance, can also be performed [87].

However, a focus group is inherently a constructed environment with a dynamic between the research(er) and participants and between the participant-peers that will shape the output. As *de facto* leader of the group, the facilitator sets the tone for the conversation [111]. It has been suggested that sensitive and highly personal topics that people do not typi-

cally discuss with strangers (even if the strangers are peers) may be difficult to research using focus groups [87], however this may depend also on the group dynamics and the researcher's positionality [69, 54]. The focus group environment is artificial in that the discussion would not occur naturally in this setting and inferences surrounding the flow of the conversation must acknowledge this; it differs in this way from participant observation, wherein a researcher is explicitly attempting to observe typical dynamics between subjects [87]. Extensive discussion exists about how to create a focus group setting which enables the most natural discussion [111].

The primary data output of a focus group is typically a verbatim transcript with each person's speech annotated by their name or another indicator. The transcript may have non-verbal information as well, such as participant body language or tone cues noted by the transcriber (or by a notetaker who is present for the focus group session but does not engage directly) [111]. Once prepared, the transcript can be analyzed according to one of several qualitative methods informed by the qualitative school of the researcher and the research question.

2.3.2 The potential of topic models in the analysis of focus groups

The volume of text data available at the end of even a small focus group study means that traditional qualitative methods may take a long time to implement; a complementary machine-based analysis may be integrated into a study for rapid, quantifiable insights [125].

Some discussion of the unique features of focus groups as a source of text for quantitative methods occurs in [74]. First, transcripts can be long, meaning a significant amount of textual data is available as the output of a focus group study. Second, focus group discussions are held between multiple people, which can provide interesting variety in word co-occurrence as the participants connect their experiences and opinions to the discussion

and try to articulate their thoughts in a semi-public setting. Participant speech will often be relatively casual and conversational– although this will depend on the participants, environment, and topic. Third, since the transcripts are verbatim, they are a record of locution: there is no editing or refining word choice like for written text, and all self-correction is incorporated rather than removed. Finally, the semi-structured nature of the discussion, where the facilitator will occasionally redirect participants towards the guiding questions, means that the subject of discussion is predetermined. However, the extent to which the responses relate directly to the question depends both on participants and on the role of the facilitator. Attractive to both qualitative research and topic modelling is the potential for additional patterns to arise in the narratives and opinions of the participants outside of the questions asked.

The Structural Topic Model of Roberts et al. [100, 97] in particular may yield interesting insights in the context of a focus group analysis.

For one, like the CTM [16], the STM estimates a correlation structure between the topics. This aligns well with our expectations: the conversational nature of focus groups means that people may draw connections across many discussed themes in a single text segment. Furthermore, the topics are unlikely to be independent on principle, since the focus group guide is made up of related questions. To the extent to which the topics (as qualified by their most probable words and most related documents) represent themes of the focus group discussion, the correlation between topics could be seen as a measure of the relatedness of these themes.

For another, the STM is relatively adept at handling rare words and short documents when compared to similar topic models– although this is not its primary purpose. Since the transcript of focus groups is largely dialogue, the potential for both colloquialisms (or person-specific uses of language) and interjections is high [74]. To account for rare words, the STM enables the option to use a SAGE (Sparse Additive Generative model) in the esti-

mation of topic-word proportions, which has been shown to better attribute rare words to topics than LDA [38, 99]. Furthermore, the STM pools information about topic prevalence across documents using covariates: even in the absence of many co-occurrences between words in some documents, information about the expected topic prevalence of documents of similar covariate levels is provided to the model and could have a stabilizing effect. However, this stabilization relies on the assumption that the covariates are somewhat accurately representing the relationships between documents in terms of language use, and that not all documents are small (some co-occurrence information is required to inform the topic prevalence).

Furthermore, the STM provides a way of utilizing the rich external information about the focus group and its text [97]. At the very minimum, the prevalence of the topics within a focus group may be expected to vary between different groups, since the amount a given question from the guide is discussed and the words used to discuss it depend on the unique participants of that group. Other covariates about the participants, guide questions, or groups may also be of interest. Using covariates to estimate the model may help in assigning words to topics, and could be used in post hoc validation of the topics by estimating associations with metadata.

Applying an STM to focus group data could thus provide evidence to answer research questions such as:

- What are the topics discussed in the focus groups? How do these topics tend to co-occur in documents? What inferences can we make about the relationship of language to the research question of the focus groups, using the estimated topics?
- Does the amount a given topic is discussed (topic prevalence) differ according to the characteristics of the facilitator (age, sex, relative social position)? According to the characteristics of the participants or focus group session (age, sex, role, relationship to topic)?

- How do the most probable words in a topic (topic content) change according to the characteristics of the participant? Are different words used to describe the topics depending on the focus group session and facilitator choices regarding questions and direction?

As with any topic model analysis, careful delimiting and preprocessing of text must occur before the topic model can be used. The preprocessing involved in topic models can have dramatic implications for the results of the model [30], but must be implemented on a case-by-case basis as it relies on knowledge of the corpus and the desired outcomes of the topic model. We will highlight a number of important considerations for the preprocessing of text sourced from a focus group or other multi-person interview, specifically.

2.3.2.1 Defining documents in focus groups

In many corpora used for topic modelling, the document is naturally embedded in how the text was collected. In the case of focus group transcripts, however, the choice of "documents" is not self-evident [4].

An initial proposal for the document could be each focus group session's transcript. This may be a reasonable proposal in interview studies, where the number of transcripts can be large and each has a clear unification by being generated from the same participant. Focus group studies, however, tend to generate fewer transcripts in total since they pool participants into groups [87], and it can be difficult to infer distinct topics when few documents exist to contrast against or when the vocabulary is largely shared [74]. Although these transcripts contain many ideas articulated by different people, the topic models considered here collapse all text into one "bag of words". It becomes impossible to disentangle the relationship of individual participants' speech to the topics when using whole transcripts as documents. Also, when the documents are too broad, it can complicate the ability to make inferences or validate results from a semantic or external lens (discussed further in section 2.3.2.1.II). As such, we advise against using whole focus group transcripts as documents

in topic modelling.

I. Length and partitioning of transcripts. Determining how to define documents in an existing corpus is often viewed as a balance between number and length of documents. It is well understood that topic models require a sufficient amount of data in order to be fit with any amount of confidence [30], although the amount of data required is debated and differs between topic modelling strategies. No empirical studies around sample size yet exist for the STM that we know of, however Tang explored the implications of document number and length for LDA, and concluded that topic distinction suffers when too few documents exist in the corpus (even if these documents are long) [114, 74]. On the other hand, many short documents are also not easily fit to using LDA: LDA assumes that each document is a mixture over topics, and thus when limited co-occurrence values are available within a document, there is insufficient information to infer precise document-topic proportions. The information provided to the CTM and STM in the form of between-topic correlations and pooling across documents using covariates supplement the word co-occurrences. So, even with fewer words per document than typically minimally expected for an LDA model, topic modelling may be able proceed relatively well on a corpus with some short documents in these models– although this merits additional empirical research.

The question of documents in focus groups thus becomes one of how to segment the transcripts. There is huge diversity in the approaches to document definition. Atkins et al. were among the first to apply topic models to multi-person interviews and use a person-level partitioning of text [4]. Miyaoka et al. also defined documents as all words said by one participant [84], Liang et al. partitioned each focus group session transcript by the facilitator-asked question [77], and Debnath divided a single focus group transcript by considering sentences [33]. Burrow et al. partitioned their focus group transcripts by talk turns (i.e., all words said by one participant between the speech of any others) from participants [22]; Mols et al. applied a qualitative approach to defining documents, by selecting text

fragments that had a direct relationship to the research question [85]. The number of documents in these studies ranged from 16 [84] to 568 [22], but studies very rarely specified document length. In practice, even among the few studies which have applied a topic model to focus group transcripts, not all specify at which level they define their documents. For the interpretation of results, this is a serious omission: it becomes impossible to perform the already difficult task of judging whether the assumptions of the topic model vis-à-vis data are met and to understand the context of the results.

II. Document partitioning in the context of interpretation. Another side to the transcript segmentation question is the meaning of the document partition in the context of the research question and topic interpretation.

Different segmentations can yield different high probability words: Le performed a quantitative comparison of LDA implemented on focus group transcripts partitioned into documents by question and by participant, and found that splitting focus group transcripts by speaker resulted in improved semantic coherence, but less exclusive topics [74].

Depending on the partitioning, the top documents for topic k (those with the highest document-topic$_k$ proportions) will differ, and so will document-level conclusions. Consider two cases of a topic model fit to transcripts: one where the documents are defined by the facilitator questions, the other where the documents are defined by the focus group participant. In the first case, these top documents are, "responses to facilitator questions with a high proportion of topic k"; in the second case: "participants whose speech has a high proportion of topic k". If the sample of participants across focus groups is relatively homogeneous, the former partitioning may be desirable as a way to explore topic prevalence and top documents by question. If on the other hand, the focus groups were held with disparate populations (e.g., residents of different neighbourhoods), the second document partitioning may be more valuable, as comparisons by topic can be made between population groups. This relates to the idea of external validation proposed by DiMaggio et al. [36]: in most

cases, the external delimiter must fall between documents, not within them, to be able to contrast the estimated topic prevalence or content.

In the STM, the document partitioning has an explicit relevance for both the additional data that can be incorporated into the model and for post hoc inferences. The covariates in the topic content and prevalence models can only be incorporated at the document-level or higher [97, 99]: if the covariate is measured on a unit of text that is a higher granularity (smaller partition) than the defined documents, it must be aggregated to the document level for incorporation. However, this may not always be possible: suppose the covariate "participant occupation" is of interest to the analysis in that it is hypothesized to affect topic prevalence. Then, if the documents are defined at the participant level, this covariate is easily incorporated into the STM (one occupation per document). However, if the documents are defined by question, it becomes more challenging to specifically associate the covariate with the correct elements of text, since each document contains responses from multiple participants. Furthermore, the outcome variable of the post hoc regression estimating the effect of covariates on topic prevalence are *document*-topic proportions, meaning the conclusions are drawn at the document-level. Consider the difference between, "female participants spoke X% more about topic A" and "following question 1, participants spoke X% more about topic A". This regression too can be used for external validation, meaning to enable comparisons, one must define documents such that variables of interest can be incorporated into the regression.

As such, document partitioning should be decided in concert with the other information that researchers will use to estimate and validate the model, and the level of desired conclusions.

III. For the analyst. In a practical light, the multitude of ways in which documents can be defined means that no reader has an innate understanding of which has been chosen for the topic model analysis. It becomes important, then, that the analyst not only defines this to the

reader at the outset of their topic model, but justifies the segmentation. We have provided three methods of justification, namely: size and number of documents, interpretation and validation of the segmented text, and, for some topic models, the incorporation of metadata. We recommend that the analyst provide a justification in the context of one or more of these for a principled approach to topic modelling focus group transcripts.

2.3.2.2 Facilitator speech

As with much of qualitative research, the researcher or researcher's actor are embedded in the research context. Facilitator speech is part of focus group transcripts, and the analyst must decide whether and how to use it.

I. As part of the documents. On one hand, the facilitator speech can contain valuable text: they will ask the intended questions, clarifying questions, and summarize or reiterate what participants say to ensure a good understanding. In this respect, keeping facilitator speech and integrating it into the documents can help contextualize short responses and provide more meaningful words per document. This makes sense particularly if the documents are partitioned by question or by talk turn, and was the approach taken by Manych et al. for topic modelling interview transcripts [80].

On the other hand, most topic models which rely on the bag-of-words assumption cannot distinguish facilitator from participant words[1], so the resulting inferences can depend on how the facilitator speaks. If there is a distinction between the way in which the facilitator describes an issue and the words the participants use, the highest probability words in a topic may no longer represent the population of interest. Similarly, the facilitator may provide contextualizing narratives or examples which, although relevant, are not shared

[1]The Multi-Field CTM of Salomatin et al. [103] and the Pairwise Topic Model of Jiang et al. [66] are examples of topic models designed to capture multiple sources of text data. These too could have interesting applications for focus groups, where one could model facilitator and participant speech separately in one model, but our interest is in the STM.

by the participants whose experiences are being studied– and would be excluded from a traditional qualitative analysis in favour of the participant responses to the story [85]. Including facilitator text will change the interpretation of and conclusions drawn from a topic model.

When documents were defined by the speaker, Le found using the facilitator text had little bearing on the performance or coherence of an LDA model [74]. However, it is difficult to generalize this result, since, as discussed, the position of the researcher with respect to the topic and the participants (i.e., how actively they guide and interact with the discussion) and the length and number of the documents with or without the facilitator speech will influence the resulting topics, model fit, and interpretation.

In the STM, no missing values are permitted in prevalence or content covariates [99], which complicates facilitator inclusion: the facilitator's speech can only be included if the covariates can be attributed to the resulting documents.

II. As a covariate. The semi-structured focus group in particular has a predefined set of themes which will be explored as the facilitator directs the discussion. The questions asked prior to a participant's contribution influence the content of the response, which, in the context of topic modelling, implies a change in topic prevalence. This is thus a natural source of pooling across documents and motivates the use of facilitator speech as a covariate.

Identifying the questions from facilitator text in semi-structured focus groups is in and of itself an exercise in theme extraction. Being semi-structured, the questions take on many forms and may not always reflect the order or wording suggested on the guidance sheet. Furthermore, additional questions outside of the initial scope of the guidance document may be used, when they become relevant and to explore themes that arise as important to the participants.

One method to include facilitator text would be hand-coding: for each facilitator talk turn, the text could be coded thematically (using an iterative, qualitative, theme-building approach) or strictly according to the number and presence of the question on the focus group guide (using a more objective, categorical approach). In either case, there is some subjectivity involved in tagging the questions. The resulting covariate would be a categorical variable representing the presence of a certain question.

Another method would be to use topic modelling or another semi-supervised or unsupervised machine learning method on the facilitator text. This would be able to flexibly extract topics of the questions either on a continuous scale of topic prevalence or using a categorical presence/absence of the topic in each document after some thresholding. However, this approach struggles in its interpretability: the resulting clusters of words or text must first be interpreted before being included as covariates. Furthermore, it would be challenging to incorporate the uncertainty around the resulting "facilitator topic" estimates into their use as covariates.

2.3.2.3 Other considerations for focus group transcript processing

Memos and contextual information: Focus group transcripts may contain notes that provide additional information about a given piece of speech, like body language or tone [111]. When taking a qualitative approach to the analysis, this information can provide valuable context. However, using this information as text in the STM and other topic models discussed here carries some of the same implications for analysis as using facilitator speech: it will influence word co-occurrence, since the model does not distinguish between two sources of text. One option to differentiate word context in a bag-of-words model could be through annotation. For example, the word "laughter" could be changed to "Xlaughter" when used in notes, and simply "laughter" when used in speech. Then, both sources could be interpreted separately in terms of co-occurrence. Alternately, the presence of a certain note could be used as a covariate in the STM, if it were assumed to affect topic preva-

lence or content. The incorporation of this information merits additional consideration and requires a standardized approach across the corpus.

Anonymization: A challenge in data cleaning for qualitative research is that of anonymization. Redacting participant identifying information is a necessary step to protect participants and their privacy. However, anonymization will change the corpus itself: the omission of names, locations, and other identifying information affects the length of documents, the word frequencies, and the size of the vocabulary, and reduces the words that are different between focus groups. A completely redacted text might have very little variation between documents or units of analysis, making topic modelling challenging. On the other hand, it would be inadvisable to run a topic model on a text before anonymization, as it might become difficult to describe the topics– in terms of the highest probability words or documents– without revealing sensitive information. Also, two methods of redacting information can provide different levels of context (e.g., using "(F1)" in place of a name could tell the analyst the participant gender, but "[name]" would not). Standardization is key here as well so that the words do not vary systematically due to a factor other than participant word use. When a mix of practices is used, the most concealing one will be the only resulting option.

2.3.3 Motivating the mixSTM through focus groups

Since the documents in the case of focus groups are partitions of the transcripts, they are inherently grouped hierarchically within the focus group sessions. We anticipate that the documents within a given focus group share some dependency in their content (concretely in terms of words and more generally in terms of themes), because they are related through the conversation's progression and facilitator's guidance. At the very minimum, the prevalence of the topics within a focus group may be expected to depend on this grouping, since the amount a given question from the guide is discussed and the words used to discuss it depend on the unique participants of that focus group. The relationship of other variables to

topic prevalence may also vary between focus groups, and the relationships between topics may change depending on the session and experiences of the participants.

Explicitly incorporating the grouped structure of focus group documents into a model's estimation and interpretation may align more realistically with our assumptions about topic prevalence in these corpora.

Chapter 3

3 Variational Expectation-Maximization in the STM

3.1 Statistical background

Understanding the semi-collapsed variational expectation-maximization algorithm used to fit the Structural Topic Model requires concepts from Bayesian inference.

3.1.1 Principles of Bayesian modelling

In the Bayesian paradigm, rather than obtaining point estimates of the parameters that correspond to the values obtained if the experiment were repeated many times, parameters are expressed in terms of their distribution given the observed data and any previous beliefs about the parameter values. To write this in terms of conditional distributions, for parameter v and observed data t:

$$p(v|t) = \frac{p(t|v)p(v)}{p(t)} \implies p(v|t) \propto p(t|v)p(v) \tag{3.1}$$

The first equation is known as Bayes rule, where $p(v)$ is the prior and represents a distributional prior belief about the parameter of interest v; $p(t|v)$ is the likelihood and represents how probable the observed data is given the parameter; and $p(v|t)$ is the posterior, the distribution of the parameter given the data and prior. The denominator of (3.1), $p(t) = \int p(t|v)p(v)\mathrm{d}v$, is a normalization constant for the posterior distribution and referred to as the marginal likelihood of the data. Obtaining the form of the posterior distribution is a central question in Bayesian inference.

3.1.1.1 Priors

Both the likelihood and the prior influence the shape of the posterior distribution. If one has few observations, prior beliefs about the parameters are weighted more heavily in the calculation of the posterior distributions of the parameters. Conversely, the more data available in the likelihood function, the less influence the prior has on the posterior. A choice between priors is often a choice between informativity and non-informativity, where informative priors deliberately give a larger weight to some region in the parameter's domain and non-informative priors are often broad distributions representing weak prior beliefs about the parameter's most probable region.

Finding the posterior distribution of a parameter can be complicated, as it involves calculating the marginal likelihood which may not always be a tractable integral. However, some combinations of distributions and likelihoods have an appealing property that always permits an analytical, exact solution to the posterior distribution: conjugacy. For a given likelihood function, a prior is called a "conjugate prior" if it takes the same form as the posterior. For an example relevant to topic modelling, if the likelihood takes the form of a multinomial distribution and the prior on the parameters is a Dirichlet, the posterior distribution over the parameters given the data will also be a Dirichlet. In contrast, the logistic normal distribution, which is defined as the probability distribution of $\beta = \frac{e^b}{\sum_i^K e^{b_i}}$ when $b \sim \text{Normal}_{K-1}(\mu, \Sigma)$, is not conjugate to the multinomial [1, 16].

3.1.1.2 Posterior quantities and inference

We can use the posterior distribution of a parameter, v, to quantify the uncertainty in v and, if desired, obtain point estimates for our parameter.

The maximum a posteriori or MAP estimate of a variable is the mode of its posterior distribution [11]: the "most probable" value of the parameter under the posterior. The MAP estimate for a parameter v with prior distribution $p(v)$ given data t with likelihood

$p(t|v)$ is:

$$\hat{v}_{\text{MAP}} = \arg\max_{v} p(v|t) = \arg\max_{v} \frac{p(t|v)p(v)}{\int_v p(t|u)p(u)\mathrm{d}u} \tag{3.2}$$

In some cases, including when the posterior distribution over v is normal, the MAP estimate coincides with the mean of the posterior distribution. Note that finding the MAP estimate for v does not require the potentially intractable marginal likelihood of the data, since the marginal likelihood is positive by definition.

Another distribution that is often of interest is the posterior predictive distribution: the probability distribution of an outcome variable over new data, given the posterior distribution of the parameters and the distribution of the outcome variable given the parameters [11]. For some new data, t_{new}, the posterior predictive distribution is defined as [11]:

$$p(t_{new}|t) = \int p(t_{new}|v)p(v|t)\mathrm{d}v \tag{3.3}$$

Defining and using the posterior distribution becomes challenging when the marginal likelihood is intractable, so approximate methods must be used. A common and widely used approach to approximate inference is the class of Markov chain Monte Carlo (MCMC) methods. MCMC methods take a random walk through the parameter space, such that every step only relies on the previous step's information, and each step hopefully results in a collection of sampled parameters that is closer to the posterior distribution of interest, although there are no guarantees [47]. MCMC is powerful in that, given enough iterations, the distribution sampled from at each step will often converge upon the true posterior distribution, so the samples can be used to determine measures of central tendency and variance about the parameters. However, MCMC methods can be slow to converge, particularly for large, complex models: commonly, the first steps (as many as several thousand draws) in the chains are discarded as "burn-in" or warm-up draws and not used in posterior inference. Monitoring and assessing the convergence of an MCMC fit requires additional steps.

As an alternative to MCMC methods, variational inference can be used.

3.1.2 Variational inference

Variational methods find a more easily computable approximation to an intractable distribution and use this approximation for downstream inference. Identifying the form of the approximate distribution involves iterating to obtain optimal parameter values which maximize the approximation's similarity to the true distribution of interest [11, 13].

The approximating distribution can take any form. In practice, the "mean-field approximation" is often used, which approximates the distribution using a product of independent distributions over disjoint partitions of the latent variables of the model [13]. More formally, consider a variable of interest, t, and some latent variables ϕ and v, which parameterize the distribution of t. The latent variables (v, ϕ) are not restricted to be individual variables and can be vectors, though for the simple example here, we will restrict our case to two latent univariate variables. The goal of variational inference is then to compute an approximation to the posterior distribution of interest, $q(\cdot)$, instead of computing the true posterior, $p(\cdot)$ [11]. The notation $q(\cdot)$ is used rather than $p(\cdot)$ when referring to the variational distributions; $q(\cdot)$ may thus refer to any number of different probability distributions and the specific variational distribution at hand is implied through its argument(s).

To do so, mean-field variational Bayes proceeds in two main steps [13]:

1. Use the mean-field approximation and assume the approximate posterior over the latent variables can be factorized into a product of independent distributions [13].

$$p(v, \phi) \approx q(v, \phi) = q(v)q(\phi)$$

   The partition of the latent variables into disjoint subsets is often chosen such that the resulting functional forms in $q(v)q(\phi)$ can be mapped to known distributions. The

partitioning can contain as many disjoint subsets of parameters as desired. Not all parameters in the model need be part of the partitioning, for example, when one is not working in a fully Bayesian context and parameter values are fixed rather than being assigned a prior.

2. Minimize the Kullback-Leibler ("KL") divergence from the approximate to the true posterior to find the closest approximation to the posterior under this partitioning of the variables. The KL divergence between two distributions $p(\cdot)$ and $q(\cdot)$ is defined as [70, 11]:

$$\mathrm{KL}(q(v,\phi)||p(v,\phi|t)) = -\int q(v,\phi)\log\frac{p(v,\phi|t)}{q(v,\phi)}\mathrm{d}v\mathrm{d}\phi = \int p(v,\phi|t)\log\frac{p(v,\phi|t)}{q(v,\phi)}\mathrm{d}v\mathrm{d}\phi \tag{3.4}$$

The KL divergence can also be interpreted as a statistical distance between any two vectors of probabilities, but it is not symmetric.

It can be shown (e.g., [11] Section 10.1) that minimizing the KL divergence between the factorized variational posterior $q(\phi)q(v)$ and the true posterior $p(\phi, v|t)$ is equivalent to maximizing the lower bound on the (log) marginal likelihood. We can define a lower bound on $\log p(t)$ using Jensen's Inequality, which states that if X is a random variable and f is a concave function, then $f(E(X)) \geq E(f(X))$. Applying this to our example, we can obtain a bound as follows [11]:

$$\begin{aligned}
\log p(t) &= \log\left[\int\int p(\phi,v,t)\mathrm{d}v\mathrm{d}\phi\right] \\
&= \log\left[\int\int p(\phi,v,t)\frac{q(\phi,v)}{q(\phi,v)}\mathrm{d}v\mathrm{d}\phi\right] \\
&= \log\left[E_q\left(\frac{p(\phi,v,t)}{q(\phi,v)}\right)\right] \\
&\geq E_q(\log(p(\phi,v,t))) - E_q(\log(q(\phi,v))) \\
&\triangleq L(q)
\end{aligned} \tag{3.5}$$

where E_q denotes the expectation with respect to the variational approximation to the posterior $q(\phi)q(v)$. $L(q)$ is referred to as the Variational Objective or the Evidence Lower Bound (ELBO) [13]. By maximizing the ELBO with respect to the parameters of the distributions of the latent variables of interest, we can find the form of the approximate posterior which is closest in KL divergence to the true posterior.

The question remains how to choose the form of the approximating distributions for each partition of the latent variables. It can be shown that an optimal form for $q(\phi)$ and $q(v)$ (the "variational distributions") exists; we denote the optimal form $q^*(\cdot)$. To find $q^*(\phi)$, we use the following identity [11]:

$$L(q) = -KL(q(\phi)|| \exp(E_{q(v)} \log p(\phi, v, x) + \text{const}))$$

Analogously for $q^*(v)$, the ELBO can be written as the negative KL divergence between the variational distribution $q(v)$ and the exponential of the expectation with respect to the other latent variables of the log joint probability. The minimum of $L(q)$ is found when the magnitude of this KL divergence is minimized. The KL divergence, from its definition in (3.4), is zero if and only if the numerator and denominator are equal which means the optimal form of the variational distribution $q^*(\phi)$ can be found using:

$$q^*(\phi) \propto \exp(E_{q(v)} \log p(\phi, v, x)) \tag{3.6}$$

where for any partition of the latent variables (here, ϕ), the expectation of the joint probability is taken with respect to all latent variables not in that subset (here, v) [13].

Given the complete form of the ELBO, variational inference iterates through updates to the parameters of the variational distributions in order to maximize the ELBO and find the approximating distribution closest to the true distribution of interest.

3.1.2.1 Limitations of variational inference

Some limitations of variational inference as outlined here must be acknowledged:

- The approximate posterior with optimized parameters that we obtain at the end of the variational inference updates will only ever be an approximation. Unlike MCMC, even with infinite iterations, variational inference will not converge upon the exact posterior or distribution of interest. There is no guarantee that this distribution resembles the true posterior in form, although it hopefully has density in similar regions [11].

- Mean-field variational inference cannot capture the correlation between latent variables in disjoint partitions. This has the related problem that the marginal variances of the approximation tend to underestimate the variance of the distribution being approximated [11].

- There are generally many local optima for the ELBO, so the variational inference algorithm may arrive at different approximations to the posterior depending on the initialization. Re-initializing the algorithm with different starting values is often recommended [99]. Fortunately, even when a local optimum on the ELBO is found rather than the global one, it is still a lower bound on the marginal likelihood, and all the same conclusions apply [11, 97].

3.2 The Correlated Topic Model

The key difference between the Correlated Topic Model (CTM) and LDA is the logistic normal distribution used in place of the Dirichlet distribution for the topic proportions in the generative model. The non-conjugacy of the logistic normal and the multinomial distributions means a variational method is often used to fit the CTM.

3.2.1 Introduction to the CTM

For a corpus with D documents $d = 1, ..., D$ with N_d words in document d ($n = 1, ..., N_d$), K latent topics $k = 1, ..., K$, and V ($v = 1, ..., V$) unique words composing the vocabulary, we can write the generative model of the CTM as the following:

$$\begin{aligned} \eta_d &\sim \text{Normal}_{K-1}(\mu, \Sigma) && \text{for } d = 1, ..., D \\ z_{d,n} &\sim \text{Categorical}_K(\theta_d) && \text{for } n = 1, ..., N_d \\ w_{d,n} &\sim \text{Categorical}_V(\beta_{z_{d,n}}) && \text{for } n = 1, ..., N_d \end{aligned}$$

In order to define the parameter θ_d on a simplex, we set $\eta_K = 0$ and set $\theta_d = \frac{\exp(\eta_{d,k})}{\sum_i^K \exp(\eta_{d,i})}$, a $K \times 1$ vector of topic proportions for each document. Since η_d follows a normal distribution, θ_d follows a logistic normal distribution [1]. The correlation between topics is introduced through the covariance matrix on document-topic weights, Σ, which is shared across all documents and implies the correlations between topics are the same for all documents in the corpus.

In words, the generative model assumes that, for each document, d, a vector of document-topic weights is drawn from a normal distribution with mean μ and variance-covariance matrix Σ, and is projected to the $K - 1$ simplex θ_d. For each word n in document d, a topic assignment $z_{d,n}$ is drawn from a categorical distribution (a multinomial distribution with one draw) with probability θ_d, and a word is drawn from a categorical distribution over the vocabulary with a mean probability β_z that depends on the topic assignment [14, 16].

In comparison to LDA, the CTM shows an improvement in predictive perplexity and held-out likelihood, which measure the ability to infer the remaining words in a document when some are omitted during fit [14, 16]. This makes intuitive sense: since the CTM has more information about the topics thanks to the between-topic correlations, it can make more accurate predictions about the topic assignment of held-out words [16].

3.2.2 Variational Expectation-Maximization for the CTM

As noted in section 3.1.1.1, the logistic normal distribution is not conjugate to the multinomial, because of the interdependence of the topic proportions and the topic assignment [16]. Blei and Lafferty additionally point out that the fast Gibbs sampler employed by LDA is no longer possible for the CTM and the alternative Metropolis-Hastings MCMC algorithm would be prohibitively slow given the dimensions of the data on which topic models are fit [16]. This motivates the development and use of a variational expectation-maximization (EM) algorithm to rapidly obtain posterior estimates in the CTM [16]. We focus here on some key definitions and results from the CTM which will be relevant to the STM, although interested readers are directed to the appendix of [16] for complete details.

The goal of variational EM is to find the values of the variational and non-variational parameters which maximize the bound on the likelihood of the documents [16]. Variational EM replaces the typical E-step of the expectation-maximization algorithm (which uses the posterior distribution of the data given the parameters) with a variational approximation to this posterior [16], and then updates the parameters of the distributions accordingly. In the M-step, the model parameters are chosen such that they maximize the bound, holding the variational parameters fixed. These two steps are iterated through until convergence.

The posterior distribution over the latent variables given the observed data w for the CTM in any document d is proportional to [16]:

$$p(\eta_d, z_{n,d}|w_{n,d}) \propto N(\eta_d|\mu, \Sigma) \prod^{N_d} \left[\text{Categ}(z_{n,d}|\theta_d) \times \text{Categ}(w_{n,d}|\beta_{d,k=z_{n,d}})\right]$$

The CTM approximates this per-document posterior using the mean-field approximation, such that for each document we assume the posterior over the latent topic weights (η) and topic assignments (z) can be factorized into $N_d + 1$ disjoint subsets of the latent variables

[16]:

$$p(\eta_d, z_{n,d}) \approx q(\eta_d) \prod^{N_d} q(z_{n,d}) \tag{3.7}$$

On a per-document basis, using equation 3.5 leads to the following formulation of the lower bound on the log marginal likelihood [16]:

$$L(q) = E_q(\log p(\eta_d, z_{n,d}, w_{n,d})) - E_q(\log q(\eta_d, z_{n,d})) \tag{3.8}$$

Using variational inference and a coordinate ascent algorithm, the variational parameters can be updated to maximize this per-document ELBO [16]. This involves iteratively maximizing over each of the variational parameters using gradient-based optimization. Applied to each document, this is the variational E-step and allows us to infer the topic weights for each document and the topic assignment of each word index. The overall bound on the likelihood, obtained by summing (3.8) across all documents, is then maximized with respect to the model parameters in the M-step using maximum likelihood estimation [16].

Since its introduction, modifications to the original variational EM algorithm for the CTM have been suggested to improve the bound or accelerate convergence [67]. Furthermore, some collapsed Gibbs samplers have been developed for classes of topic models which use logistic normal distributions [29, 81].

3.3 The Structural Topic Model

The Structural Topic Model uses the logistic normal distribution as in the CTM, with the addition of document- or higher-level covariates in the estimation of topic prevalence and content. Roberts et al. use a semi-collapsed version of the variational expectation-maximization algorithm used to fit the CTM in order to fit the STM [97, 99].

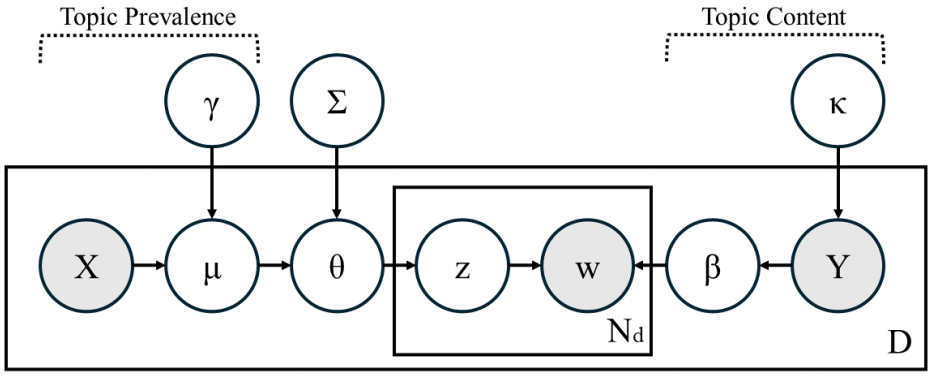


Figure 3.1: Plate diagram representing the Structural Topic Model (STM)

3.3.1 Introduction to the STM

Adopting the formulation from Roberts et al., in addition to the variables defined in section 3.2.1, the STM allows for P topic prevalence covariate terms ($p = 1, ..., P$) stored in the design matrix $\mathbf{X}$ to provide information to the document-topic proportions θ, and R categorical content covariate levels stored in the matrix $\mathbf{Y}$ to provide information for the topic-word proportions β [97]. The STM then assumes the following generative model [97]:

$$\begin{aligned}
\gamma_k &\sim N_p(0, \sigma_k^2 \mathbf{I}_p) && \text{for } k = 1, ..., K-1 \\
\eta_d &\sim \text{Normal}_{K-1}(\mathbf{\Gamma}' x'_d, \mathbf{\Sigma}) && \text{for } d = 1, ..., D \\
z_{d,n} &\sim \text{Categorical}_K(\theta_d) && \text{for } n = 1, ..., N_d \\
w_{d,n} &\sim \text{Categorical}_V(\mathbf{B} z_{d,n}) && \text{for } n = 1, ..., N_d
\end{aligned} \tag{3.9}$$

To allow covariates to affect topic prevalence, the STM includes topic-specific covariate weights, γ_k, applied to the covariates stored in $\mathbf{X}$ in order to obtain the mean of the distribution of document-topic weights per document (η_d). The hyperprior on σ_k^2 (the variance of the topic-specific distribution of covariate weights) is an inverse gamma distribution [97]. As in the CTM, topic proportions are obtained from the weights: $\theta_d = \frac{\exp(\eta_{d,k})}{\sum_i^K \exp(\eta_{d,i})}$

[97]. Additional details of the topic prevalence model will be discussed in the following chapters.

The topic content model in the STM, which is outside the scope of the present work and will not be addressed at length, involves a single categorical covariate in $\mathbf{Y}$ which is allowed to affect topic content. The parameter of the categorical distribution over words, β, is defined in terms of deviations from a baseline transformed frequency of the occurrence of word v across the whole corpus (m_v), where deviations are permitted to arise due to the assigned topic $\kappa^{(t)}$, levels of the content covariates $\kappa^{(c)}$, or the interaction between the topic and the covariate $\kappa^{(i)}$ [97]. For a given document d, word v and topic k, the probability distribution over the vocabulary is parameterized by [97]:

$$\beta_{d,k,v} = \frac{\exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})}{\sum_i^V \exp(m_i + \kappa^{(t)} + \kappa^{(c)} + \kappa^{(i)})}$$

$\mathbf{B}$ is the the $V \times K$ matrix whose columns are β_k.

The overall posterior distribution over the latent variables given the observed data (w, Y, X) for the STM is proportional to [97]:

$$p(\eta_d, z_{n,d}, \kappa, \Gamma, \Sigma | Y, X, w_{n,d}) \propto$$
$$\prod^{D} \left[\text{Normal}(\eta_d | X_d\gamma, \Sigma) \prod^{N} \left[\text{Categ}(z_{n,d}|\theta_d) \times \text{Categ}(w_n|\beta_{d,k=z_{n,d}}) \right] \right] \prod p(\kappa) \prod p(\Gamma) \quad (3.10)$$

As in the CTM, the STM uses variational EM to find an approximation to the true posterior distribution that minimizes the KL divergence between the truth and the approximation, and uses this approximation for downstream inference.

3.3.2 Obtaining the STM's objective function

On a per-document basis, the latent variables whose posterior distributions are of interest are η_d and the N_d latent variables $z_{d,n}$. To approximate the posterior distribution using mean-field variational inference, the STM uses the same partitioning as the CTM (equation 3.7). In contrast to the CTM, the STM does not iteratively perform updates for each parameter. Instead, to speed convergence, the STM uses a semi-collapsed version of the variational inference algorithm during the E-step [97]. This is made possible by the fact that the joint probability distribution of the latent and observed variables in the STM can be collapsed by integrating out $z_{d,n}$ [67], so we can rewrite the joint distribution of the data and latent variables in terms of just w and η:

$$p(w, \eta, z) = p(w|\eta)p(\eta) \tag{3.11}$$

3.3.2.1 Finding the $q^*(\cdot)$ distributions

Wang and Blei provide a method for finding approximations to the optimal variational distributions in non-conjugate models where the distribution of the data given the latent variables is part of the exponential family [126]. Since the categorical distribution is a member of the exponential family, we know the STM fits this profile. Using (3.6) and (3.11), we can write [126]:

$$\begin{aligned} \log q^*(\eta) &\propto E_{q(z)}[\log p(w|\eta)p(\eta)] \\ &= \log p(w|\eta)p(\eta) \\ &\triangleq f(\eta) \end{aligned} \tag{3.12}$$

To isolate $q^*(\eta)$ we would need the distribution of $\exp(f(\eta))$ which cannot be normalized to integrate to one [126]. Instead, Wang and Blei propose that the variational distribution

$q^*(\cdot)$ can be approximated using a second order Taylor approximation of $f(\eta)$ around the MAP estimate of η, $\hat{\eta}$ [126]:

$$\begin{aligned} & f(\eta) \approx f(\hat{\eta}) + -\frac{1}{2}(\eta - \hat{\eta})^\intercal \nabla^2 f(\hat{\eta})(\eta - \hat{\eta}) \\ \implies & q^*(\eta) \tilde{\propto} e^{f(\hat{\eta})} e^{-\frac{1}{2}(\eta-\hat{\eta})^\intercal \nabla^2 f(\hat{\eta})(\eta-\hat{\eta})} \\ \implies & q^*(\eta) \tilde{\propto} e^{-\frac{1}{2}(\eta-\hat{\eta})^\intercal (-\nabla^2 f(\hat{\eta})^{-1})^{-1}(\eta-\hat{\eta})} \\ \implies & q^*(\eta) \tilde{=} \text{Normal}(\hat{\eta}, -\nabla^2 f(\hat{\eta})^{-1}) \end{aligned} \tag{3.13}$$

So, the optimal form of the variational distribution $q^*(\eta_d)$ can be approximated by a Gaussian. Taking on the notation from Roberts et al. [100, 97], we call its variational parameters $\lambda_d = \hat{\eta}_d$ (a vector of length $K-1$) and $\nu_d = -\nabla^2 f(\hat{\eta}_d)^{-1} = -\nabla^2[\log p(w|\hat{\eta}_d)p(\hat{\eta}_d)]^{-1}$ (which is a $(K-1)\times(K-1)$ matrix). Note that these parameters are both document-specific and depend on the MAP estimate $\hat{\eta}$.

Following Wang and Blei [126], the optimal form of the variational distribution $q^*(z)$ will have the same distributional form as $p(z|\eta)$. So, $q^*(z)$ takes the form of a categorical distribution with some variational parameter, which we will henceforth call ψ.

Importantly, by using the Laplace approximation for the variational distribution of η, we are no longer directly maximizing the ELBO but an approximation to the ELBO [97]. The full *approximate* objective function across all documents is thus [97]:

$$\begin{aligned} \text{ELBO} = & \sum^{D} E_q(\log p(\eta_d|\mu_d, \Sigma)) + \sum^{D}\sum^{N} E_q(\log p(z_{d,n}|\eta_d)) \\ & + \sum^{D}\sum^{N} E_q(\log p(w_{d,n}|z_{d,n}, \beta_{d,k=z_{d,n}})) - \sum^{D} E_q(\log q(\eta_d|\lambda_d, \nu_d)) \\ & - \sum^{D}\sum^{N} E_q(\log q(z_{d,n}|\psi_{d,n}))) \end{aligned}$$

Where:

$$E_q[\log p(\eta_d|\mu_d,\Sigma)] = E_{q(\eta)}[-\frac{K-1}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\eta_d-\mu_d)^\intercal\Sigma^{-1}(\eta_d-\mu_d)] \tag{3.14}$$

$$E_q[\log p(z_{d,n}|\eta_d)] = E_{q(z)q(\eta)}[\sum_i^K z_{d,n,i}\log(\theta_{d,i})] \tag{3.15}$$

$$E_q[\log p(w_{d,n}|z_{d,n},\beta_{d,k=z_{d,n}})] = E_{q(z)}[\sum_i^V w_{d,n,i}\log(\beta_{d,z_{d,n,i}})] \tag{3.16}$$

$$E_q[\log q(\eta_d)] = E_{q(\eta)}[-\frac{K-1}{2}\log(2\pi) - \frac{1}{2}\log|\nu_d| - \frac{1}{2}(\eta_d-\lambda_d)^\intercal\nu_d^{-1}(\eta_d-\lambda_d)] \tag{3.17}$$

$$E_q[\log q(z_{d,n})] = E_{q(z)}[\sum_{i=1}^K z_{d,n,i}\log(\psi_{d,n,i})] \tag{3.18}$$

Since we are limiting our scope to the topic prevalence model surrounding η, we will develop out only (3.14) and (3.17). Both of the expressions for $E_q[\log p(\eta_d|\mu_d,\Sigma)]$ and for $E_q[\log q(\eta_d)]$ depend on η_d. From typical matrix results, we have:

$$E_\eta(\eta-\mu)^\intercal\Sigma^{-1}(\eta-\mu) = \mathrm{Tr}(\nu_d\Sigma^{-1}) + (\lambda-\mu)^\intercal\Sigma^{-1}(\lambda-\mu)$$

$$E_{\eta_d}(\eta_d-\lambda_d)^\intercal\nu_d^{-1}(\eta_d-\lambda_d) = \mathrm{Tr}(\nu_d\nu_d^{-1}) + (\lambda_d-\lambda_d)^\intercal\nu_d^{-1}(\lambda_d-\lambda_d) = 1$$

which results in the following forms:

$$E_q[\log p(\eta_d|\mu_d,\Sigma)] = -\frac{K-1}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}[\mathrm{Tr}(\nu_d\Sigma^{-1}) + (\lambda_d-\mu_d)^\intercal\Sigma^{-1}(\lambda_d-\mu_d)] \tag{3.19}$$

$$E_q[\log q(\eta_d)] = -\frac{K-1}{2}\log(2\pi) - \frac{1}{2}\log|\nu_d| - \frac{1}{2} \tag{3.20}$$

3.3.3 E-step: Maximizing the variational distributions

Rather than iterate between maximizing $q(\eta)$ and $q(z)$ as is done for the original variational E-step for the CTM [16], the STM first finds the MAP estimate of η which maximizes the joint probability over the latent variables, and then performs a closed-form update for ψ using this joint optimum [97].

We must first obtain the MAP estimate of η ($\hat{\eta}$), as the mean and variance of the variational distribution $q(\eta)$ are both functions of $\hat{\eta}$ (see equation 3.13). Using (3.2), finding the MAP estimate of η is equivalent to maximizing the (log) joint probability over latent variables and data, (3.11). Thus, for any given document d, the MAP estimate which maximizes the joint probability $p(w, z, \eta)$ can be found by optimizing [97]:

$$\begin{aligned}\log p(w|\eta)p(\eta) &= \log\Bigg[\frac{\prod^{V}\left(\sum^{K}\beta_{v,k}e^{\eta_k}\right)^{c_v}}{\left(\sum^{K}e^{\eta_k}\right)^{N}}) + \log(\frac{1}{(2\pi)^{\frac{K}{2}}|\Sigma|^{\frac{1}{2}}}) + \exp(\frac{-1}{2}(\eta-\mu)^{\intercal}\Sigma^{-1}(\eta-\mu))\Bigg] \\ &\propto \sum^{V} c_v \log[\sum^{K}\beta_{v,k}e^{\eta_k}] - N\log\sum^{K}e^{\eta_k} - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\eta-\mu)^{\intercal}\Sigma^{-1}(\eta-\mu)\end{aligned} \tag{3.21}$$

In the STM's implementation, Roberts et al. use quasi-Newton methods to optimize this function [97].

The update for ψ (the parameter of variational distribution $q(z)$) can then be solved for in closed form. Analogously to in the CTM [16], since we are operating within the exponential family [13], we can obtain an expression for the maximum value of the variational parameter: $\psi_{d,n,k} \propto e^{\lambda_{d,n,k}}\beta_{w_{d,n,k}}$ [97]. Since $\psi_{d,n,k}$ is a proportion such that $\sum^{K}\psi_{d,n,k} = 1$, we can normalize this quantity for each topic to obtain the optimal update for the parameter of the categorical variational distribution over z:

$$\psi = \frac{e^{\lambda_{d,n,k}}\beta_{w_{d,n,k}}}{\sum_k e^{\lambda_{d,n,k}}\beta_{w_{d,n,k}}}$$

So, the optimal update for ψ which will maximize the lower bound relies on λ and β. β is held fixed in the E-step and is maximized during the M-step of each iteration of the EM algorithm (via κ and the topic content model). We have already established the optimal value for $\lambda = \hat{\eta}$ determined through quasi-Newton optimization, so this can be used to solve for the optimal ψ [97].

3.3.4 M-step: Update model parameters (μ and Σ)

With the forms of our variational approximations chosen and the optimal updates for the variational parameters determined, we can proceed to the M-step of the EM algorithm, where the remaining (non-variational) parameters (μ, Σ, κ) are maximized in order to tighten the lower bound on the marginal probability. We direct the reader to Roberts et al. [97] and [99] for discussion κ and the topic content model.

3.3.4.1 Incorporating covariates in the updates to μ

We can motivate our updates for μ in the STM using maximum likelihood. Taking the derivative of the ELBO with respect to μ_d and setting it to zero yields:

$$\begin{aligned}
\frac{\partial L(q)}{\partial \mu_d} &= \frac{\partial}{\partial \mu_d}\Big[D(K-1)\log(2\pi) + D\log|\Sigma| + \sum_i^D \mathrm{Tr}(\nu_i\Sigma^{-1}) + \sum_i^D(\lambda_i-\mu_i)^\intercal\Sigma^{-1}(\lambda_i-\mu_i)\Big] \\
&= \frac{\partial}{\partial \mu_d}\sum_i^D(\lambda_i-\mu_i)^\intercal\Sigma^{-1}(\lambda_i-\mu_i) \\
&= -2\Sigma^{-1}(\lambda_d-\mu_d) \\
0 &\overset{set}{=} -2\Sigma^{-1}(\lambda_d-\mu_d) \\
\iff \hat{\mu}_d &= \lambda_d
\end{aligned} \tag{3.22}$$

The maximum of μ_d is obtained when $\lambda_d - \mu_d = 0$. In order to incorporate covariates, the STM sets $\mu_d = \mathbf{\Gamma}^\intercal\mathbf{x}_d^\intercal$, where x_d represent the values of topic prevalence covariates

for document d [97]. As such, the goal becomes to find the optimal values of $\mathbf{\Gamma}$ which will minimize the distance between λ_d and $\mathbf{\Gamma}^\intercal \mathbf{x}_d^\intercal$. Roberts et al. do this by solving the regression equation $\lambda_k = X\gamma_k + \varepsilon$ using variational linear regression with a regularizing prior on γ_k [97, 99]. We will elaborate further on this estimation in Chapter 4.

In contrast, in its M-step, the CTM uses MLE to obtain the update for a single, global μ which maximizes the ELBO [14, 16]:

$$\begin{aligned}
\frac{\partial L(q)}{\partial \mu} &= \frac{\partial}{\partial \mu} \sum_i^D (\lambda_i - \mu)^\intercal \Sigma^{-1} (\lambda_i - \mu) \\
&= -2\Sigma^{-1} \sum_i^D (\lambda_i - \mu) \\
0 &\overset{set}{=} \sum_i^D (\lambda_i - \mu) \\
\iff \hat{\mu} &= \frac{1}{D} \sum_i^D \lambda_d
\end{aligned} \tag{3.23}$$

3.3.4.2 Maximum likelihood estimation for Σ

Similarly, we can find the optimal update for Σ using MLE. Taking the derivative of the ELBO with respect to Σ^{-1} and setting it to zero gives:

$$\begin{aligned}
\frac{\partial L(q)}{\partial \Sigma^{-1}} &= \frac{\partial}{\partial \Sigma^{-1}} \Big[D(K-1)\log(2\pi) + D\log|\Sigma| + \sum^D \mathrm{Tr}(\nu_d \Sigma^{-1}) + \sum^D (\lambda_d - \mu_d)^\intercal \Sigma^{-1} (\lambda_d - \mu_d) \Big] \\
&= -D\Sigma + \sum_d^D \nu_d + \sum_d^D (\lambda_d - \mu_d)(\lambda_d - \mu_d)^\intercal \\
0 &\overset{set}{=} -D\Sigma + \sum_d^D [\nu_d + (\lambda_d - \mu_d)(\lambda_d - \mu_d)^\intercal] \\
\iff D\Sigma &= \sum_d^D [\nu_d + (\lambda_d - \mu_d)(\lambda_d - \mu_d)^\intercal] \\
\iff \hat{\Sigma} &= \frac{1}{D} [\sum_d^D \nu_d + (\lambda_d - \mu_d)(\lambda_d - \mu_d)^\intercal]
\end{aligned} \tag{3.24}$$

So the globally shared between-topic covariance matrix is maximized by taking the mean covariance matrix formed from a combination of the per-document variational update ν_d and the squared difference of λ_d and μ_d.

In contrast, in the CTM, the maximum likelihood estimate of Σ arising in the variational update takes the form [16]:

$$\hat{\Sigma} = \frac{1}{D}\sum_{d} \nu_d \mathbf{I} + (\lambda_d - \hat{\mu})(\lambda_d - \hat{\mu})^{\intercal} \tag{3.25}$$

This difference arises because in the CTM, there is one global update to μ (the column mean across all λ_d), and the variational distribution of η is univariate with just one ν_d per document (a $K-1$ vector with one $\nu_{d,k}$ per topic). Any covariance between topics is obtained from the second term: the square of the difference between λ and μ for each document.

3.3.5 Bringing it all together: iteration

The algorithm that fits the STM iterates between updates for the variational parameters η and ψ and maximization of the model parameters Γ, Σ, κ until convergence [99]. In a typical variational inference strategy, convergence would be assessed as relative change to the ELBO per step. Since the STM is fit using a semi-collapsed variational E-step, Roberts et al. assess convergence of the algorithm in terms of the relative change to the collapsed per-document log likelihood $p(w|\eta, \beta)$ which omits z and its updates [97]:

$$L = \sum^{D}\left(\left[\sum^{V} w_{d,v}\log(\beta_{\mathbf{d},\mathbf{v}}\theta)\right] - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\lambda_d - \mu_d)^{\intercal}\Sigma^{-1}(\lambda_d - \mu_d) + \frac{1}{2}\log|\nu_d|\right) \tag{3.26}$$

3.3.6 Benefits of the STM

Relative to the LDA, the STM more often retrieves the true effect of covariates on topic proportions, and the variance of the covariate slope estimates decreases with an increased number of documents [100, 97]. Roberts et al. further provide evidence that the STM does not induce nonexistent covariate effects, through simulations with no true effect and a permutation analysis [100]. In comparison to its predecessor models, the STM has the benefit of a completely flexible covariate incorporation in both the topic prevalence model and influencing topic content [97].

Chapter 4

4 Incorporating grouping structure into the estimation of the mean topic weights

4.1 The STM's incorporation of covariates

The generative model of the STM assumes that, for each document d, the document-topic proportions θ_d are sampled from a logistic normal distribution with a document-specific mean and global covariance [97]:

$$\eta_d \sim \text{Normal}_{K-1}(\mu_d, \Sigma) \iff \theta_d \sim \text{LogisticNormal}_K(\mu_d, \Sigma) \tag{4.1}$$

Note the distinction between η_d which are "document-topic weights" ($\eta_d \in \mathbb{R}^{K-1}$) and θ_d which are "document-topic proportions" ($\sum_K \theta_d = 1$, i.e., belong to the $K-1$ simplex). For brevity, these will be shortened to "topic weights" and "topic proportions" when there is no ambiguity with the topic-word proportions, β, from the topic content model.

Recall that the goal of the M-step of the variational EM algorithm used to fit the STM is to maximize the evidence lower bound with respect to the model parameters, given the form of the variational posterior found during the E-step. In its M-step, the STM updates μ_d [97].

4.1.1 Introduction to regression updates to μ

To obtain some intuition about the STM's update to μ_d, consider the matrix $\mathcal{M}$ which is a $D \times (K-1)$ matrix with entries $\mu_{d,k}$ corresponding to the weight of topic k in document d. The columns of this matrix, μ_k, are $D \times 1$ vectors of topic weights associated with topic

k:

$$\mathcal{M} = \left[\mu_1 \quad | \quad \ldots \quad | \quad \mu_{K-1}\right]$$

Similarly, λ_k is a $D \times 1$ vector of the variational updates to topic weights for topic k across all documents. Like μ_k, $\lambda_k \in \mathbb{R}^D$. Finally, we have covariate data, $\mathbf{X}$, a $D \times P$ matrix where P is the number of columns in a model matrix of covariates assumed to affect topic prevalence, and x_d is the vector of covariate values associated with the dth document.

During the M-step, the STM obtains a $P \times 1$ vector of posterior mean coefficients, γ_k, by conducting a linear regression of the covariate information, $\mathbf{X}$, on the variational parameter update values, λ_k, independently for each topic [97, 99]. These coefficient estimates are obtained such that they minimize the columnwise (per-topic) sum of the squared differences $\sum^D(\lambda_k - \mathbf{X}\gamma_k)^2$ subject to an additional penalty on the coefficients. The mean vector of the predictive distribution from this regression ($\mathbf{X}\gamma_k$) for each topic k is thus the closest approximation to λ_k under this error criterion, given covariate information and data.

Across all topics, we thus obtain a matrix of coefficients, $\mathbf{\Gamma}$. Row d of $\mathcal{M} = \mathbf{X}\mathbf{\Gamma}$ is $\mu_d = \mathbf{\Gamma}^\intercal x_d^\intercal$, which is used as the mean of the topic prevalence distribution in (4.1). Crucially, only the predicted values in $\mathcal{M}$ are used downstream. This makes μ_k a key point of entry for incorporating additional structure and information into the topic prevalence model of the STM.

Because λ_k is different for each topic, the coefficient weights γ_k are different for each topic. However, the covariate information contained in $\mathbf{X}$ is the same for each topic. This means the same covariate information is assumed to be associated with the topic prevalence of each topic, but can affect different topics' prevalence in different ways. This formulation also means that two documents with similar levels of the covariates share similar fitted values in a given topic, which the STM presents as analogous to partial pooling [97].

4.1.2 The STM's variational linear regression

The STM implements a fast variational linear regression on λ with a ridge-like penalty on the coefficients in order to obtain the mean of the posterior predictive distribution given the per-document topic weights [99].

4.1.2.1 Regression details

For simplicity, we consider a single topic, k and drop the subscripts in our description. This linear regression with outcome λ, coefficients γ, data x, and variance parameters for the residual variance and coefficient variance respectively β^{-1} and α^{-1} is formulated as a hierarchical Bayesian model in the following manner [99]:

$$\begin{aligned} p(\lambda|\gamma,\beta) &= \text{Normal}(\gamma^\intercal x, \beta^{-1}\mathbf{I}) \\ p(\gamma|\alpha) &= \text{Normal}(0, \alpha^{-1}\mathbf{I}) \\ p(\beta) &= \text{Gamma}(c_0, d_0) \\ p(\alpha) &= \text{Gamma}(a_0, b_0) \end{aligned} \tag{4.2}$$

with possibly some hyperprior distribution over the parameters for the gamma distributions left unspecified by Roberts et al. [97, 99].

The normal distribution on γ grants some informativity in the form of shrinking the coefficients towards 0 to a degree tuned by α, as in ridge regression (which we will discuss below) [11, 57]. Importantly, in the implementation of the STM, the intercept is left out of this penalization [99].

Since the posterior distribution of γ does not have a closed form when there are hyperpriors on the variances, an approximate method must be used. In order to efficiently estimate the values of the coefficients, the STM uses a mean-field variational approximation to the posterior which takes a fully factorized form: $q(\gamma,\alpha,\beta) = q(\gamma)q(\beta)q(\alpha)$. The resulting vari-

ational distributions are a normal distribution on the coefficients with gamma distributions on the precisions [99, 11]. The parameters of the variational regression have the following updates [99]:

$$
\begin{aligned}
q^*(\gamma) &= \text{Normal}(m_N = E(\beta)V_N X^\intercal \lambda, V_N = \left[E(\alpha)\mathbf{I} + E(\beta)X^\intercal X\right]^{-1}) \\
q^*(\beta) &= \text{Gamma}(c_N = \frac{1+D}{2}, d_N = \tfrac{1}{E(\beta)_{N-1}+d_0} + \tfrac{1}{2}[\text{Tr}(V_N X X^\intercal) + (m_N^\intercal X - \lambda)^\intercal (m_N^\intercal X - \lambda)]) \\
q^*(\alpha) &= \text{Gamma}(a_N = P, b_N = \tfrac{4}{E(\alpha)_{N-1}+b_0} + \tfrac{1}{2}[m_N^\intercal m_N + \text{Tr}(V_N)])
\end{aligned}
\tag{4.3}
$$

Because the variational posterior on γ is normal, its posterior mean coincides with its posterior mode: the MAP estimates of γ are the mean of the variational posterior, m_N. Using the fact that the conditional distribution of λ and the approximate posterior distribution of γ are both normal, the mean of the posterior predictive distribution is $m_N^\intercal \mathbf{X}$ [11], and is the update for μ_d. In a frequentist sense, these represent the "fitted values" of the regression; in a more Bayesian sense, these are the most probable values of the outcome λ, given the data and the posterior distribution of the coefficients.

The variational distribution parameters are solved for iteratively using the variational updates until the change in the sum of the coefficients is less than 0.0005[1]. This variational regression is able to be implemented in base R and converges quickly even with complex models and many documents [99]. Since this regression is fit $K-1$ times per iteration of the EM algorithm (whose iteration limit is set by the user and can be in the tens or hundreds), it is important to the STM's implementation that this regression runs relatively quickly.

[1]The original implementation of the STM uses 0.0001 as the threshold value. To match the implementation of the mixSTM method discussed in this chapter, the threshold was slightly increased. Implementation settings and thresholding are discussed further in Appendix A.1.

4.1.2.2 Regularization

The hierarchical model described in (4.2) implicitly penalizes large coefficients by drawing them towards zero. Penalization of regression coefficients, also called "regularization" is a model fitting technique that is particularly useful when overfitting a model to the current dataset is a concern [11].

In the linear regression case, regularization imposes a penalty on the objective function to be minimized, transforming it from a sum of square residuals, $\sum(y - XB)^2$ to a sum of square residuals plus a weighted term over the coefficients

$$\sum(y - X\beta)^2 + \rho \sum |\beta|^q \tag{4.4}$$

The second term, $\rho \sum |\beta|^q$ is a regularization term, where ρ is a tuning parameter or regularization coefficient (which induces more shrinkage towards zero for higher values) and q dictates the form of the regularization [11]. When $q = 2$, this regularization strategy is known as "ridge regression" [57]. The question naturally arises of how to choose ρ, the strength of the regularization. While ridge regression will not induce sparsity by pulling coefficients to exactly zero (unlike the $q = 1$ case, called the "Lasso" [119]), the strength of the regularization still determines how sensitive the present regression will be to the current dataset relative to if the data were generated multiple times [11]. Methods for choosing an appropriate ρ include using cross-validation, empirical Bayes, or information criteria [122, 53].

As noted by Hoerl and Kennard [57], the ridge regression also arises from a Bayesian hierarchical formulation of a regression with a normal prior on the coefficients with mean zero and some variance parameter. For some choices of other priors, the posterior over the coefficients is then also normal, and the log posterior takes the following recognizable form: $\sigma^2 \sum(y - \beta^\intercal X)^2 + \alpha\beta^\intercal\beta$, where σ^2 is the noise variance and α is the precision of the normal

prior on the coefficients [11]. Then, the tuning parameter is determined automatically as $\rho = \frac{\alpha}{\sigma^2}$ [11], which is the value for which the expected generalized cross-validation error is minimized [49].

In the STM, the regression on topic weights is used to find predicted values at each iteration of the EM algorithm making regularization an important part of this model. In early iterations, overfitting the model to preliminary values of λ could pull the M-step to maximize an effect that is only a noisy artifact of the initialization, not of the data. In later iterations, overfitting is still a concern where there are few documents at a particular level of a covariate or in a particular group, which might result in extreme values of μ. In addition, ridge-like regressions can reduce the identifiability issues that arise in cases of multicollinearity between covariates [3]. When the independent variables in a regression are linearly dependent or highly correlated, the variance of coefficient estimates becomes large and subsequent inferences may become unstable. Finally, regularization allows the model to be fit even when it is overparameterized, that is, has a large number of covariates relative to the number of documents (e.g., does not meet the 10 event observations per covariate minimum recommended in regression literature)[2]. In the exploratory setting of topic models, it is desirable to have the flexibility to incorporate many covariates of potential interest, even if they could be collinear or represent relatively small groups.

The STM also implements another option for this regression: the Lasso using R's `glmnet` package [45]. In cases where there are many potentially sparse covariates, the STM suggests moving to an L1-penalty model instead [99]; using `glmnet` also permits tuning for an estimation between the L2- and L1-penalization ("elastic-net") [45]. The authors of the STM note that using the Lasso is less computationally efficient and not recommended unless the covariates in the prevalence model are highly sparse (for example, with hundreds of categories of a factor variable) [99]. The regularization coefficient must be tuned by

[2]Note that we still expect the overall number of covariates to be lower than the number of documents, here and throughout.

the user. In practice, although the `glmnet` Lasso will successfully fit an STM, it rarely results in estimated topic proportions closer to the true values in simulation than the ridge regression, especially if all included covariates do have an effect on topic prevalence.

4.2 Clustered document models for μ

We turn our attention now to the question of performing inference when documents are grouped, such as in the case of focus groups where the sessions or participants form natural groupings of documents. We expect the documents within a group to share more similarity with each other than with documents in another group, and that the relationship between topic prevalence and covariates may change depending on the group.

4.2.1 Motivating a change to the estimation of μ_k

The penalized linear regression used in the M-step of the STM can account for grouping variables and interactions with the grouping variables using categorical dummy-coded variables which will be estimated and penalized like any other covariate in the model. However, positing a hierarchical regression which separates group-varying and non-group-varying effects may more accurately represent our beliefs about the data structure and result in models whose posterior predictive estimates more closely resemble the relationship between the metadata and topic prevalence. Note that we are not suggesting a change to the distributions that make up the ELBO, just to how μ, the mean of the model distribution for the topic proportions, is maximized. This allows much of the inference to continue as usual.

Some desirable properties of a hierarchical regression for μ_k include:

- Ability to handle a large number of groups: Even though it can account for many covariates, the STM's regression may struggle to converge when there are many parameters to estimate relative to the number of observations; for example in the case

of multiple group-specific slopes and/or when there is a large number of groups. The automatic prompt when convergence fails encourages the user to try the L1-penalized model instead, but this may not be desirable. For one, if each grouping covariate included has an effect on the topic prevalence (as is assumed in the case of grouped documents) and none will be shrunk to zero via this penalization, the introduced bias means the overall fit will be less accurate. For another, the Lasso implemented in the STM requires the choice of a tuning parameter, which is challenging in the context of a regression that is conducted many times with different outcome values. Instead, one could opt for a mixed model which estimates fewer parameters for the group effects (variance components, rather than coefficients) and thus avoids the need to perform L1-penalization on the group coefficients.

- Separate penalization for group effects: By setting a separate prior on any group effects, penalization can happen differently for the group-associated and non-grouped effects. In the original STM implementation, for all coefficients including group intercepts or group-by-slope interactions, the level of penalization is tuned by the single parameter α. Separating the tuning of penalization into two or more parameters allows more flexbility in the model.

- Interdependence of covariates: Although penalization can help when covariates are multicollinear or dependent, mixed models provide an explicitly modelled covariance for the group-varying predictors. This may more closely represent our beliefs about the complex relationships between variables and topic prevalence. Though ridge regression can handle some amount of collinearity between covariates, the frequentist Lasso is not as adept and may result excessively shrunk coefficients or the selection of just one of several collinear variables [122]. This is particularly troubling for the cases when several group-varying covariates are included for a large number of groups.

- Uneven group sizes: Naturally, one level of any categorical variable has to be omitted in a fixed effects model and is treated as the "reference", but this can cause issues when that group is particularly small or extreme in its observations. Mixed models account for this by partial pooling, i.e., borrowing information from the global mean to weight group estimates. Although the STM's regression can do some pooling thanks to the regularization approach, the intercept (and thus reference group of a grouping variable) is unpenalized, which may still result in extreme values for the intercept if it corresponds to a small group.

4.2.2 Potential regression models

There are several properties that a regression model for grouped documents would need to suit the purposes of our mixSTM. Like the variational regression in the STM, it has to be able to quickly converge and obtain the posterior distribution of coefficients and posterior predictive distribution, as well as incorporate some degree of penalization on the coefficients (ideally without need for manual tuning). Furthermore, it should model a correlation structure between group-varying effects and involve some division of the grouped "random" and non-grouped "fixed" effects.

Linear mixed models as fit by `lme4` [6] or `nlme` [93, 94] in R fulfill these criteria except regularization. This makes them vulnerable to overfitting, to unstable estimates when multicollinearity between covariates is an issue, and, because there is no pooling between the fixed effect coefficients, require more data to have a sufficient degrees of freedom for estimating coefficients reliably when many covariates or levels of a categorical covariate are of interest[3].

The frequentist approach to penalization exists in many forms. L1-penalties for mixed model selection have been proposed by Foster et al. [44], Ni et al. [90], Schelldorfer et al.

[3] In preliminary implementations, a version of mixSTM using nlme during the M-step sometimes failed to initialize during fit.

[105], and Groll and Tutz [53], for example, and Wang proposed an adaptive Lasso approach [127]. To account for multicollinearity among fixed effect coefficients, Özkale and Can provided the derivation for a ridge estimator for fixed effect coefficients [92], modelled after results from Liu and Hu [78] and Eliot et al. [39], and provided a recommendation for choosing the tuning parameter when cross-validation is deemed computationally intensive. Kuran and Özkale [71] proposed a ridge-like solution to multicollinearity through stochastic linear restrictions (which have been expanded on to, for example, account for measurement error [48]) but this approach requires the incorporation of prior information from similar problems. The disadvantage to each of these approaches is the selection of an appropriate tuning parameter, particularly as the regression outcome changes on a per-topic and per-iteration basis, so a "one-size fits all" approach may be lacking.

In line with the original STM's regression, a hierarchical Bayesian model thus presents interesting avenues. A Bayesian regression formulation can place a hyperprior on the tuning parameter to control the expected amount of regularization in a way that can adapt to the data. The properties of shrinkage-inducing priors for regression in the Bayesian context have been extensively studied [122, 115, 68]. For mixed models, the exact posterior over the coefficients may not have a closed form, but implementing a mixed model with shrinkage-inducing priors is easily done using the flexible frameworks of `brms` [21] or `Rstan` [110] in R. However, MCMC methods like those used to fit the models in `brms` and `Rstan` are slow, especially for large datasets, and typically require assessment of the convergence of chains, which limits their utility in the context of an iterated regression in the STM. Yang et al. developed a fast Gibbs sampling approach to jointly select fixed and random effects in a model [131], based off the popular parameterization of Chen and Dunson [31], but we have no interest in selecting subsets of the random effects. Some authors have turned to variational Bayes to fit penalized mixed models: Yi and Tang proposed an approach under spike-and-slab priors [132], while Degani et al. proposed an algorithm with global-local priors on the fixed effects [34]. The coefficients and random effect pa-

rameters when these variational inference strategies have closed form updates and thus can be obtained rapidly. Degani et al. further streamline the updates by taking advantage of the highly sparse structure of the matrices involved [34].

4.3 Penalized mixed models fit using streamlined variational Bayes

Degani et al. propose a method for fitting a two- or three-level penalized mixed models efficiently using streamlined variational Bayes [34]. Although their model and algorithm are motivated by a desire to perform selection on fixed effects in linear mixed models, we propose using this regression in the mixSTM to incorporate both random and fixed effects. As we shall see, this model fulfills all the criteria outlined in the previous section for a model to update μ_k.

4.3.1 Global-local priors

The penalizing priors used by Degani et al. on the fixed effect coefficients are from the class of "global-local" shrinkage priors [34]. These priors are continuous on their domain with a high density near zero, resulting in varying amounts of penalization depending on the specific distributions chosen and their hyperparameters [34, 115]. Degani et al. choose to focus on global-local priors for shrinkage because they tend to perform well in high dimensions and because they can be rewritten as scale-mixtures of normals which can lead to conjugate, closed-form updates [34]. Notably, these global-local priors do not automatically perform selection of variables by shrinking coefficients to exactly zero, which is ideal for the mixSTM implementation case where selecting a subset of covariates is not the goal.

For an example, consider setting a Horseshoe prior [24] over a subset of $h = 1, ..., H$

coefficients β_h subject to selection: $\beta_h|\tau \sim \text{Horseshoe}(0, \tau)$. The Horseshoe distribution, as a member of the global-local family, can be rewritten (per [34] Table 1):

$$
\begin{aligned}
p(\beta_h|\zeta_h, \tau^2) &= N(0, \tau^2/\zeta_h) \\
p(\zeta_h|a_{\zeta_h}) &= \text{Gamma}(\frac{1}{2}, a_{\zeta_h}) \\
p(\tau^2|a_{\tau^2}) &= \text{Gamma}(1, 1/a_{\tau^2})
\end{aligned}
$$

where the hyperparameters a_{τ^2} and $a_{\zeta_h^2}$ also have hyperpriors. Relative to other shrinkage priors in the literature, the Horseshoe distribution has thicker tails allowing little penalization for coefficients that are truly large, while strongly shrinking others towards zero [24, 122].

In the scale mixture representation, it becomes clear why these priors are called "global-local": they have two parameters controlling the penalization for each coefficient. One, τ^2, is "global" because it is shared across all penalized coefficients and controls a general amount of shrinkage; the second parameter, ζ_h, is "local" because it tunes the amount of shrinkage on each coefficient individually [34, 115]. Many members of this family of shrinkage priors exist [115]. Besides the Horseshoe, Degani et al. also implement the Laplace and Negative Exponential Gamma (NEG) priors on coefficients in their regression [34]. Without loss of generality, we will use the Horseshoe prior going forward, both for its ability to keep large coefficients large and to avoid specifying the additional parameter necessary for the NEG prior.

4.3.2 Mean-field variational updates for γ and u

For a two-level linear mixed model with an outcome λ, p^S coefficients γ^S subject to penalization and $q = p^R$ coefficients γ^R associated with the random effects, N observations of data $X = [X^R|X^S]$; $i = 1, ..., M$ groups each with q random effect estimates u_i associated with the random effects design matrix $Z_i = X_i^R$, and a residual variance β, Degani et al.

write the following conditional distribution for the outcome [34]:

$$p(\lambda|\gamma^S, \gamma^R, u, \beta) = \text{Normal}(X^S\gamma^S + X^R\gamma^R + Zu, \beta\mathbf{I}) \tag{4.5}$$

Each of the regression parameters, γ^R, γ^S, u, and β are assigned prior distributions according to Equation 17 of [34]. In brief:

$$\begin{aligned}
\beta|a &\sim \text{Inverse-}\chi^2(\nu_\beta, 1/a_\beta) \\
\begin{bmatrix}\gamma^R \\ \gamma^S\end{bmatrix}|\zeta, \tau^2 &\sim N\left(\begin{bmatrix}\mu_{\gamma^R} \\ 0\end{bmatrix}, \begin{bmatrix}\Sigma_{\gamma^R} & 0 \\ 0 & \tau^2\text{diag}(\zeta)^{-1}\end{bmatrix}\right) \\
u_i|\mathbf{\Sigma} &\sim N(0, \mathbf{\Sigma}) \\
u|G &\sim N(0, G = \mathbf{I}_m \otimes \mathbf{\Sigma}) \qquad \text{where } u = \text{stack}(u_i)
\end{aligned}$$

where the hyperparameters a_β, ν_β, τ^2, ζ, and $\mathbf{\Sigma}$ have hyperpriors specified in [34] according to the choice of global-local prior and random effects structure (two- or three-level).

Degani et al. approximate the posterior distribution over all latent variables using the mean-field restriction. Importantly, the resulting joint variational posterior over the regression coefficients γ and predicted random effects u takes the form of a normal distribution [34]:

$$q^*(\gamma, u) \sim \text{Normal}(\mu_{q(\gamma,u)}, \Sigma_{q(\gamma,u)}) \tag{4.6}$$

where variational updates for $\mu_{q(\gamma,u)}$ and $\Sigma_{q(\gamma,u)}$ are, as presented in Equation 21 of [34]:

$$\mu_{q(\gamma,u)} = \begin{bmatrix}\gamma^R \\ \gamma^S \\ u\end{bmatrix} \leftarrow \Sigma_{q(\gamma,u)}\left(\mu_{q(\beta^{-1})}\begin{bmatrix}X & Z\end{bmatrix}^\intercal \lambda + \begin{bmatrix}\Sigma_R^{-1}\mu_{\gamma^R} \\ 0 \\ 0\end{bmatrix}\right) \tag{4.7}$$

$$\Sigma_{q(\gamma,u)} \leftarrow \left(\mu_{q(\beta^{-1})} \begin{bmatrix} X & Z \end{bmatrix}^{\intercal} \begin{bmatrix} X & Z \end{bmatrix} + \begin{bmatrix} \Sigma_R^{-1} & 0 & 0 \\ 0 & \mu_{q(\alpha^{-1})}\text{diag}(\mu_{q(\zeta)}) & 0 \\ 0 & 0 & E(G^{-1}) \end{bmatrix} \right. \tag{4.8}$$

The (approximate) posterior predictive distribution is thus found analogously to in the STM, as it is once again a convolution of normals. The mode of the posterior predictive distribution is:

$$\mu_k = \begin{bmatrix} X & Z \end{bmatrix} \mu_{q(\gamma,u)} = X^R \gamma^R + X^S \gamma^S + Zu$$

Although closed-form updates to the variational parameters arise from the mean-field variational approximation, their computation involves the inversion of several large, sparse matrices (see equation 4.8). Inverting these matrices is time-consuming, so Degani et al. opt to instead adapt Nolan et al.'s streamlined updating procedure for mixed models fit with variational inference to the case with non-normal priors on the coefficients [34, 91]. Since each EM iteration fits a regression $K - 1$ times, the relative rapidity of Degani et al.'s implementation was among the most important considerations for our integration of a mixed effects regression into the mixSTM.

4.3.3 Comparison to the STM's variational regression

To understand how differences arise between the fit using Degani et al.'s variational regression in the mixSTM and the original variational regression in the STM, we can contrast the variational posterior updates to the mean and variance for the covariate vector as seen in (4.3), (4.7), and (4.8).

First, consider the penalized fixed effect coefficients represented by γ in the STM's model and by γ^S in the mixSTM's model. In both cases, the mean update can be written as

$$E(\beta)\mathcal{V}X^{\intercal}\lambda$$

which is a product of the expectation of the error precision $E(\beta)$, the product of the observations and the outcome $\mathbf{X}^\intercal\lambda$, and the variational update to the variance, $\mathcal{V}$. This variance update takes the form

$$\mathcal{V} = V_N = \left[E(\beta)\mathbf{X}^\intercal\mathbf{X} + E(\alpha)I\right]^{-1}$$

for the single variance-covariance matrix V_N in the STM's regression. In the mixed effects regression we have implemented in the mixSTM, the corresponding update to the submatrix of $\Sigma_{q(\gamma,u)}$ associated with the penalized coefficients is:

$$\mathcal{V} = \left[E(\beta)\mathbf{X}^\intercal\mathbf{X} + E(\alpha)\text{diag}(E(\zeta_h))\right]^{-1}$$

The difference between the updates is in the second term of the expression for $\mathcal{V}$. While V_N is a diagonal matrix with entries corresponding to the expectation of the coefficient precision, the global-local prior over the penalized coefficients results in an additional term in the update to the submatrix $\Sigma_{q(\gamma,u)}$: the per-covariate expectation of the local variance parameter $E(\zeta_h)$. This allows the mixed effects model to regularize these fixed effect coefficients on a per-coefficient basis.

Similarly, u in the mixSTM has the same update as γ in the STM for the mean of the variational posterior, but differs in the specification of the variational covariance matrix. In a two-level model like the cases considered here, the submatrix of $\Sigma_{q(\gamma,u)}$ associated with the random effects takes the form

$$E_q(G^{-1}) = I_m \otimes M_{q(\Sigma^{-1})}$$

where $M_{q(\Sigma^{-1})}$ is a positive definite $q \times q$ matrix initialized as an identity matrix but able to take on off-diagonal values during updates [34, 79] and $\otimes$ is the Kronecker product. Importantly, this means that (unlike in the STM's regression which only has a diagonal matrix at this step), Degani et al.'s regression explicitly models a covariance between random effect

terms in each group.

Finally, one can contrast the updates for γ^R, the fixed coefficients associated with the random effect components, with the updates for γ in the fixed effects regression. Rather than a hyperparameter tuning the amount of penalization on γ^R, a hyperparameter matrix (Σ_R) is user-specified for the second term in the variational posterior variance. In our implementation, this matrix is chosen to be arbitrarily non-informative, with large diagonal values and no covariance information, although other choices could be made.[4] Also, the mean of the variational posterior over γ_R has an additional term: $\Sigma_R^{-1}\mu_{\gamma^R}$ (i.e., the unique solution to the system $\Sigma_R x = \mu_{\gamma^R}$). In our implementation, where coefficients are assumed to be centred, $\mu_{\gamma^R} = 0$ and thus $x = 0$.

4.3.4 The mixSTM's generative model with a novel mean update

We can now rewrite the generative model of the mixSTM, given our proposed regression implementation for the mean topic weights:

$$\begin{aligned}
(\gamma_k, u_k) &\sim \text{Normal}(\mu_{q(\gamma_k,u_k)}, \Sigma_{q(\gamma_k,u_k)}) \\
\eta_d &\sim \text{Normal}_{K-1}(\mathbf{\Gamma}' x_d' + \mathbf{U}' z_d', \Sigma) \qquad \text{for } d = 1, ..., D \\
z_{d,n} &\sim \text{Categorical}_K(\theta_d) \qquad \text{for } n = 1, ..., N_d \\
w_{d,n} &\sim \text{Categorical}_V(\mathbf{B} z_{d,n}) \qquad \text{for } n = 1, ..., N_d
\end{aligned} \tag{4.9}$$

where x_d is the row for document d of the matrix $X = \begin{bmatrix} X^R & X^S \end{bmatrix}$; z_d is a row from the matrix $\mathbf{Z}$; and $\mu_{q(\gamma_k,u_k)}$ and $\Sigma_{q(\gamma_k,u_k)}$ are defined from the variational updates in equation (4.7) and (4.8). As before, $z_{d,n}$ are the topic assignments of word indices and $w_{d,n}$ the words in the document.

[4] See Appendix A.1.

4.4 Simulation studies for the mixSTM's mean update

We performed a number of simulations to assess the performance of the novel updates to μ which allow for grouped covariate effects in an STM. As noted, the original and new proposal for the regression are similar in their structure and assumptions with respect to the distributional forms of the variational posterior. We hypothesized that the difference between the fit of these two models would be small in simple cases such as when there is only a group intercept (so there is no benefit to being able to model a correlation between random slopes and intercepts), when the group size is relatively large, and when there are few covariates (so overparameterization is less of a concern).

4.4.1 Simulation methods

For each simulation, we generated 1000 documents with a mean of 150 words from a 600-word vocabulary using 5 topics with 10 continuous covariates and some grouping structure.

Covariates subject to selection were sampled from a multivariate normal distribution with mean 0 and covariance matrix sampled from a Wishart($\mathbf{I}$) distribution to introduce arbitrary correlation in the fixed effects. The random coefficients (intercepts or slopes) were simulated from a $\mathrm{N}(0, \Sigma)$ distribution on a per-topic basis, where the variances of group effects were set to 1 and correlation between random effects to 0.1 or 0.95. In the case of random slopes, covariates associated with the random effects were generated from independent $\mathrm{N}(0, 1)$ distributions. The per-topic mean of topic weights was calculated as a linear combination of the generated data and covariates: $\mu = X\beta + Zu$. For each document, θ_d was sampled from a logistic normal distribution with mean μ_d and covariance $0.25\mathbf{I}$. In each corpus, β was simulated from a $\mathrm{Dirichlet}_{600}(0.25)$ distribution. For each word in each document, the topic assignment $z_{n,d}$ was sampled with probability θ_d from a categorical distribution, and the word label $w_{n,d}$ was sampled from a categorical distribution with mean

probability $\mathbf{B}z$.

We developed six primary simulation scenarios:

Set A: The true model involves only a groupwise random intercept. We varied the number of equal-sized groups to compare how group size affected the fit in a simple model.

A1. 20 groups of documents.

A2. 50 groups of documents.

Set B: The true model involves a groupwise random intercept and three continuous random slopes with 50 equal-sized groups of documents. We varied the correlation between the random effects to examine how correlation of effects affected the fit in more complex models.

B1. Correlation of 0.1.

B2. Correlation of 0.95.

Set C: The true model is the same as simulation B1 but we specified a model that did not match the true generation mechanism.

C1. Addition of 20 fixed continuous covariates with no true effect on the topic proportions.

C2. Incorrect specification of the random effects as only a random intercept model.

These scenarios are relatively ideal in that they have strong covariate effects, large documents, and relatively little variance for the topic proportions. We also explored two situations which more closely resembled real-life scenarios. In scenario R1, we considered a case like B1 but where the average length of a document was only 30 words, $\Sigma_\theta = 0.5\mathbf{I}$, and the variance of the random slopes and intercepts was also 0.5. This represents a scenario where there is a lot of residual variance relative to the between-group variance, and also where documents are short. In scenario R2, we reduced the topic weight variance back to

0.25 but kept the short documents and lower between-group variability of R1.

Although generating "realistic" representations of text is difficult, the parameters here were chosen to provide sufficient data to explore how the original STM and the mixSTM compare on some quantitative measures of fit. Although 1000 documents in a corpus may seem small for a topic model, in practice this is sufficient to obtain good estimates and also represents a corpus size that could reasonably be obtained from qualitative research. In preliminary simulations, even 200 documents was sufficient. The vocabulary size of 600 is lower than might be expected realistically in a conversation among adults, but provides a sufficient number of unique words to be distributed across 5 topics. Finally, not all qualitatively collected corpora may have a mean of 150 words per document, so we relax this and other assumptions in our final two scenarios (R1 and R2).

To initialize the topic model, we used the "Spectral" option as recommended by Roberts et al. [99].[5] We specified the γ estimation to use the variational linear regression shown in (4.2) for the STM and to use a Horseshoe prior for the mixed model implementation in the mixSTM. Preliminary results showed minimal difference when using the Horseshoe prior compared with the Laplace prior and avoided the need to additionally specify the NEG parameter λ.

4.4.1.1 Simulation comparisons

To compare simulation results, we examined how close each model's MAP estimates of the topic proportions were to the true document-topic proportions using two distances across all documents and topics: the L1 distance $|\hat{\theta}_d - \theta_d|$ and the KL divergence $\sum^K \hat{\theta}_d \times \log_2 \frac{\hat{\theta}_d}{\theta_d}$. We also calculated the difference in L1 distance ($|\hat{\theta}_{d,mix} - \theta_d| - |\hat{\theta}_{d,stm} - \theta_d|$) and in KL divergence ($\text{KL}(\hat{\theta}_{d,mix}, \theta_d) - \text{KL}(\hat{\theta}_{d,stm}, \theta_d)$) to the true topic proportions between the mixSTM

[5]The "Spectral" option in the R package is the default and avoids the need to re-initialize the model at several starting values as would be recommended when using another initialization method such as random starting points or first running a short LDA model, since it is deterministic [99].

and STM. Although the L1 distance is more intuitively interpretable, the KL divergence uses a ratio rather than difference and thus captures relative distance in proportions close to the extremes better than the L1 distance. We present the median (Q1-Q3) distance obtained by the mixSTM and the median (Q1-Q3) difference between distances obtained by the mixSTM and STM across all 100 simulations (500,000 topic proportions per scenario for the L1 distance, or 100,000 documents per scenario for KL divergence). We calculated the proportion of times out of 100 simulations for each scenario that the mixSTM had a smaller median distance to the true topic proportions than the STM (representing a median closer fit to the truth for that simulation). Finally, we present the overall percentage of differences in distance in which the mixSTM had a smaller distance to the true document-topic proportions.

One challenge that arises when comparing the results of topic models is that the number and order assigned to the topics is not meaningful: the same topic may be discovered as "topic 1" in one simulation and "topic 2" in another. In order to calculate and compare the distance to the true topics, we aligned the topics by jointly maximizing the dot product of the estimated $\hat{\theta}_k$ from the original STM model with the true θ_k across all topics, using `lpSolve` in R.

We also considered the 5-fold cross-validated held-out likelihood for a subset of 10 of the 100 simulated corpora for two scenarios, using a modified version of the `eval.heldout` function from the `stm` package [99]. For each of the ten corpora, all words of five randomly chosen subsets of 200 documents (20%) were sequentially held out and a mixSTM with Horseshoe prior, a STM with the pooled prior, and a CTM (no covariates) were trained on the remaining 800 documents. The performance of the training set posterior on the test documents was assessed by calculating the average per-word likelihood per test document. Held-out likelihoods are presented using mean and standard deviation of per-document likelihood and difference in per-document likelihoods on the held-out documents across all

10 corpora.

Finally, we used the provided `exclusivity` [10] and `semanticCoherence` [82] functions in the `stm` package to compare the topic models on measures designed to capture topic quality (see section 2.1.2). We set $M = 10$ for both measures (to calculate scores using the top 10 words of each topic), and kept the weight parameter for the FREX score for exclusivity at its default value of 0.7 (which prioritizes exclusivity but also takes word frequency into account) [99].

4.4.2 Simulation results

4.4.2.1 Retrieval of true topic proportions θ

Table 4.1 presents the results of the eight simulations comparing the mixSTM (which uses a mixed effects model to estimate μ) to the STM (which uses a fixed effects model). The fit was good across all simulations, with both the mixSTM and STM returning small median L1 distances and KL divergences, particularly when the data was generated with many covariates with strong effects (as in the case of the model which has random slopes in addition to the random intercept) and when there were a larger mean number of words per document. The overall median difference between the mixSTM and STM in L1 distance and KL divergence to the true topic proportions was negative (in favour of the mixSTM) for each model except a random intercept model with few groups or a random slope model incorrectly specified as a random intercept model. However, the median difference was universally (for both negative and positive differences) close to zero: the largest median difference in L1 divergences occurred in scenario B2 and was -0.000072 (Q1=-0.0035, Q3=0.0027), or about 0.5% of the median L1 divergence for the mixSTM. When considering the KL divergence, the largest between-model median difference was -0.00074 (Q1=-0.0051, Q3=0.0026), or about 3% of the median KL divergence between the mixSTM and the true values.

The reason for the closeness to zero of the median difference is two-fold: first, the majority of differences to the true document-topic proportions are small, and second, the number of distances to the true document-topic proportions which are smaller for the mixSTM than the STM is around 50% in all cases. In Table 4.1, we can see that for scenarios B1 and B2, 53% and 52% of L1 distances to the true document-topic proportions were smaller for the mixSTM than the STM: this represents 14,134 and 12,108 document-topic proportions above the equality threshold, respectively. Also, even when looking in terms of KL divergence, the mixSTM improved the fit in 54% and 57% of document-topic proportions for scenarios B1 and B2 respectively.

Acknowledging that there is variance in the fit between simulations within each scenario, it

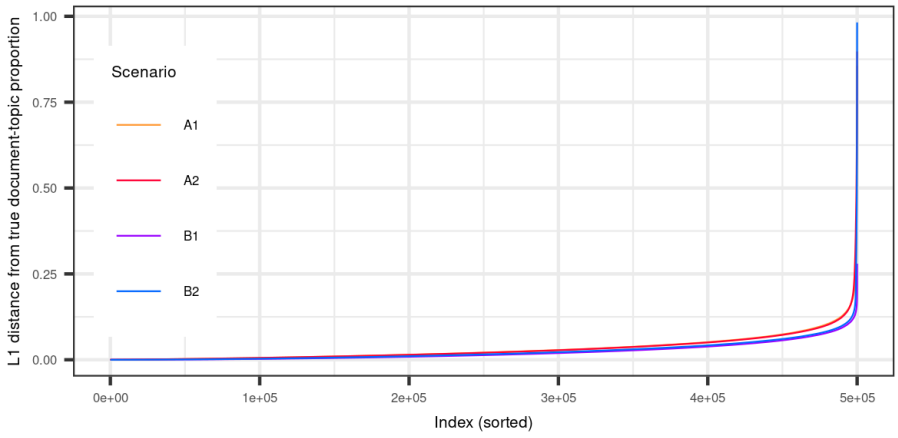
Table 4.1: Comparison of mixSTM and STM using the difference of L1 distances and KL divergences to true topic proportions.

	Scn.	mSTM Median (Q1, Q3)	Overall median (Q1, Q3) diff mSTM-STM ($\times 10^{-4}$)	%sims $\tilde{d}$<0	%diffs d<0
L1 Distance	A1	.021 (.0075, .044)	.39 (-6.9, 9.2)	34	46.9
	A2	.020 (.0074, .043)	-.17 (-9.0, 8.9)	61	51.7
	B1	.013 (.0037, .032)	-.18 (-7.8, 6.7)	86	52.8
	B2	.016 (.0046, .036)	-.72 (-35, 27)	91	52.4
	C1	.013 (.0037, .033)	-.18 (-15, 13)	89	52.1
	C2	.017 (.0065, .037)	.001 (-4.0, 3.9)	52	50.0
	R1	.033 (.0098, .078)	-.10 (-36, 35)	61	50.4
	R2	.031 (.0090, .075)	-.39 (-30, 25)	73	51.8
KL Divergence	A1	.033 (.017, .058)	1.9 (-7.4, 15)	30	42.8
	A2	.032 (.016, .056)	-.70 (-12, 11)	59	52.5
	B1	.021 (.010, .037)	-1.0 (-11, 7.8)	87	54.2
	B2	.023 (.011, .040)	-7.4 (-51, 26)	97	56.9
	C1	.021 (.011, .038)	-1.9 (-21, 15)	94	53.9
	C2	.038 (.021, .068)	-.70 (-8.3, 7.6)	59	50.4
	R1	.095 (.045, .180)	-1.1 (-78, 78)	62	50.7
	R2	.084 (.039, .162)	-3.6 (-61, 51)	71	52.9

Comparison of mixSTM and STM using the difference of L1 distances and KL divergences to true topic proportions across 8 scenarios. The median (Q1, Q3) L1 distance and KL divergence from the true document-topic proportions are presented for both topic models. The median difference of L1 distances across all simulations, documents and topics, and the median (Q1, Q3) difference of KL divergences across all documents and simulations is presented, where a negative value is in favour of the mixSTM's fit. The percentage of the 100 simulations ("% sims") in which the median ($\tilde{d}$) L1 distance or KL divergence between estimates and the truth was smaller (better) for the mixSTM than the STM, as well as the percentage of all differences ("% diffs") in L1 distance and KL divergence for which the mixSTM had a smaller (better) distance than the STM are presented.

becomes interesting to contrast the fit of the mixSTM and STM on a per-simulation basis. When the true model included random slopes, the mixSTM retrieved a closer median L1 distance to the true values of θ in 86% of simulations when the correlations between random effects were small (0.1) and 91% of simulations when they were large (0.95). When we added many fixed effect covariates with no effect on the outcome, the mixSTM retrieved a median L1 distance closer to the true values in 89% of cases, but only 52% of cases when the model was incorrectly specified as an intercept-only model rather than the true random slopes model. This echoes the results when the true model was groupwise intercept only: the L1 distance to the truth was infrequently better for the mixSTM when fewer groups were in the model (34% of simulations), and still not as frequently when there were 50 groups (61% of simulations). Relative to a more ideal scenario with high between-group variability and long documents, the mixSTM's improvement over the STM in finding the true values of the topic proportions is attenuated when documents are shorter: in 61% of simulations, the median L1 distance to the true values across all topics was better for the mixSTM than the STM, even with high residual variation in the topic proportions. When the residual variation was smaller but the between-group variability still low, the mixSTM performed better in 73% of cases with shorter documents. Similar results are visible for the KL divergence.

Figure 4.1: Sorted L1 distances between estimated and true topic proportions for 100 simulations of scenarios A1, A2, B1, and B2.

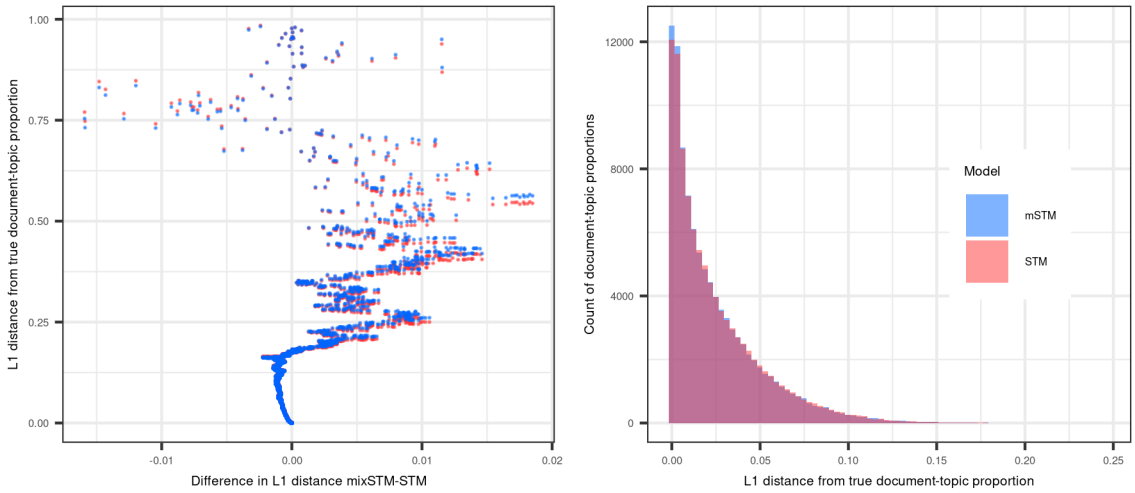


The distribution of L1 distances (and KL divergences) is highly skewed towards 0, with the large majority of estimated document-topic proportions near the true values for both the STM and mixSTM. Figure 4.1 shows the L1 distances from the true topic proportions sorted in ascending order for the first four scenarios (A1-2, B1-2) for the mixSTM. At the lowest of those presented, in simulation B1, the proportion of L1 distances to the true topic proportions above 0.1 was 1.49% for the STM and 1.48% for the mixSTM, or approximately 15 document-topic proportions of the 1000 documents for each topic in each simulation. Similarly, the L1 distance to the true topic proportions was greater than 0.1 in scenario B2 for only 2.04% of document-topic proportions in the STM and 1.94% of document-topic proportions for the mixSTM; in scenario A1, this proportion was 4.11% for the STM and 4.15% for the mixSTM; and in scenario A2, 4.61% for the STM and 4.51% for the mixSTM. The relatively higher amount of document-topic proportions above the arbitrary threshold of an L1 distance of 0.1 for the intercept-only scenarios makes sense: there is less covariate information available for each document.

It is especially for these smallest difference values (closest to the truth) that the mixSTM showed an improvement relative to the STM. We explored this using several plots of the L1 distance and KL divergence for scenario B2 (the one in which mixSTM most consistently outperformed the STM across simulations in its retrieval of the true topic proportions).

Figure 4.2(A) plots the sorted L1 distances for the mixSTM and STM (blue and red, respectively) against the difference between these vectors of sorted values (not necessarily against the difference within a given document). In contrasting the overall spread of the L1 distances across documents and simulations, we see that the mixSTM's smallest distances are nearly always closer to zero than the STM's smallest distances. Figure 4.2(B) is a histogram of the document-topic proportions from the mixSTM and STM for the first 20

Figure 4.2: (A): Sorted L1 distances between MAP estimates and true document-topic proportions versus the difference in L1 distances (mixSTM-STM) for 100 simulations of scenario B2. (B): Histogram of L1 distances between MAP estimates and true document-topic proportions for the first 20 simulations of scenario B2.

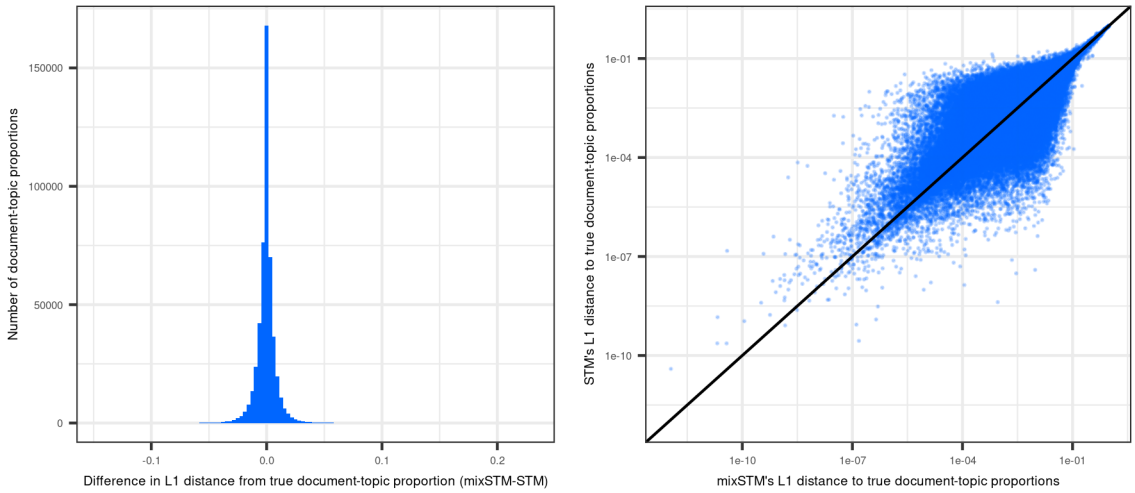


simulations[6] (100,000 document-topic proportions). Like in Figure 4.2(A), the difference in document-topic proportions is most visible for the smallest values, where the mixSTM has more mass than the STM.

Figure 4.3(A) shows the difference in L1 distances to the true θ for each document-topic proportion as a histogram. The distribution is skewed: the majority of the mass falls slightly in favour of the mixSTM (just below zero), showing that for the majority of cases, particularly when the difference is small, the mixSTM obtains a closer document-topic proportion on a given document than the STM. Figure 4.3(B) compares the L1 distances by plotting the mixSTM's distance to the truth versus the STM's distance to the truth, with axes on the $\log_{10}$ scale: as in Figure 4.3(A), the mass falls above a line passing through the origin and (1,1) and in favour of the mixSTM particularly for small values (closer fit to the truth). Interestingly, there appears to be a longer tail in favour of the STM in Figure 4.3(A) and more points in favour of the STM for large L1 distances in Figure 4.3(B), which suggests

[6] So chosen to be visible on the axes' scale.

Figure 4.3: (A): Histogram of difference in L1 distance (mixSTM-STM) between MAP estimates and true document-topic proportions for scenario B2. (B): L1 distance between the MAP estimates and the true document-topic proportions for the mixSTM plotted against the same for the STM, with axes on the $\log_{10}$ scale for scenario B2.



that when the distance to the true value is large in scenario B2, the STM finds a closer value to the truth than the mixSTM.

These conclusions are echoed when using the KL divergence for scenario B2 and related conclusions can be reached for other simulation scenarios under study here (for the analogous graphs, see Appendix A.2).

4.4.2.2 Held-out likelihood

Table 4.2: Mean (SD) per-document held-out likelihood for the mixSTM (mSTM), STM, and CTM.

	Per-document mean (SD)			Mean (SD) per-document difference	
Scen.	mSTM	STM	CTM	mSTM-STM	mSTM-CTM
A2	-5.854 (0.21)	-5.854 (0.21)	-6.048 (0.08)	.000516 (0.0052)	.194 (0.18)
B2	-5.793 (0.25)	-5.799 (0.26)	-6.032 (0.084)	.00664 (0.052)	.240 (0.23)

Mean (SD) per-document held-out likelihood across 5 folds and 10 simulated corpora in scenarios A2 and B2, for the mixSTM (mSTM), STM, and CTM, and mean (SD) of the difference in likelihoods between the mixSTM and the two other topic models considered. A positive difference is in favour of the mixSTM.

Table 4.2 presents the mean (SD) per-document 5-fold cross-validated held-out likelihood

for the mixSTM, STM, and CTM across 10 simulated corpora and the mean difference in held-out likelihood of the mixSTM relative to the other two topic models. The mixSTM has a held-out likelihood closer to 0 than the CTM in both scenarios: a mean (SD) increase in held-out likelihood of 0.19 (0.18) and 0.24 (0.23) for the intercept-only and intercept-and-slopes models, respectively. Both in the groupwise intercept-only and intercept-and-slopes model, mixSTM shows an improvement in average per-document held out likelihood compared to the original STM implementation: a mean (SD) difference in held-out likelihood 0.00052 (0.0052) for the intercept-only model and 0.0067 (0.052) for the slopes model. As expected, the difference is larger for scenario B2 where the mixSTM showed a closer fit to the true θ_k.

4.4.2.3 Topic quality

Much of the comparison of topic models often comes down to the interpretability of topics in light of the research question and corpus. In simulation, the generated "words" have no actual semantic relationship, only a simulated one. However, one can still use quantitative scores for topic exclusivity and semantic coherence or contrast the highest probability word indices for each topic.

Table 4.3 presents the per-topic sum of the semantic coherence scores and FREX scores using the top ten words for the STM and mixSTM as well as the summed differences (mixSTM-STM) across all 100 simulations in scenarios A2 and B2. The semantic coherence only differed between the two models for 5.4% of the 500 calculated scores in scenario A2 and 10.8% in scenario B2. Across all simulations and topics, the mixSTM had a lower mean semantic coherence and higher mean exclusivity than the STM in scenario A2, and a higher mean semantic coherence and lower mean exclusivity than the STM in scenario B2. On a per-topic basis, there is more variation: although the overall mean semantic coherence across all topics and simulations is in favour of the mixSTM for scenario B2, the semantic coherence is only better on average for two of five topics. Similarly, the mixSTM has more

Table 4.3: Per-topic and overall sum of semantic coherence scores and frequency-weighted exclusivity scores for two scenarios.

Scen.	Topic	$\sum C_{stm,k}$	$\sum C_{mix,k}$	$\sum C_{mix,k} - C_{stm,k}$	$\sum_{k=1}^{K} \sum C_{k,mix} - C_{k,stm}$
A2	1	-2970.3	-2971.8	-1.43	-11.66
	2	-3007.1	-3010.8	-3.70	
	3	-3063.5	-3063.6	-0.104	
	4	-3107.1	-3110.4	-3.30	
	5	-3145.0	-3148.1	-3.13	
B2	1	-2900.1	-2898.6	1.51	1.66
	2	-2808.9	-2810.7	-1.80	
	3	-3054.6	-3055.7	-1.11	
	4	-2999.3	-3002.1	-2.81	
	5	-3071.2	-3065.3	5.87	
Scen.	Topic	$\sum E_{stm,k}$	$\sum E_{mix,k}$	$\sum E_{mix,k} - E_{stm,k}$	$\sum_{k=1}^{K} \sum E_{k,mix} - E_{k,stm}$
A2	1	947.00	947.99	-0.0033	0.254
	2	946.34	946.43	0.091	
	3	949.17	949.12	-0.043	
	4	946.22	946.37	0.155	
	5	945.99	946.05	0.055	
B2	1	946.79	946.78	-0.0093	-0.109
	2	946.76	946.76	-0.0077	
	3	948.53	948.46	-0.071	
	4	944.88	944.99	0.112	
	5	946.27	946.14	-0.134	

Per-topic and overall sum of semantic coherence scores, C, and frequency-weighted exclusivity scores, E, for 100 simulations of scenario A2 and B2. The difference (mixSTM-STM) in C and E is also presented for each topic and overall. A semantic coherence score closer to zero represents better semantic coherence, and a higher (further from zero) exclusivity score represents more exclusive topics; a positive difference is a result in favour of the mixSTM.

exclusive topics overall in scenario A2, but has a better mean exclusivity for only three of five topics. In general, the mean difference in semantic coherence or exclusivity between the two topic models is small, and the median difference is near or at zero for each topic when taken across all simulations. For the simulated scenarios here, the size of the mean difference may not represent a practical difference in topic quality. This is supported by an examination of the most probable words for each topic (an example of which is presented in Appendix A.3), where the same words are often the most probable when adjusting for grouping structure using fixed or random effects, although there are some changes in the rank of specific words.

4.4.3 Discussion

In general, the mixSTM with a penalized mixed effects regression performs equally as well as the original STM in terms of fit, but provides an intuitive incorporation of grouping structure that better reflects the assumptions we have about some data collection methods such as focus groups. When there are differences in fit, topic retrieval and quality appear to be similar.

As hypothesized, the mixSTM does show a benefit in terms of its ability to retrieve the true document-topic proportions when groups of documents are smaller in size and when there are multiple random slopes, particularly if said random slopes are highly correlated. This was found to be the case whether or not additional covariates without a true effect were added to the model, and whether or not the full true random effects structure was specified correctly. Decreasing the size of the documents and the amount of variability between groups attenuated the benefit of using the mixSTM. Interestingly, in scenarios where the mixSTM's MAP estimates of θ were closer to the true topic proportions on average, the smallest distances to the true values that the mixSTM achieved tended to be smaller than those found by the STM, providing some intuition as to where the benefit is achieved. In real text, it is unlikely that the relationships are as straightforward as simulated here. There is a complex interplay of measured and unmeasured covariates having an impact on text at any time, but our results are encouraging for cases where there is a true effect, and even in cases where the true structure of group effects is not given to the model.

The mixSTM also had an improved per-document held-out likelihood relative to both the STM and the CTM. The improvement over the CTM arises naturally from the inclusion of additional information about covariates and the subsequent partial pooling. The improvement relative to the STM suggests that the separation of fixed and random effects during the penalization and regression model fit for μ leads to a posterior over the topic weights that is more generalizable to unseen instances, perhaps due to less overfitting specifically

for small groups. It may also be a result of the prior distribution choice: while the ridge regression penalizes all coefficients according to a single penalty to a degree affected by the amount of data available, the Horseshoe prior allows truly large coefficients to stay large during penalization thanks to the coefficient-specific tuning.

The mixSTM and STM had similar exclusivity and semantic coherence for each topic, without a clear benefit for either model. In the grand scheme of things, this makes sense: the proposed advantage of the mixSTM relative to the STM is in its incorporation of covariates into the document-topic prevalence model and representation of analyst beliefs, not necessarily the topic-word proportions that are the focus of the topic content model and metrics like semantic coherence and exclusivity. Further simulations which experiment with a range of numbers of topics and covariate settings would inform more widespread applications of this part of the mixSTM.

Chapter 5

5 Incorporating groups into the estimation of topic covariances

5.1 The STM's global topic covariance matrix

The STM, like the CTM, proposes a global covariance matrix for the relationship between the topics [16, 97]. In the generative model of the STM (see section 3.3.1), the covariance between topics is incorporated into the model through Σ, which is found in the distribution of θ (and η):

$$\theta_d \sim \text{LogisticNormal}_{K-1}(\mu_d, \Sigma)$$

Each document's topic proportions are assumed to be sampled independently from the above distribution. Choices about the covariance matrix have implications for the STM's estimation of document-topic proportions and for posterior sampling.

5.1.1 Sampling from the variational distribution of θ

In the STM (and mixSTM), recall that for each document, d, the variational posterior distribution of each η_d, $q(\eta_d)$, is a normal distribution parameterized by λ_d (a document-specific mean) and ν_d (a document-specific covariance matrix). Although the MAP estimates of θ_d and λ_d obtained in the last iteration of the STM's EM algorithm are stored in the output of the STM, the per-document covariance matrices ν_d are not: as $K-1 \times K-1$ matrices, storing each would be excessively demanding on memory, especially when they are not used individually in downstream EM iterations [97]. As such, the question of accessing and sampling from the variational posterior on θ_d becomes a question of how to estimate ν_d. The STM authors provide two methods [99].

The first, less computationally demanding method is to obtain an approximation to the document-specific covariance from the global covariance matrix, Σ, which *is* stored in the STM output. Recall from section 3.3.4.2 that in the STM, the global Σ for all documents is updated during the M-step as the following:

$$\hat{\Sigma} = \frac{1}{D}\sum_{d} \nu_d + (\lambda_d - \hat{\mu}_d)(\lambda_d - \hat{\mu}_d)^\intercal$$

By subtracting the mean squared difference of λ_d and μ_d (each obtained from the final iteration of the EM algorithm) from the global covariance matrix Σ, one can get an average approximation to ν_d across the documents, which we will call $\bar{\nu}_d$. This "global" method uses only estimates which are stored in the STM output, but maintains a reasonable amount of accuracy [99].

The second, more demanding method of obtaining ν_d, is to re-perform the final E-step update for each document. While this "local" method guarantees more precision about the uncertainty by using document-specific covariances, it involves the calculation and storage of D potentially large matrices. For this reason, Roberts et al. recommend the "global" approximation [99].

5.2 Two-level nested groupings provide a new opportunity for Σ

In the case of innately grouped documents, we propose that the generative model allow for several different covariance matrices between topics, rather than one globally shared matrix.

We consider a two-level nested grouping structure of documents, where each document uniquely belongs to one of G groups, $g = 1, ..., G$. Instead of assuming a global covariance

in the generative model, consider instead a *groupwise* covariance Σ_g:

$$\theta_d \sim \text{LogisticNormal}_{K-1}(\mu_d, \Sigma_g) \qquad \text{for } d = 1, ..., D, d \in g \tag{5.1}$$

The corpus of documents is partitioned into G groups of Δ_g documents apiece, each group with a different Σ_g.

The estimation of groupwise covariances could be applied equally to the STM or to the CTM, since both use a logistic normal distribution for the document-topic proportions.

5.2.1 Justification for partitioning Σ

Estimating topic relationships in a groupwise manner has a clear motivation. Since the documents within a group are assumed to come from the same source or share some innate property that clusters them together, it follows that the relationship between topics might differ between these groups. That is, the amount that any two topics are discussed in the same document might be different depending on the group. Insomuch as the generative model of a topic model attempts to model mechanisms by which text is created, groupwise covariances represent a real belief we may have about the text.

By incorporating grouped covariance matrices into the estimation, we also supplement the existing information (word co-occurrences) that the model can use to estimate expected document-topic proportions. This is one of the benefits provided by the CTM relative to LDA [16]. In the case of estimating Σ_g, this information about the relationships between topics is specific to the group at hand, which may then lead to a better fit.

We also expect that the global approximation to the posterior distribution of θ_d could become more precise by using grouped Σ_g. When using an overall global covariance to approximate the posterior, we are pooling potentially dissimilar information across the groups contained in the ν_d. If there is systematic difference in the covariances between groups, it

cannot be captured by a global Σ and the overall precision of posterior draws will be lower. In estimating Σ_g and using it in the global approximation, we expect that the variances will become closer to those under the local method (i.e., smaller). However, if the documents within a group are dissimilar, the variance estimates will explode due to a lack of overall pooling.

5.2.2 Σ_g in the variational EM algorithm

We will now explore the details of incorporating groupwise covariances for the variational inference algorithm of the STM.

We can write overall posterior distribution over the latent variables for the STM with group covariances as proportional to:

$$p(\eta_d, z_{n,d}, \kappa, \Gamma, \Sigma_g | Y, X, w_{n,d}) \propto \prod^{G} \prod_{\delta \in g}^{\Delta} \left[\text{Normal}(\eta_\delta | X_\delta \gamma, \Sigma_g)\right] \times \prod^{D} \prod^{N} \left[\text{Categ}(z_{n,d} | \theta_d) \times \text{Categ}(w_n | \beta_{d,k=z_{n,d}})\right] \prod p(\kappa) \prod p(\Gamma) \quad (5.2)$$

where the product of independent normal distributions over the documents is now a nested product of documents within groups. This amounts to the same number of terms in the product as in (3.10) since we are using the same mean-field approximation, but the normal distributions no longer have the same covariance matrices for all documents. This results in changes to the ELBO, specifically to the term which involves the covariance parameter in (3.14).

In the E-step of the variational expectation-maximization algorithm, we obtain the variational parameter updates (which rely on the MAP estimate of η_d) by maximizing (3.21) with respect to η_d. Since this is done on a per-document basis, the change to the E-step is a

substitution of Σ_g in place of Σ:

$$\sum^{V} c_v \log[\sum^{K} \beta_{v,k} e^{\eta_k}] - N \log \sum^{K} e^{\eta_k} - \frac{1}{2} \log |\Sigma_g| - \frac{1}{2}(\eta - \mu)^\intercal \Sigma_g^{-1} (\eta - \mu)$$

The M-step update for μ_d does not rely on Σ and thus does not change. To motivate our M-step update for the group covariance, Σ_g, we can use maximum likelihood estimation. We have, for each Σ_g associated with group g which has documents $\delta = 1, ..., \Delta_g$:

$$\begin{aligned}
\frac{\partial L(q)}{\partial \Sigma_g^{-1}} &= \frac{\partial}{\partial \Sigma_g^{-1}} \left[\Delta_g \log |\Sigma| + \sum^{\Delta_g} \text{Tr}(\nu_\delta \Sigma_g^{-1}) + \sum^{\Delta_g} (\lambda_\delta - \mu_\delta)^\intercal \Sigma_g^{-1} (\lambda_\delta - \mu_\delta) \right] \\
&= -\Delta_g \Sigma_g + \sum^{\Delta_g} \nu_\delta + \sum^{\Delta_g} (\lambda_\delta - \mu_\delta)(\lambda_\delta - \mu_\delta)^\intercal \\
0 &\overset{set}{=} -\Delta_g \Sigma_g + \sum^{\Delta_g} [\nu_\delta + (\lambda_\delta - \mu_\delta)(\lambda_\delta - \mu_\delta)^\intercal] \\
\iff \Delta_g \Sigma_g &= \sum^{\Delta_g} [\nu_\delta + (\lambda_\delta - \mu_\delta)(\lambda_\delta - \mu_\delta)^\intercal] \\
\iff \hat{\Sigma}_g &= \frac{1}{\Delta_g} [\sum^{\Delta_g} \nu_\delta + (\lambda_\delta - \mu_\delta)(\lambda_\delta - \mu_\delta)^\intercal]
\end{aligned} \tag{5.3}$$

Essentially, rather than taking the mean of the combination of the per-document covariance ν_d and the square difference of λ_d and μ_d across all documents (as is done in the STM and the CTM) groupwise updates occur by taking the mean of this sum for each group.

This new update for Σ_g can also be leveraged for draws from the variational posterior of η_d, using a modified global method from section 5.1.1 which we will call "global-grouped". Rather than estimating a single $\bar{\nu}_d$ across all documents, we now have the ability to compute G average approximations to the document-specific variational covariances ($\bar{\nu}_g$) and use these in our sampling from the variational posterior.

We note that the memory cost of storing grouped covariance matrices is larger than storing a single, global covariance (one $K-1 \times K-1$ matrix per group, rather than just one overall), but it will nonetheless be fewer matrices than storing each ν_d.

5.2.3 Generative model for the mixSTM

We can rewrite a full generative model for the mixSTM, which involves both a novel estimation on the mean topic weights (from Chapter 4) and a groupwise covariance:

$$
\begin{aligned}
(\gamma_k, u_k) &\sim \text{Normal}(\mu_{q(\gamma_k,u_k)}, \Sigma_{q(\gamma_k,u_k)}) \\
\eta_d &\sim \text{Normal}_{K-1}(\mathbf{\Gamma}' x_d' + \mathbf{U}' z_d', \Sigma_g) \qquad \text{for } d = 1, ..., D,\ d \in g \\
z_{d,n} &\sim \text{Categorical}_K(\theta_d) \qquad \text{for } n = 1, ..., N_d \\
w_{d,n} &\sim \text{Categorical}_V(\mathbf{B} z_{d,n}) \qquad \text{for } n = 1, ..., N_d
\end{aligned}
$$

Importantly, we still assume in the generative model that each document is an independent draw from a multivariate normal distribution with a mean determined by covariate relationships and a certain covariance. The new covariance does not introduce direct dependency between θ_d for different rows, but does restrict the exchangeability assumptions in a similar way as the incorporation of covariates into topical prevalence does.

5.3 Simulation studies for the mixSTM's covariance update

We explored the effect of grouped covariances through a number of simulations. We hypothesized that incorporating Σ_g into the estimation would improve the fit of the STM, but may face issues related to attempting to estimate a very large number of parameters relative to the amount of data.

We explored the effect of grouped Σ_g on two topic models: one, the mixSTM with random effects in the updates to η and two, the CTM using a covariate-free `stm` package implementation [99]. We were interested in implementing it on the CTM to see if, in the absence of incorporated covariates, grouped Σ_g led to better fit when there were true covariate ef-

fects.

5.3.1 Simulation methods

As before, for each simulation study, we simulated 1000 documents with a mean of 150 words from a 600-word vocabulary with the mean topic proportion defined by 10 continuous covariates (sampled from $\text{Normal}_{10}(0, \text{Wishart}(\mathbf{I}))$ and random coefficients (intercepts or slopes) drawn from $\text{Normal}(0, 1)$ distributions on a per-topic basis with a correlation between random effects of 0.1. In the case of random slopes, covariates associated with the random effects were generated from independent $\text{Normal}(0, 1)$ distributions. The per-topic mean of topic weights was calculated as a linear combination of the generated data and covariates: $\mu = X\beta + Zu$. For each document, θ_d was sampled from a logistic normal distribution with mean μ_d and one of two covariances:

1. $\Sigma_g \sim \text{Wishart}(0.25\mathbf{I})$ for $g = 1, ..., G$, representing truly different topic covariance matrices per group.

2. $\Sigma_g = \Sigma = 0.25\mathbf{I}$ representing identical topic covariance matrices in each group.

In each corpus, β was simulated from a $\text{Dirichlet}_{600}(0.25)$ distribution. For each word in each document, the topic $z_{n,d}$ was sampled with probability θ_d from a categorical distribution, and the word $w_{n,d}$ was sampled from a categorical distribution with probability $\mathbf{B}z$.

We specified four scenarios:

1. $G = 10$, $K = 5$, random intercept only.

2. $G = 50$, $K = 5$, random intercept only.

3. $G = 50$, $K = 20$, random intercept only

4. $G = 50$, $K = 5$, random intercept and three random slopes.

For each model, including the CTM, we specified the "Spectral" initialization of [99]. In the mixSTM, we additionally set a Horseshoe prior for the penalized coefficients.

5.3.1.1 Simulation comparisons

We examined how close each model's MAP estimates of the topic proportions were to the true document-topic proportions using the mean KL divergence $\frac{1}{K}\sum^{K}\hat{\theta}_d \times \log_2 \frac{\hat{\theta}_d}{\theta_d}$. We calculated the difference in mean KL divergence to the true topic proportions between the a model with grouped covariances and with a global covariance. We use the mean per-topic KL divergence (rather than the overall KL divergence as in Chapter 4) to make the fit comparable between models with different numbers of topics. To present the results, we show the median (Q1-Q3) distance obtained by the mixSTM and the median (Q1-Q3) difference between distances obtained by the two methods of estimating the covariance across all 100 simulations (100,000 KL divergences). We calculated the proportion of times out of 100 simulations for each scenario that the grouped covariances resulted in a closer median KL divergence from the MAP estimates to the true document-topic proportions and we present the overall percentage of differences in mean KL divergence in which the grouped Σ_g resulted in a smaller distance to the true document-topic proportions than the global Σ.

We also considered the performance in terms of the five-fold cross-validated held-out likelihood for a subset of 10 of the 100 simulated corpora for two scenarios. In each corpus, we sequentially held out a randomly chosen subset of 200 documents and fit a topic model (a mixSTM with either groupwise covariances or a single, global covariance) on the remaining 800. We calculated the mean per-word held-out likelihood in each held-out document using the quantities inferred from the training set. To report on the simulations, we present the mean (SD) document-level held-out likelihood across all test documents for each model and the mean (SD) of the difference between the held-out likelihood in the two models.

As a final comparison to see the effect on posterior draws for θ, we consider the variance (diagonal entries) in $\bar{\nu}_d$ (via the global approximation), $\bar{\nu}_g$ (via the global-grouped approximation) and ν_d (via the local approximation re-estimated from the original, global-covariance case). We present, for each scenario, the median (Q1, Q3) variance across all $K-1$ topics for the single $\bar{\nu}_d$, all G estimated $\bar{\nu}_g$, and all D re-estimated ν_d and plot the distribution of variances. To give context to the above results about changes to ν_d estimation, we resample from the variational posterior distribution over θ using each method. For each of 10 simulations, we draw 100 samples from the posterior of η_d for all 1000 documents and project them to the simplex to obtain θ_d. Then we compute the median (Q1, Q3) per-topic variance of θ_d for each document. We do not compare θ estimates directly, just the per-document-topic variance, since the MAP estimates of the variational mean are anticipated to differ between the approaches.

5.3.2 Simulation results

5.3.2.1 Σ_g in the mixSTM

Table 5.1: Comparison of the mixSTM estimated with separate covariances per group (mSTM_G) to the mixSTM estimated with only one global covariance (mSTM_{NG})

	Scn.	mSTM_G: Median (Q1, Q3) ($\times 10^{-3}$)	mSTM_G - mSTM_NG Median (Q1, Q3) ($\times 10^{-3}$)	%sims $\tilde{d}$<0	%diffs d<0
True group cov	1	5.1 (2.6, 9.2)	-0.75 (-2.7, 0.43)	97	66.2
	2	5.7 (3.0, 10.0)	-0.40 (-2.4, 1.2)	91	57.7
	3	22.4 (15.1, 32.3)	0.026 (-3.8, 4.0)	42	49.8
	4	5.4 (2.8, 9.7)	0.032 (-1.4, 1.5)	52	49.1
No group cov	1	7.6 (3.8, 14.6)	0.086 (-1.2, 1.3)	42	47.0
	2	6.6 (3.5, 11.3)	0.15 (-1.3, 1.7)	32	45.8
	3	33.6 (22.8, 48.8)	10.3 (3.3, 19.9)	0	14.3
	4	4.5 (2.2, 7.9)	0.11 (-0.66, 1.1)	7	45.4

Comparing the mixSTM estimated with separate covariances per group (mSTM_G) to the mixSTM estimated with only one global covariance (mSTM_{NG}). We present the median (Q1, Q3) mean KL divergence to the true document-topic proportions for mSTM_G, the median difference in mean KL divergence between mSTM_G and mSTM_{NG} (negative in favour of estimating Σ_g), as well as the percentage of 100 simulations where the overall median KL divergence $\tilde{d}$ for that simulation was smaller for the grouped covariance model, and the percentage of all documents in which the KL divergence was smaller for the grouped covariance model.

We present the results of our simulations for the mixSTM in Table 5.1. When the number of groups was small, the model had only a random intercept, and there was a true groupwise difference in topic covariances (scenario 1), incorporating a groupwise covariance improved the fit of the topic model: in 97% of simulations and for 66% of per-document differences, the median KL divergence was better for the model with Σ_g than with a global Σ; this corresponded to a median difference in mean KL divergence of -0.00075. When the number of groups increased to 50 (scenario 2), 91% of simulations and 58% of all KL divergence differences were in favour of Σ_g. The median KL divergence between the mixSTM with groupwise covariances' MAP estimates of θ and the true θ were comparable. When we increased number of topics from 5 to 20 (scenario 3), the improvement to the fit was attenuated: the groupwise topic covariance improved the median KL divergence in only 42% of simulations, and improved the KL divergence across 50% of all documents. Similarly, for a five-topic model with random intercepts and random slopes (scenario 4), the groupwise topic covariance improved the median KL divergence in 52% of simulations, and improved the KL divergence in 49% of all documents.

In a case where there was no simulated difference in covariances between the groups, incorporating a groupwise covariance matrix into the mixSTM did not improve the fit. Particularly for scenario 3 (with 50 groups, 20 topics, and a random intercept) and scenario 4 (with 50 groups, 5 topics, and both random intercept and random slopes), the fit was always or almost always worse when performing estimation with groupwise covariances.

5.3.2.2 Σ_g in the CTM

The results of our comparison of covariate-free topic models (the CTM) with and without estimating groupwise covariances are presented in Table 5.2. When there was truly a grouped covariance, incorporating groupwise estimation of Σ_g improved the fit in all four scenarios, but particularly in scenario 3 (50 groups and 20 topics) where 64% of all documents had MAP estimates which achieved a mean KL divergence closer to the true

Table 5.2: Comparison of a CTM estimated with separate covariances per group (CTM_G) to a CTM estimated with only one global covariance (CTM_{NG})

	Scn.	CTM_G: Median (Q1, Q3)	CTM_G - CTM_{NG} Median (Q1, Q3) ($\times 10^{-3}$)	%sims $\tilde{d}$<0	%diffs d<0
True group cov	1	8.4 (4.6, 14.7)	-0.39 (-2.4, 1.3)	96	56.5
	2	8.6 (4.7, 14.8)	-0.43 (-3.0, 1.7)	96	55.8
	3	23.0 (15.5, 33.2)	-2.0 (-6.0, 1.8)	100	64.1
	4	9.8 (5.2, 17.3)	-0.31 (-2.8, 1.6)	99	55.1
No group cov	1	11.4 (5.9, 22.2)	0.33 (-2.0, 3.1)	32	45.9
	2	8.9 (4.8, 15.8)	-0.28 (-3.2, 2.5)	88	53.1
	3	34.3 (23.7, 49.4)	-0.90 (-9.8, 6.4)	78	53.4
	4	9.1 (4.9, 15.9)	-0.44 (-3.1, 1.5)	100	56.6

Comparing a correlated topic model (no covariates in the estimation) estimated with separate covariances per group (CTM_G) to a CTM estimated with only one global covariance (CTM_{NG}). We present the median (Q1, Q3) mean KL divergence to the true document-topic proportions for CTM_G, the median difference in KL divergence between CTM_G and CTM_{NG} (negative in favour of estimation Σ_g), as well as the percentage of 100 simulations where the overall median KL divergence $\tilde{d}$ for that simulation was smaller for the grouped covariance model, and the percentage of all documents in which the mean KL divergence was smaller for the grouped covariance model. Note that although there were no covariates included in the estimation, the simulation of the documents assumed a relationship between topic prevalence and some metadata.

document-topic proportions. In the situation where there was no simulated difference between the groups in topic correlations, estimating a groupwise covariance still improved the fit relative to a model with only a global covariance matrix in models with a larger number of groups (but not when there were few groups). The fit was particularly improved for the most complex model with 50 groups and random slopes along with the random intercept, where 100% of the simulations had a median KL divergence closer to the truth and 57% of documents had a smaller KL divergence to the truth.

5.3.2.3 Held-out likelihood

Table 5.3 presents the mean five-fold cross-validated held-out likelihood for scenario 1 (10 groups) and scenario 2 (50 groups) fit using the mixSTM with and without partitioned groupwise covariances. In both cases, the estimation of a topic model with grouped covariances performed worse in held-out likelihood than with a single global covariance: a mean (SD) difference between the two models of -0.0023 (0.016) when there were 10 groups and of -0.0028 (0.018) when there were 50 groups.

Table 5.3: Mean (SD) per-document held-out likelihood for a mixSTM with and without a groupwise covariance matrix.

	Per-document mean (SD)		Mean (SD) per-document difference
Scen.	mSTM$_G$	mSTM$_{NG}$	mSTM$_G$-mSTM$_{NG}$
1	-5.879 (0.23)	-5.876 (0.23)	-0.0023 (0.016)
2	-5.875 (0.23)	-5.872 (0.23)	-0.0028 (0.018)

Mean (SD) per-document held-out likelihood across 5 folds and 10 simulated corpora in scenarios 1 and 2 for a mixSTM estimated with grouped covariances (mSTM$_G$) and with a global covariance (mSTM$_{NG}$) , and mean (SD) of the difference in likelihoods between the mixSTM under the two modelling assumptions. A positive difference is in favour of estimation with a grouped covariance.

5.3.2.4 Variance estimates using ν re-estimation

Table 5.4: Comparison of distribution of posterior variational variance estimates ν_d estimated under local, global-grouped, and global methods for simulation scenario 2.

50 groups	Mean (SD)	Median (Q1, Q3)
ν_d (N=4000/sim)	0.46 (0.35)	0.36 (0.19, 0.62)
$\bar{\nu}_g$ (N=200/sim)	0.40 (0.33)	0.30 (0.19, 0.50)
$\bar{\nu}_d$ (N=4/sim)	0.46 (0.11)	0.46 (0.38, 0.54)

Table 5.4 and Figure 5.1 present the distribution of the variational posterior variance estimates for 100 simulations under each method of approximating ν_d in simulation scenario 2. The global method results in variance estimates tightly concentrated around the mean of ν_d. The distribution of the global-grouped variance estimates more closely resembles those of the local method in terms of dispersion, although the overall mean and median of $\bar{\nu}_g$ are biased downwards relative to the local method.

Figure 5.1: Distributions of variational posterior variances ν under "global" (NG), "global-grouped" (G), and "local" (L) estimation methods for scenario 2 in the mixSTM.

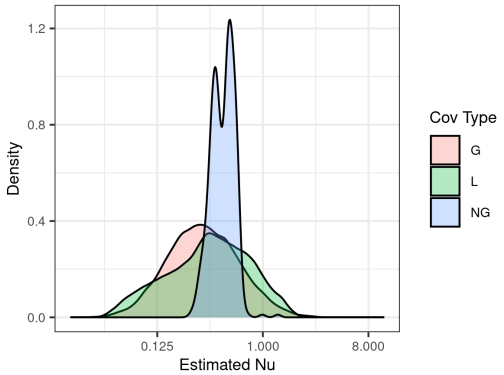


Table 5.5 presents the variance of document-topic$_1$ proportions for a subsample of 10,000 documents (of the 100,000 simulated) for four scenarios fit using the mixSTM. For a given topic in scenarios 1 (10 groups), 2 (50 groups), and 4 (50 groups, random slopes), the median variance of $\theta_{d,k}$ is smallest when the local re-estimation strategy for ν_d is used, intermediate when the global-grouped $\bar{\nu}_g$ is used, and largest when the full global $\bar{\nu}_d$ is used. In each of these scenarios, the global-grouped variance for the document-topic samples is closer to the local variance than the global variance. In scenario 3 which had 20 topics rather than 5, the median document-topic proportion variance was still smaller when using global-grouped approximation than global approximation and the local re-estimation yielded the highest variance of the three methods. Across all 10,000 documents, a smaller variance for $\theta_{d,k}$ than the one estimated using the local method was obtained for 42.9% of documents when using the global-grouped method and 31.2% of documents when using the global method.

Table 5.5: Median (Q1, Q3) variance of 100 samples drawn from the estimated variational posterior of $\theta_{d,k}$, using the local, global-grouped, and global approximations to ν_d in the mixSTM.

	Median (Q1, Q3) variance of $\theta_{d,k}$ ($\times 10^{-2}$)		
Scen.	ν_d	$\bar{\nu}_g$	$\bar{\nu}_d$
1	0.14 (0.058, 0.23)	0.19 (0.035, 0.61)	0.30 (0.045, 1.1)
2	0.13 (0.041, 0.22)	0.14 (0.021, 0.46)	0.26 (0.029, 0.95)
3	0.040 (0.012, 0.10)	0.017 (0.0028, 0.11)	0.035 (0.0084, 0.18)
4	0.087 (0.017, 0.19)	0.10 (0.0091, 0.53)	0.16 (0.010, 0.95)

5.3.3 Discussion

Estimating groupwise covariances in a model with innate groups of documents can improve model fit in a few cases. First, when used with the mixSTM, the grouped covariances benefit the model when there is a true difference between the groups in the topic correlations, when there are few topics, and when the model on the mean topic weights is simple. Second, when used with a CTM, grouping the covariances benefits the model fit in

almost all studied cases. However, even when the grouped covariances improve fit for the mixSTM, they do not improve the held-out likelihood. On a per-document-topic basis, the estimated posterior variance of θ when using a global-grouped method falls between the single-covariance global estimation and the local re-estimation, for most scenarios.

Given these results, we can infer that the main drawback to estimating groupwise covariances in this manner is overparameterization. The two modifications we have proposed to the mixSTM (to the mean and covariance) act on the same part of the topic model: the document-topic proportions. In models which are already provided with information about the expected topic proportions, the amount of additional information provided by pooling the correlations within groups does not tend to have a benefit, especially if there are many parameters being estimated (e.g., more topics, or more varying slopes). This is likely because we have already saturated the model with parameters and used our available degrees of freedom. Future experiments with a range of corpus sizes may explore this issue in more depth.

In contrast, when there is an underlying grouping structure but we do not provide it to the model (using a CTM), the group covariances do improve the fit. In this case, the only information being provided to the model about document-topic proportions is in the covariances, and since there is a true difference in the document-topic means, this information is useful for model fit. This also explains why estimating Σ_g improves the fit on a CTM even in the absence of a true between-group difference in covariance: the covariances are used to relay information about the expected document-topic proportions that is not available elsewhere. However, when the groups are large, the benefit is somewhat attenuated, likely because the covariances are less specific to the individual documents. Broadly in the context of the CTM, the estimation of groupwise covariances opens up options for topic models on grouped corpora such as focus groups, even in the absence of comprehensive covariate information: without knowing any more than the grouped structure of documents, the analyst

could potentially improve the fit of the model by assuming per-group covariance matrices in the generative model.

It is unsurprising that the grouped covariances led to worse performance in terms of held-out likelihood, even when the overall fit was improved. By separately estimating covariance matrices in partitions of the data, we are limiting the amount of information available to estimate these quantities, which already makes the estimated covariances more sample-specific. In removing some documents from the groups, we lessen the amount of information available to estimate the relationships between topics in the group even further. It follows naturally that the resulting covariance estimates are less generalizable to a held-out sample.

Comparing draws from the variational posterior of θ using three methods of estimating the variational posterior variance reveals that the mean and median global-grouped estimates tend to fall between the local and global methods. This suggests, as expected, that a benefit of estimating groupwise covariances might be in pooling less dissimilar information into the variance estimates which grants more precision within each group. Although the mean and median variance for the global-grouped method is slightly larger than for the local method (more uncertain on average), for individual document-topic proportions, the variance is smaller in nearly half the documents– an increase from the one-third of documents where this is true in the global case. That the overall density of the variances is somewhat shifted towards zero suggests that draws from a posterior using the global-grouped approximation may underestimate the variance of document-topic proportions in some documents. This seems to particularly be the case when there is a larger number of topics, where in the presented example, both the global and global-grouped methods have lower document-topic variance estimates than the local estimation.

Some caution in using grouped covariances may be warranted in cases where there is already an abundance of information incorporated into the model about the mean document-

topic proportions or when there is relatively little data. In practice, one must motivate the use of group-specific covariance matrices with domain- and corpus-specific knowledge about the nature of the groups and expected differences between groups. After all, outside of simulation, the ability to test the fit of the model in this way is nonexistent; we must instead rely on the knowledge of the data we have.

Chapter 6

6 Estimating the effect of covariates on topic proportions in a setting with grouped documents

6.1 The STM's strategy for a regression on topic prevalence

Beyond the estimation of topics with prior information granted through covariates, the STM has a second functionality: the ability to estimate the association of these covariates with the inferred document-topic proportions using an exploratory regression.

6.1.1 The Method of composition

The method of composition is an iterative Monte Carlo method that provides a way to obtain independent samples from the marginal distribution of a variable, $p(\beta)$, using the conditional density of the variable of interest (β) conditioned on another variable, y, and the distribution of y, $p(y)$ [116]. Formally, suppose one wants to sample from the distribution $p(\beta) = \int_y p(\beta|y)p(y)dy$. Then the method of composition is:

1. Sample $y^{(i)} \sim p(y)$
2. Sample $\beta^{(i)} \sim p(\beta|y^{(i)})$
3. Repeat steps 1 and 2 m times, obtaining m samples of both $y^{(i)}$ and $\beta^{(i)}$.

$\beta^{(1)}, ..., \beta^{(m)}$ are independent samples from the marginal distribution $p(\beta)$, and each pair $(\beta^{(i)}, y^{(i)})$ is a sample from the joint density $p(\beta, y)$ [116].

This method is useful when the integral for the distribution $p(\beta)$ is intractable. It also has another benefit, highlighted in practical use by Treier and Jackman [121]: it provides a way

to incorporate measurement uncertainty in the variable y into the estimation of the variable β.

Suppose that one has a regression outcome, y, which is measured with some uncertainty. This uncertainty is quantified in the known distribution of y (obtained from repeated sampling, for example, or else explicit distributional assumptions about the variable). If one performed a linear regression of data X on a single sample of y, one could obtain point estimates of the regression coefficients $\hat{\beta}$, which also have some variance. However, this variance around $\hat{\beta}$ is with respect to the regression and does not include the uncertainty with respect to y.

Trier and Jackman suggest an approach to the method of composition applicable to regressions with measurement uncertainty on an independent variable [121], which can be extended to the case of a dependent variable. The steps (outlined in the appendix of [121]) are as follows:

1. Sample $y^{(i)} \sim p(y)$ where $p(y)$ is the distribution of y.
2. Run the regression of interest using each sample $y^{(i)}$ to obtain m estimates of the coefficients $\hat{\beta}^{(i)}$ and variance-covariance matrices of the coefficients $\hat{V}^{(i)}$.
3. Sample $\tilde{\beta}^{(i)}$ from the asymptotic distribution of the regression coefficients for each of the m regressions: $N(\hat{\beta}^{(i)}, \hat{V}^{(i)})$.

The samples $\tilde{\beta}^{(i)}$ are draws from the marginal posterior density of β which incorporate the uncertainty from both y and each regression on y.

Since step 2 of [121]'s procedure uses a frequentist regression, the final step draws from the asymptotic distribution of the maximum likelihood estimator of β. Using this in place of a posterior density implies two assumptions [106]. First, we assume there is enough data for the likelihood to wash out the prior distributions on the coefficients, leading to a normal form for the posterior [106, 124] . Second, the prior distribution on the coefficients, $p(\beta)$ is

implicitly and somewhat unreasonably assumed to be uniform [106, 123].

In any topic model, including the STM, the document-topic proportions θ are latent variables which have some distribution, and with that distribution comes an estimate of uncertainty. If the document-topic proportions are used as the outcome variable in a post hoc regression, not incorporating the estimation uncertainty around said proportions might lead to an underestimation of the uncertainty in the regression coefficients [99]. The STM provides the option to use the method of composition, modelled after [121], to estimate the coefficients and their standard errors [99].

6.1.2 Accessing θ, our dependent variable

In order to estimate the effect of covariates on document-topic proportions, one must first obtain θ, the dependent variable of the regression. In the STM (and mixSTM), for each document, d, the variational posterior distribution of η_d is a normal distribution parameterized by a document-specific mean λ_d and document-specific covariance matrix ν_d.

Recall the three methods shown in section 5.1.1 for sampling from $q(\eta_d)$:

- "Global", which involves only quantities available in the output of the STM and involves calculating a single mean approximation to the covariance matrices, $\bar{\nu}_d$, by subtracting off the mean squared difference of λ_d and μ_d from the single global Σ.
- "Global-grouped", which is similar to the global method except it utilizes groupwise estimated Σ_g in the approximation, so we obtain $\bar{\nu}_g$ (one estimate per group).
- "Local", which involves re-computing the E-step of the variational algorithm, given the results of the STM. It is more computationally demanding but more accurate, since it yields one ν_d per document.

The `stm` package in R also offers an option to use the MAP estimates of θ as the dependent variable of the regression without recomputing an (approximate) variational posterior. This

method (chosen by setting `uncertainty="None"` when sampling from θ's posterior in the `stm` package) incorporates none of the uncertainty about θ into downstream inference [99].

In any of the methods which do incorporate the uncertainty over θ, once ν_d has been obtained, samples can be drawn from the variational posterior over η_d for each document. These samples of the topic weights are projected to a simplex, transforming them into θ_d [99]. In this way, we can make draws from the posterior distribution of θ to be used in the method of composition.

6.1.3 Inference on θ

Once the outcome variable is determined, the STM then performs a linear regression of the covariates of interest on the vector of topic proportions for topic k, θ_k [99]. The `stm` package recommends that the covariates used in the regression be a subset of those used to fit the model [99]. In the case where no latent variable uncertainty will be incorporated, this is amounts to one regression per topic of interest. In the case where we are sampling from the posterior over θ to incorporate this uncertainty, we take m draws of the outcome and must run a linear regression on each of them. The coefficient estimates and variance-covariance matrix of the coefficients of each of the m regressions are stored.

In the simple linear regression case, the asymptotic distribution of the maximum likelihood estimate of β is a multivariate normal with mean $\hat{\beta}$ and variance-covariance matrix equal to the estimated covariance matrix of the coefficients from the regression [124]. Draws are taken from this asymptotic distribution Normal($\hat{\beta}$, $\hat{V}$), and the mean and standard deviations of the draws are used to obtain the final estimates of regression coefficients, their standard errors, and t-values.

6.1.3.1 Choosing a regression family and link function

The original STM uses a linear regression to model the relationship of the covariates with the topic proportions. The authors note in the documentation that this can mean that predicted or estimated topic proportions may occasionally be negative, since they do not use a regression bounded on the interval between 0 and 1 [99]. The use of linear regression for proportion outcomes is far from unheard of and will typically result in reasonable estimates; when the predictions lie outside the realm of possibility, Roberts et al. recommend using splines on the independent variables [99]. Unlike in an experimental setting, performing an observational, exploratory regression means the consequences of unbounded proportions are minimal– but they still must be interpreted with caution.

This being said, Schulze et al. point out that several users of the `stm` package have reported negative topic proportions in their publications and that alternate regression formulations should be recommended [106]. At the cost of interpretability for the broadest audience, a non-linear model may be desired to better incorporate flexibility and to obtain non-negative confidence intervals and proportions across the whole span of possible covariate values. Schulze et al. propose a beta regression[1] [106].

Beta regression is a flexible-shape, distributional regression (rather than a generalized linear model). It can model outcomes that lie strictly between 0 and 1 (not 0 and not 1, although zero/one-inflated beta regressions exist). In the case of the STM, because of the transformation of η using exponentials, the topic proportion of a given topic in a document can never be exactly zero. This makes a beta regression a theoretically strong contender for an alternative model for the association of covariates with topic proportions. However, the interpretation of coefficients from a beta regression is non-intuitive. In any multiple beta

[1] Schulze et al. also mention quasibinomial regression as an option, although their clear preference is for beta regression. Quasibinomial regression is similar to a binomial (logistic) regression in that its coefficients are on the log odds scale, but has an additional dispersion parameter in its likelihood. The results can be interpreted for changes in proportions [55].

regression, the only way to get interpretable marginal values of regression coefficients is to simulate data at the median values of all other coefficients, use the simulated data to obtain the estimate, and transform the estimate from the logit scale to proportions [55]. A beta regression also includes a precision parameter which can be specified on a per-coefficient basis, further complicating the interpretation but allowing for more flexible relationships of covariates with the proportions.

As an alternative to the linear, beta, or quasibinomial regressions, the truncated linear model (or 'linear probability model') is as interpretable and flexible as the linear model, but puts bounds on the outcome [55].

6.1.3.2 Choosing priors for the coefficients

As mentioned above, sampling coefficients using the asymptotic distribution of the maximum likelihood estimates of the coefficient estimators in place of the the posterior distribution of the coefficients implies an improper uniform prior. This is not only implausible, but can be informative about coefficient estimates and may lead to issues when there is insufficient data to learn the model parameters [123, 19].

For inference in linear models, some have suggested the use of a broad normal prior with mean 0 and some choice of the variance for all coefficients [123]. However, in order to incorporate specific prior assumptions in this manner, one must move away from the frequentist framework and into Bayesian hierarchical models to fit the regression. Schulze et al. implement a fully Bayesian beta regression for use in the estimation of covariate associations with topic proportions [106].

6.2 Implementation in a setting with grouped documents

In a setting where we have groups of documents, we propose estimating the association between covariates and topic proportions using a mixed effects regression.

6.2.1 Motivating the addition of random effects

When documents are inherently grouped, it becomes important to incorporate this structure into the analysis. If incorporated as fixed effects, grouping variables will be treated as levels of a factor, interacting with the intercept or other covariates. If incorporated as random effects, groupings are instead parameterized by variance components which describe the amount of variation across groups. In either model, if the group is not included and the relationship between the outcome and a covariate differs by group, we may confound other effects of interest. In the analysis of topic prevalence, the grouping variable is an important descriptor for the patterns under study.

We will enumerate some potentially relevant motivations for using random effects when document-topic proportions are the outcome, with the example of focus groups.

- When observations cannot be collected independently, we can no longer assume uncorrelated residuals. Fixed effects models implicitly assume that the between and within-group variance are equal [5], but a mixed model explicitly models the between-group correlation for any group-varying covariates. This represents a more realistic assumption about the covariates that are anticipated to vary between groups, since they will contribute together to develop the effects.

- The random and fixed effect models lead to different interpretations. In fixed effect models, each group obtains a coefficient but it is difficult to summarize the effect of the grouping as a whole. In contrast, the variance components of a random effects model can be used to describe how much of the variance in topic proportions for topic k is due to the grouping structure of the documents versus between-document residual variance, giving insight into the importance of differences between groups. Furthermore, there are scenarios where it would be undesirable to interpret the effect of the grouping variable in terms of the specific groups. Consider the case of focus

groups, where, if each focus group session or participant is a "group", it could be inappropriate for anonymity's sake to interpret the change in topic proportion at this level. It may also not address a question of interest in some studies: whether the observed variation is due in large part to the innate grouping structure.

- Random effects models also have a more straightforward interpretation for new levels of the grouping variable. Predictions for a fixed effects model assume that all relevant group levels have been measured: any new observation will belong to an existing grouping level. In random effects models, the levels of the grouping variable are considered samples from a distribution, which permits predictions for topic proportions to be made for out-of-sample grouping levels. In focus groups, one might be interested in generalizing the topic prevalence to additional locations or groups of individuals, although we must be aware of the potential to overfit to the current sample [37].

- Mixed models can be a helpful framework if the number of documents in one or more groups is small. The uncertainty of the coefficients for small groups' dummy variables in a fixed effects model could be large and the estimates vulnerable to overfitting. In contrast, in a mixed model, fewer parameters need to be estimated to account for each group, and when groups are small, mixed models "borrow" information from the grand mean through partial pooling to predict the group effects. This avoids some of the overfitting concerns and results in more stable estimates. In focus groups, the number of documents per group may vary depending on circumstances and the participants, so flexibility around group size is needed.

- In a fixed effects model, it is challenging to incorporate covariates at the same level as the grouping, as these will be perfectly collinear with the dummy grouping variables and their effect will be impossible to disentangle from those of the group itself [32]. Also, if a covariate varies little within a group but has high residual variance, this

can cause a destabilizing effect on coefficient estimates in a fixed effect model [32]. Mixed effects models can handle additional group-invariant variables in the fixed effects portion since the groups are modelled in the random effects portion, while still achieving stable coefficient estimates. The group is an important division in focus group analyses: additional participant or focus-group level covariates may be of interest when exploring changes in topic proportions.

For these reasons and more, we adapt the method described in section 6.1.1 to model topic proportions using mixed effects regression.

6.2.2 The Method of composition for mixed effect models

In order to obtain samples from the posterior of β (the fixed effects coefficients) and ω (the random effects variance components) while incorporating estimation uncertainty about the topic weights η, we implement the method of composition as follows:

1. For each document, sample m times $\hat{\eta}_d \sim q(\eta_d) = \text{Normal}(\lambda_d, \nu_d)$ with ν_d obtained using a method discussed in section 6.1.2. Project each $\hat{\eta}_d$ to the simplex and obtain $\hat{\theta}_d$.

2. For each topic of interest, use the posterior distribution of the coefficients and variance components given the outcome from sample i obtained in step 1 and the covariate information $\mathbf{X}$ to draw $(\hat{\beta}, \hat{\omega}) \sim p(\beta, \omega | \hat{\theta}_k^{(i)}, \mathbf{X})$ many times.

3. Obtain the final estimates of β and ω for a topic by taking the mean and standard deviation across all samples from step 2.

Using this strategy, we can incorporate uncertainty about the latent document-topic proportions into the estimation of both random effect variance components and fixed effect coefficients. This provides, in addition to the typical interpretations under the original fixed effects implementation, a method to describe variation between groups and correlations

between group-varying predictors.

6.2.2.1 Approaches to a joint posterior and their assumptions

To obtain draws from the posterior in step 2 of the method of composition above, we consider three avenues: 1) sampling from the asymptotic distribution of the likelihood; 2) a pseudo-Bayesian approach where we use MCMC sampling starting from the normalized likelihood of a fit model; or 3) a full Bayesian regression approach which samples from the posterior during model fitting.

Avenue 1 most closely resembles the original implementation in the `stm` package. Like for the MLE of the linear model, some asymptotics exist for linear mixed models fit by REML or MLE (see [65] for a comprehensive summary). In particular, under some conditions which vary by the likelihood choice, the joint distribution of β and ω is asymptotically normal [65]. Jiang [65] rightfully points out, however, that determining the appropriate sample size at which asymptotic results hold is difficult when there are many components to the sample size (group size, number of groups, and overall sample size) and to the number of predictors. Also, calculating the covariance matrix of the asymptotic distribution is non-trivial, as it involves the square root and inversion of a number of matrices [65]. Furthermore, the same concerns about the implicit prior forms holds as mentioned in section 6.1.3.2.

In Avenue 2, the likelihood and parameters of a frequentist fit are used as starting points for an MCMC procedure. We can use the MCMC sampling to draw from the posterior distribution of β and ω, however, these samples are not independent which can bias the variance of estimates downwards. At minimum, the effective sample size for each parameter under consideration should be verified to be large enough before interpreting the resulting estimates. Avenue 2's key advantage relative to Avenue 1 is that it makes no assumptions about the required sample size for asymptotics to be valid. However, this strategy makes the same assumptions regarding the likelihood outweighing the priors and about the priors

taking on an improper form (unless otherwise specified) [19]. Taking MCMC samples can be slow, requires monitoring convergence, and may not successfully converge when there is insufficient data or when the random effects covariance matrix is near-singular [19].

Avenue 3 is the extension of the fully Bayesian method suggested by Schulze et al. for the mixed model scenario [106]. Using MCMC, a regression model can be fit with specific priors on coefficients and random effect variance/covariances. As in Avenue 2, one obtains samples from the posterior distribution of all model parameters during fitting, but these draws are not all independent. This fully Bayesian method makes no assumptions about the relative weight of the likelihood versus the priors, and can set any form for the priors. Like all MCMC approaches, it can be slow and requires convergence checking.

It should be noted that while Avenues 2 and 3 are easily extended to a beta regression, we are not aware of the asymptotics of a mixed effects beta regression and thus the implementation of Avenue 1 is potentially not as straightforward for that formulation.

We cast no aspersions about which of the three avenues is most appropriate for any given setting. The goal is to propagate all sources of variance to the final coefficient estimates in a way that represents the relationship of covariates with topic proportions, which is something any of these methods can accomplish.

6.2.3 Implementation

Existing packages in R can be leveraged to perform the post hoc regression on topic prevalence.

For a frequentist-style mixed model fit, we apply the `glmmTMB` package [20]. `glmmTMB` uses the Laplace approximation to the intractable integrals in the likelihoods of generalized linear models [20]; user-set priors on the parameters can be incorporated. We chose this package as both linear and beta regression models are implemented, and the associated package `tmbstan` [86] is capable of performing post hoc sampling on the posterior of the

coefficients using the likelihood, accommodating Avenue 2.

For a Bayesian-style mixed model fit, we use the `brms` package [21]. Both linear and Bayesian beta regressions are possible through the specification of a family and link function. The priors on each parameter of the regression can be set by the user. Notably, the default priors for `brms` are in fact flat priors on the regression coefficients; in our implementation, we specify Gaussian priors on the fixed effect coefficients and intercept as mentioned in section 6.1.3.2.

The decision of which implementation of a regression model to use does not necessarily have a correct answer; we propose alternate R packages that could be used in Appendix A.4. In practice, both practicality and the availability of all desired utilities (like regression posterior sampling) outweighed other considerations in our implementation.

6.2.4 Example

To illustrate the variance propagation strategy for a post hoc regression on topic prevalence, we use a single simulated dataset and topic model from Chapter 4 under scenario A2 (fifty groups, no group-varying predictors but a single group intercept).

We estimated the effect of ten fixed effects and one random intercept on document-topic prevalence for Topic 1, using a linear mixed effects model fit using `glmmTMB`. To obtain estimates of the coefficients, 4000 draws (after a 1000 sample burn-in) from 4 chains were taken from the approximate posterior for both a regression on the MAP estimates (no additional uncertainty incorporation) and regressions on 100 samples from the variational posterior over θ using the "global" approximation. To compare the amount of variance propagated by the original STM's `estimateEffect` method, we also used a fixed effect regression with the random intercept represented through dummy variables and drew 1000 samples from the asymptotic distribution of the coefficient estimators.

Table 6.1 presents the results of the mixed effect model under both a "global" uncertainty

incorporation and no uncertainty incorporation; Table 6.2 presents the same for a fixed effects model.

Table 6.1: Comparison of coefficients estimated using a mixed effect linear regression with "Global" and "None" uncertainty incorporation.

	Global		None	
	Estimate	Std. Error	Estimate	Std. Error
Intercept	0.179	0.0220	0.179	0.0215
$\beta 1$	0.00199	0.00485	0.00200	0.00381
$\beta 2$	0.0323	0.00448	0.0330	0.00354
$\beta 3$	-0.0246	0.00504	-0.0249	0.00404
$\beta 4$	-0.0269	0.00513	-0.268	0.00395
$\beta 5$	0.0158	0.00497	0.0163	0.00399
$\beta 6$	0.0210	0.00497	0.0216	0.00395
$\beta 7$	-0.0379	0.00515	-0.0380	0.00396
$\beta 8$	-0.0182	0.00483	-0.0184	0.00397
$\beta 9$	0.00162	0.00451	0.00145	0.00364
$\beta 10$	0.0185	0.00516	0.0189	0.00421
$\log_2 \omega$	-1.888	0.1063	-1.870	0.1052

Coefficient estimates and standard errors for a post hoc mixed effect regression on topic prevalence from the mixSTM. Four chains of 4000 samples were drawn from the approximate posterior distribution of the coefficients and variance components using tmbstan, and averaged across all draws from the variational posterior of topic weights (100 draws in the case of "global"; the MAP estimates/1 draw for "none"). $\log_2 \omega$ is the natural logarithm of the random intercept standard deviation (default scale for glmmTMB).

Table 6.2: Comparison of coefficients estimated using original fixed effects regression with "Global" and "None" uncertainty incorporation.

	Global		None	
	Estimate	Std. Error	Estimate	Std. Error
Intercept	0.101	0.0295	0.0946	0.0238
$\beta 1$	0.00222	0.00494	0.00217	0.00380
$\beta 2$	0.0325	0.00455	0.0329	0.00357
$\beta 3$	-0.0250	0.00515	-0.0250	0.00406
$\beta 4$	-0.0264	0.00513	-0.0269	0.00395
$\beta 5$	0.0159	0.00503	0.0161	0.00402
$\beta 6$	0.0215	0.00511	0.0217	0.00408
$\beta 7$	-0.0378	0.00511	-0.0379	0.00400
$\beta 8$	-0.0183	0.00496	-0.0185	0.00392
$\beta 9$	0.000849	0.00462	0.00120	0.00356
$\beta 10$	0.0192	0.00534	0.0190	0.00421
U_2	-0.0837	0.0402	-0.0773	0.0344

Coefficient estimates and standard errors for a post hoc fixed effect regression on topic prevalence from the STM. 4000 samples were drawn from the asymptotic distribution of the coefficient estimators and averaged across all draws from the variational posterior over η (100 draws in the case of "global"; the MAP estimates/1 draw for "none"). One of the fifty categorical grouping factors U is presented as an example.

Across the fixed effects (β), the standard error of the coefficient estimates is a mean (SD) of 0.0010 (1.1×10^{-4}) larger when incorporating variance over θ for the mixed effects model implemented using `glmmTMB` and `tmbstan`, and 0.00108 (6.3×10^{-5}) larger when using the fixed effects asymptotic implementation. In terms of the intercept and varying-intercept parameters, however, the mixed model implementation has less uncertainty (the difference between "global" and "none" standard errors is 0.000474 for the intercept, 0.00111 for the log of the random effect standard deviation) than the fixed effects model (a mean (SD) of 0.00839 (0.00189) across all 50 dummy variables representing the groups, including the intercept).

This difference for the group intercept is explained by the increased uncertainty between regressions in the coefficient estimates for the dummy variables representing the groups in the fixed effects model due to small group size, while much of the uncertainty around the intercept estimate for each mixed effects regression is accounted for by the random intercept component. Overall, the amount of uncertainty propagated is low: when there is little variance in either the samples or in the regression estimates, less uncertainty exists to be propagated through to the final "global" estimates.

In terms of implementation, the frequentist mixed effects model strategy using MCMC (Avenue 2) takes much longer than the fixed effects asymptotic strategy, especially as the number of draws, chains, and regressions increases. The number of samples from the variational posterior over η (in our case, 100) must be large enough to represent the amount of variance in the posterior; the default in `stm` is 25 [99]. The number of post-burn-in draws from the coefficient posterior (in our case, 4000 per chain) must also be large in order to obtain a sufficient sample. The limiting factor for Avenue 2 implemented as it is here is the effective sample size for the intercept coefficient in the random intercept model, which is lower than for other coefficients. Here, 1000 warm-up draws were discarded for each chain, although plotting the simulated draws suggested that convergence was rapid and fewer than

half as many could have been discarded. We recommend performing a first, preliminary regression to assess the amount of time needed, compare the effective sample sizes, and obtain information about convergence when implementing the "global" or "local" method using MCMC.

The choice of a post hoc regression should not consider only the time it takes to perform the analysis, but the resulting interpretations. In the above example, if we consider the group intercepts to be focus groups, the regression tables result in different primary interpretations of the group-related estimates. For the mixed effects model, one would conclude that the mean document-topic proportion for topic 1 across all focus groups was 0.179 (0.0220), with a standard deviation in mean document-topic proportion between focus groups of 0.151 (holding all other covariates fixed). For a fixed effects model, one would be able to conclude that the mean document-topic proportion for topic 1 in the first focus group was 0.101 (0.0295), and the mean document-topic proportion for topic 1 in the second focus group was 0.0837 (0.0402) lower than in the first focus group, or a mean of 0.0173 overall (holding all other covariates fixed)– and equivalently for all other levels of the grouping variable.

Most importantly, the "global" strategy for both of these models successfully incorporates the variance in the estimated document-topic proportions into the estimates of the regression coefficients. Although we did not provide an example of them here, the same conclusions can be made about the global-grouped and local estimation strategies. We showed in Chapter 5 that the variance of individual document-topic proportions tends to be lower under the global-grouped or local estimation strategies for ν_d than when using the global method, which implies there would be less variance propagated to the final estimates than shown in our example. However, using any of the three strategies will accomplish the goal of variance propagation; our extension to `estimateEffect` which is tailored to grouped documents is fit to use with the mixSTM.

Chapter 7

7 Case study applying the mixSTM to focus group transcripts from the Homelessness Counts study

7.1 Background

The Homelessness Counts study was a Canada-wide initiative to more accurately describe rates and patterns of homelessness [41]. The nationally published point-in-time counts of people experiencing homelessness (PEH) in Canada suggest approximately 30000 people per night were experiencing homelessness across the included communities [62]. However, these counts may not adequately capture all homelessness, particularly among those who do not access homelessness services frequently, are part of a transient population, or are otherwise "hidden" to discrete counting methods [59, 35]. Inaccurate numbers impact the resource acquisition and availability of services for people currently experiencing and at risk of homelessness.

Key to understanding the prevalence of homelessness in Canada is that of leverageable data: what data exists about PEH in Canada and who can access it? In order to explore this question, Homelessness Counts researchers undertook focus groups with stakeholders and service providers in the homelessness sector. In all, 52 meetings were held in 28 communities [43]. These communities were selected to represent a variety of population sizes, remotenesses, and resident demographics; stakeholders with varying roles with respect to the homelessness sector were invited to participate [43]. In virtual focus groups held via Zoom, questions were asked surrounding data administration and the profile of homelessness in the community. Following the focus groups, the Homelessness Counts team visited each community and performed one-on-one interviews with people currently facing homelessness [41, 42].

In the present secondary analysis, a subset of the meeting transcripts representing focus group data are analyzed using a Mixed Structural Topic Model (mixSTM) to obtain additional insights from these discussions and illustrate the mixSTM.

7.1.1 Research questions

Three primary questions arise for this application of a topic model to focus group transcripts:

1. What topics emerge using a topic modelling approach? How do these topics align with the results of previously completed qualitative analyses?

2. Does the prevalence of certain topics discussed in the focus groups vary by location, particularly the remoteness of the community, and how much of the variance in topic prevalence is explained by the grouped nature of the discussion in focus groups? What implications might this carry for the interpretation of the results?

3. How are topics distributed across facilitator questions and can this be used to validate or obtain additional information to inform topic interpretation?

Structural topic models (the STM and the mixSTM) have the ability to pool information about topic prevalence (how much a topic is discussed) across documents with respect to shared levels of a covariate and allow for exploratory post hoc inference on topic prevalence using the incorporated covariates.

To facilitate research question 2 and 3, we wished to incorporate additional variables into the model. On the estimation side, it is likely that topic prevalence is likely related to the direction of the conversation and the experiences of the participants in the group; on the post hoc inference side, we would like to estimate the effect of available variables on the topic prevalence for interpretation and validation purposes. We also expect topic prevalence to differ between focus groups, because they are separate conversations with

distinct members. For the benefit of pooling information about topics and to limit any confounding of other effects in post hoc regressions, we would like to incorporate focus group session as a grouping variable in the analysis. However, inference on each session is not of interest: we are only seeking to account for the variance in topic prevalence between focus groups. As such, the mixSTM was chosen as an appropriate topic model for the analysis of this corpus of semi-structured focus group transcripts.

7.2 Quantitative methods

7.2.1 Preprocessing for focus group transcripts

7.2.1.1 Metadata of interest

One of the questions explored in the Homelessness Counts study was the characteristics of and services for homelessness in rural and remote settings [42]. To quantify the accessibility and rurality of each community where a focus group was held, we used the remoteness index [2, 112], which is an indicator of how accessible a community is on a scale of 0 to 1 (higher indices correspond to more difficulty accessing the community). We hypothesized that documents may have different proportions of topics depending on the remoteness of the community.

Other covariates of interest for the mixSTM estimation were the population of the community and whether the community was in a province or territory, which we also hypothesized might affect topic prevalence. Focus group session was also incorporated as a random effect. Not only are documents inherently grouped within the focus groups, meaning we expected topic prevalence to be related between documents within groups, but we were also interested in quantifying the extent to which the sessions differed in the proportions of topics discussed.

Facilitator questions The facilitator's questions were also hypothesized to affect topic proportions: the prevalence of a topic is likely to be influenced primarily by the direction of the conversation, which is determined by the facilitator and the semi-structured interview guide.

For the Homelessness Counts data, we coded the interviewer text by comparing the questions asked with the focus group guide and tagging each interviewer talk turn which corresponded to a guide question with the question number. This tagging was done blinded to participant responses to limit the amount of outside influence on the categorization, for example, that participant responses may reflect several potential guide questions as they draw connections across the conversation. This tagging scheme separates the conversation into chunks of participant and facilitator talk turns.

The complete focus group guide had 10 questions. In practice, some of the questions elicited similar information: questions 1 and 8 (existing databases), questions 3 and 4 (data sharing), and questions 9 and 10 (opinions on current data systems) were merged in the coding. In almost all focus group sessions, the facilitator additionally asked about facilitating the next stage of the project: when and where to conduct additional interviews with people experiencing homelessness in the community. This was coded as a question despite not being on the guide as it was consistently asked across groups, provided a valuable segmentation of the conversation, and contained additional information about the prevalence of, services for, and management of homelessness in the community from the perspective of the participants. All in all, eight codes were applied to facilitator text representing the focus group questions:

- Question 1 (and 8): How do people locally keep track of who is homeless/local rates of homelessness? If you were looking for data regarding the homeless population in your community, where would you look? What datasets include PEH?
- Question 2: In this community, how would a person newly homeless/newly arrived

homeless be identified?

- Question 3 (and 4): How is information shared across agencies that serve the homeless population? What is the nature of intersectoral data sharing?
- Question 5: Who would have by-name lists? Who would have such lists in common?
- Question 6: In your community, who are the "invisible homeless"?
- Question 7: How would you describe the major subpopulations of people experiencing homelessness in your community?
- Question 9 (and 10): What are the biggest problems with how data is managed regarding the homeless population? What works well with data regarding the homeless population?
- Question 11: Where should we visit in your community to interview PEH, particularly those who might be "invisible"? Are there any service provider contacts we should connect with to understand homelessness in your community?

We initially attempted a different approach which qualitatively coded each of the facilitator's talk turns, but the conversational nature of focus group dialogue made the labelling challenging and the number of unique questions grow rapidly, lessening their utility for the pooling of information. Instead, to ensure consistency across the corpus, only facilitator talk turns corresponding to the above questions were labelled.

7.2.1.2 Document segmentation

In the context of the analysis of the Homelessness Counts focus groups, the incorporation of metadata into the estimation was a key consideration for defining documents. The existing metadata of interest included the facilitator's questions, the population and remoteness index of the community in which the focus group was conducted, and whether the community was in a Canadian province or in a territory. The latter three covariates are all at the focus group level or higher, and thus could be included in an STM for any segmentation of the focus group transcripts. In order to incorporate the facilitator questions, the documents

would have to be at minimum divided by the questions.

We initially considered defining our documents as the speech said by one participant in response to one facilitator question. This definition of documents allows us to incorporate the question into the topic prevalence model, without pooling potentially disparate opinions and experiences from different participants. Because the categorization of the questions is somewhat crude and relies only on the facilitator's text, any tangents or supplementary discussion that occurs between the facilitator's redirection towards the guide questions would be assembled into one document. This would result in documents that are likely to be composed of multiple topics due to their length alone, suiting our multi-membership model. However, in an initial model under this definition, the topics were broad– a fact that was reflected in the top documents, which, even when they had an overwhelming proportion of a given topic, did not always seem to align with just one topic's high probability words or just one topic's initial interpretation. The magnitude of the shared vocabulary across all documents when defined this way made extracting individual insights difficult.

We instead used participant talk turns (the speech of one participant between the speech of any other people) as the documents. These documents are naturally short, as the conversation moves between participants often, but metadata at the level of the talk turn or higher can be easily incorporated into the model. In simulation, we observed that the fit of the mixSTM was poorer when the documents contained fewer words (Chapter 4). However, the available covariate information can help the mixSTM fit to smaller documents, given some longer documents which will set out guiding means for the topic weights. It should be noted that even on small documents, the mixSTM assumes that documents are composed of multiple topics in different proportions. This is not an unreasonable assumption: even within a talk turn, a participant may discuss several ideas. However, the short documents may affect the number of topics that can be estimated.

We labelled talk turns in the following manner. First, the text was cleaned and preprocessed

according to the procedure described in the next section. Then, all segments with no words after removing stopwords (words which have function in text but are not useful for inferring meaning, such as "the") and all segments spoken by the interviewer with fewer than three words after stopword removal were excluded. In practice, facilitator text with at least 4 words occasionally was used to ask a focus group guide question and thus created natural breaks; shorter facilitator segments were more likely agreement, repetition, or simple clarifying questions. Finally, all sequential text by one person was combined into a document, creating the talk turns.

We excluded all facilitator talk turns from the corpus used in the analysis.

7.2.1.3 Text cleaning

The Homelessness Counts project involved many meetings to coordinate the focus groups and to obtain information about the communities. Of all the transcripts of these meetings, "focus group" transcripts were considered to be the subset that: 1) asked more than two of the focus group guide questions; 2) had at minimum two participants; 3) were not labelled as initial meetings with potential participants or follow-up meetings with the same participants; and 4) were transcribed verbatim.

The excluded documents did contain information on the theme of homelessness in Canada but were difficult to integrate with the proposed analysis and did not represent the intended corpus. Of the exclusions, one focus group was not able to be transcribed due to technical issues. In seven meetings, only one participant attended. Since we sought to illustrate the mixSTM in the context of focus groups specifically, we excluded transcripts of single-person interviews. In five meetings, the conversation contained two or fewer questions corresponding to the semi-structured guide. The remaining exclusions were initial meetings or follow-ups.

Before creating the documents, we removed all introductory materials before the first focus

group question was asked. This is a common approach in topic models of interviews or focus groups and restricts the content to that which is research-relevant (e.g., [84, 22]). We also removed all notes about participant behaviour and tone: in our corpus, these notes were rare and we removed them all to avoid conflating facilitator/notetaker observations with the words used by participants. Names, organizations, and identifying details were previously redacted during the primary analysis.

In many topic models, words only used rarely in the corpus can have a destabilizing effect on the topics because of the limited co-occurrence information. In the (mix)STM, covariates provide additional information about the mean proportions of topics within documents, and the SAGE model on topic content has been shown to improve the allocation of rare words to topics [99, 38]. As such, we removed no rare words. High frequency words are often removed when they are shared across a large majority of documents because their co-occurrences are less informative for resulting topics. However, in a small corpus (in terms of the overall number of words) this can reduce the size of documents. We thus elected to remove only the top 20 words (0.35% of all unique non-stopworded words before stemming) which dominated the topics in preliminary models.

All numerals, punctuation, and one-letter words were removed from the corpus and the text was converted into lowercase. We standardized some phrases (e.g., "LGBTQ+", "2SLGBTQ" were standardized to "LGBT"; "by name" and "by-name" lists were converted to "byname"). Stopwords from the `removeWords` function in the `tm` package in R [40] were removed along with several interjections such as "um" and "yeah". Any resulting documents with five or fewer words were removed, as the co-occurrence rates within these documents were not likely to be useful. We acknowledge that six-word documents represent an arbitrary minimum, but some minimum length requirement was necessary and from scanning the documents, the original text of documents longer than five words after stopword removal had potential to be content of interest. Stemming was performed on the

resulting document-term matrix after all other exclusions with the `quanteda` inbuilt stemmer [8], which trims endings of words in an attempt to reduce them to their root form (e.g., making equivalent "read" and "reading").

7.2.2 Applying the mixSTM to focus group data

7.2.2.1 Choosing K

Using `searchK` from the `stm` package [99], we scanned every K from 6 to 30 topics, followed by a qualitative appraisal of topics from favourable solutions. We chose this range looking at our corpus: there were 8 facilitator questions coded, but these questions were coarse delimiters of the conversation content. This range of topics seemed plausible in that it contained fewer than 8 topics at minimum and extended to a range that would likely be unsupported by the amount of data.

For `searchK`, we set 0.5 as the proportion of held-out words and 10% as the proportion of held-out documents for calculation of held-out likelihood. We used the top 10 words in each topic to compute semantic coherence and exclusivity scores. We specified a mixed effects model on topic prevalence but no model on topic content, as our primary interest was in the potential modification of topic prevalence and the exclusivity metric provided in `searchK` cannot be calculated for STMs with a topic content covariate.

The "best" number of topics would result in a high exclusivity, semantic coherence, and held-out likelihood, and low residual dispersion. Looking in tandem at all measures from the `searchK` results, we selected specific values of K to examine more closely by running a full mixSTM. For these models, we considered the top words (in highest probability and most exclusive when weighted by frequency (FREX, [10])) and looked at the documents with the highest document-topic proportions for that topic. To get more context for the use of a topic, we also looked at documents which shared intermediate document-topic proportions of several topics. Of interest was the apparent interpretability and quality of

the topics in terms of whether they identified a cohesive theme. We selected our final number of topics based on this evaluation.

7.2.2.2 Running the mixSTM

Once we selected the number of topics, we ran a full mixSTM analysis. We specified a mixed effects formula on topic prevalence, with fixed effects corresponding to the remoteness index and the logarithm of the community's population (continuous) and whether the focus group was held in a province/territory (categorical). We included focus group as a random intercept (as the baseline rate of the prevalence of topics was expected to vary between focus groups) and facilitator questions as random slopes (allowing the impact of the question on the topic prevalence to vary between groups). Using a categorical variable as a "slope" defines several random intercepts per group and allows for a correlation to be modelled between the levels of the variable. This random effects structure models our beliefs about the focus group data: the topics in a given participant's talk turn depend on the conversation as a whole and also on the question to which they are responding. By additionally allowing correlations between the amounts that different questions affect topic prevalence, this model maximizes the amount of information available to be pooled across documents. We elected to estimate one global covariance (rather than group-by-group topic covariance matrices), because we were already specifying a complex model on the mean and there was unlikely to be a benefit to the fit.

The corpus vocabulary was not very large and including a content covariate limited the interpretability of the topics, but the high probability documents using an LDA implementation on the topic content model were less cohesive to the themes. We thus elected to use the covariate-free SAGE model [99], which is more adept than the default model at handling rare words, but does not attempt to split the vocabulary of each topic between levels of a covariate.

To initialize the model, we used the "Spectral" option in the `stm` package [99]. We spec-

ified a Horseshoe prior for the fixed effects in our penalized mixed effects model on topic prevalence with a 200-iteration maximum for the M-step regression model fit. We kept the default convergence tolerance of 1×10^{-5} and set a maximum of 500 EM iterations.

7.2.2.3 Estimating the covariate effects

We explored the association of covariates of interest on some key topics with three primary goals. First, we were interested in the association of rurality with the various discussions that occurred in the focus groups: once adjusted for the differences between focus groups, did more remote communities weigh in on topics in different amounts than more accessible communities? Second, we included facilitator question as a deviation-coded categorical variable in this post hoc regression to attempt to validate our interpretation of the topics using an external approach: did our interpretations of topics align with the expected segment of the focus group discussion? Third, we were interested in the grouping by focus group session, since the proportion of a topic in a talk turn in any given focus group is expected to be intrinsically related to the amount that the topic is discussed in other talk turns in that focus group: did different focus groups have a different average prevalence of topics? We also included the population size of the community (on the log scale) in the regression. For the purposes of validating the topics (which are shared by the whole corpus), we used facilitator question as a fixed effect.

We used the "global" method for posterior sampling (which computes an average approximation to the topic covariance) to propagate the uncertainty in the document-topic proportions to final coefficient estimates. We implemented a linear regression with an identity link using a fully Bayesian approach in `brms` [21], with default priors on the variance components and Normal(0,10) priors on the fixed effects coefficients and intercept. We took 25 draws from the variational posterior of the topic proportions and used 1000 post-warmup draws from four chains of a regression on each to obtain final estimates of coefficients and variance components with their sample standard deviations.

We present these and the corresponding t-value for the coefficient estimates. For the variance components, we present the variance of the focus group random intercept and residual variance on the standard deviation scale, along with the amount of residual variance explained by the variation within focus groups ($\sigma^2_{fg}/(\sigma^2_{fg} + \sigma^2_r)$). Questions were labelled according to their place on the guide (see section 7.2.1.1).

7.3 Results

7.3.1 Describing the corpus

Our corpus consisted of 29 focus group transcripts from sessions conducted in 18 communities across 10 provinces and territories. The focus groups had a mean (SD) of 5.31 (2.59) participants and 5.78 (1.84) unique focus group guide questions were tagged in each focus group.

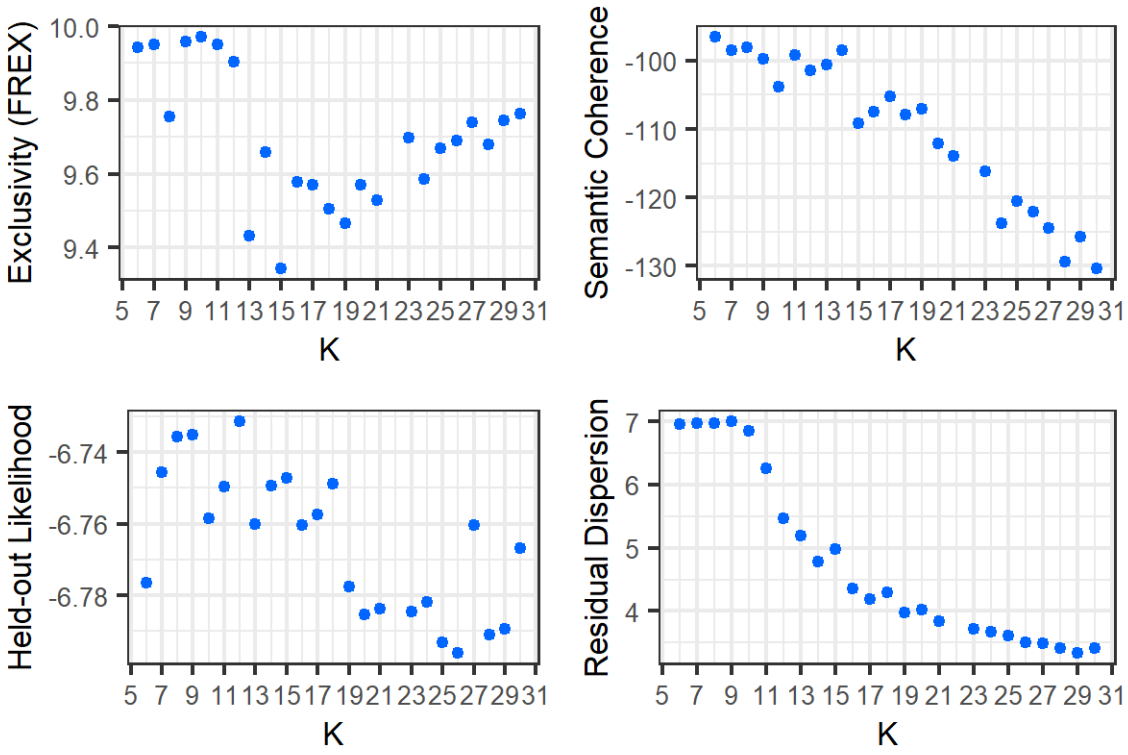
Before any data cleaning, our corpus had 8572 transcript segments across all participants and facilitators and an initial vocabulary size of 5842 words (redacted information excluded). These segments were a median (Q1, Q3) of 11 (2, 32) words long. The most common words (excluding stopwords) in the corpus were "people" (1854 uses), "know" (1836), "just" (1235), "think" (941), and "homeless" (867). We removed these and the next fifteen most frequent words (*can, get, one, say, kind, lot, go, housing, data, really, community, also, going, want, us*). After stopword and high-frequency word removal, the vocabulary was 5620 words, with each word being used a median (Q1, Q3) of 2 (1, 7) times and a maximum of 415 times.

After excluding all interviewer speech, removing all talk turns with no words after stopword removal, and combining adjacent speech by the same participant, the corpus was 1662 documents long, each with a median of 17 (6, 33) words. After excluding 370 documents with five or fewer words, the final corpus was 1292 participant talk turns long, each with a

median (Q1, Q3) length of 23 (13, 38) words. These documents represented a pre-stopword removal median length of 90 (52, 116) words. The final vocabulary size after stemming was 3140. In total, there were 39191 unigrams in the corpus.

7.3.2 Selecting K and running the model

Figure 7.1: Exclusivity, Semantic Coherence, Partial Held-Out Likelihood, and Residual Dispersion for 6-30 topics.



Exclusivity and semantic coherence were calculated from the top 10 words. Partial held-out likelihood was calculated from holding out 50% of words in 10% of documents. $k = 22$ is not presented due to convergence issues.

Figure 7.1 plots the exclusivity, semantic coherence, held-out likelihood and residual dispersion against the number of topics. With 12 or fewer topics, the exclusivity of the topics (when considering the top 10 words) is highest; it decreases sharply at 13 topics and then gradually climbs again. The semantic coherence of the topics also appears to be relatively high for fewer than 14 topics with a sharp decrease after 19 topics. The held-out likelihood is relatively similar for all solutions, but particularly large for those between 7 and 18

topics. The amount of residual dispersion decreases as the number of topics increases, particularly with more than 10 topics with a slowed trend after 16 topics. From the `searchK` results, we see that our corpus probably supports few topics, which makes intuitive sense as our documents are short in length.

We considered some models in further depth. We first selected a mixSTM with 11 topics and one with 17 topics. A model with 11 topics would appear to have good semantic coherence, highly exclusive topics, and a middling held-out likelihood, but very high (although not the highest) residual dispersion. Choosing few topics often results in topics that are broad: we thus also wanted to consider a higher granularity to see if it could improve the distinction between themes. As such, we chose a model with 17 topics, which, of the models with more than 15 topics (which had the lowest residual dispersion), had the highest semantic coherence for the top 10 words, a typical exclusivity, and a better held-out likelihood than any model with more topics. Based on these models, we additionally contrasted the results with a mixSTM with 14 topics (lower residual dispersion than the 11-topic model, but similar semantic coherence and held-out likelihood at the cost of lower exclusivity) and one with 27 topics (low residual dispersion and semantic coherence, with relatively high exclusivity and moderate held-out likelihood).

The model with 11 topics had highest proportion of distinct, high-quality topics when looking at both the words and documents assigned to them. However, some of the topics seemed to cover several concepts that were distinct in other models. The model with 17 topics achieved more specific topics, at the cost of having a larger number of less cohesive topics. The 14-topic model had fewer hard-to-define topics than the 17-topic model, but had some that covered overlapping concepts like the 11-topic model, and the 27-topic model achieved similar granularity to the 17-topic model but with more document intrusion. Interestingly, some topics emerged distinctly in all models (such as “sharing of information”), some topics emerged but always in conjunction with other topics (e.g., “evictions/tenancy”), and

some topics emerged distinctly in only a subset of the models (e.g., the 11- and 14-topic models addressed couch surfing, while the 17-topic model did not).

The model with 17 topics was chosen as the preferred model and was used in all subsequent analyses.[1] Although it had some ill-defined topics, the quality of the topics that were specific was generally high. This also suited our a priori interest in exploring the relationship between covariates and topic prevalence: an analysis that used broader topics would be less likely to yield a meaningful exploration of differences.

The mixSTM with 17 topics converged on iteration 9 after approximately 16 minutes. During fit, some iterations of the M-step had issues inverting a matrix during an intermediate step of the streamlined mixed effects regression fit. Approximate solutions were found rapidly and only the fitted values from the regression are used downstream, so this is expected to have had minimal impact on the results. In one of the M-step mixed effects regressions, the maximum allowed number of iterations was reached before the mixed effects model converged. Since this iteration was intermediary and thus unlikely to change the results for the interpretable topics, we did not halt the model.

7.3.3 Topic interpretation

We divide the topics by quality for interpretation. For each topic, we present the highest probability ("Highest Prob") words and the most exclusive words (weighted by frequency [10], "FREX") to the topic. For high-quality topics, we additionally present selected quotes from documents with a high proportion of the topic in question.

7.3.3.1 High-quality topics

These topics were considered of higher quality in that many of their high probability words formed a semantically related set, and all or nearly all of the documents with a high pro-

[1] The details of the unselected model with 11 topics (in terms of high probability words) are provided in Appendix A.5.

portion of these topics were clearly related to the identified theme.

Topic 2: Youth homelessness and intersections with school systems.

- Highest Prob: youth, famili, kid, home, school, children, servic, support, thing, hidden, homeless, come, popul, educ, time, system, stay, need, now, tran
- FREX: youth, famili, school, kid, children, home, hidden, educ, tran, popul, system, lgbt

Topic 2 surrounds youth in homelessness and how the homelessness sector interacts with school systems. Top words in this topic were related to youth (kids, children, school, educ) and families (famili, home), along with more general and less exclusive words like "support" and "services". Interestingly, "hidden" also came up as a high probability word in this topic, suggesting that not only is youth- and family-homelessness an issue, but that it was discussed in conjunction with homelessness that may be invisible to traditional data collection. Documents with high proportion of this topic identified schools as a connection point for youth and family homelessness, but also about other services that do (or don't) exist for youth in the community.

[0525] "For us as the [organization] we had an LGBT youth group for a number of years and a lot of folks came through that youth group, and then we were able to identify the fact that within that youth group over half that youth group were trans. And they were trans in need of a lot of services, including housing."

[0032] "So two of the Catholic schools here in [city] have implemented a pilot project where they have a community worker within the school system and that was actually a really great relationship actually connecting them with the [agency] because that would be affecting families."

Topic 4: Sharing information among service providers.

- Highest Prob: inform, share, agenc, consent, access, hifi, individu, system, client, servic, file, work, time, someon, make, need, sign, thing, challeng, around
- FREX: share, consent, inform, agenc, hifi, file, individu, access, system, sign, client, barrier

Topic 4 identifies high probability words relating to sharing of information (share, consent, access, inform, sign) of clientele (individuals, clients) between agencies, potentially electronically (HIFIS[2], systems, files). Documents with a high proportion of this topic described various ways in which agencies collaborated (or elected not to) with the sharing of information on PEH. In particular, stakeholders were quick to identify barriers to sharing information, including obtaining informed consent and maintaining privacy and respect for autonomy. Others emphasized the redundancy of the way information is stored under current practices where sharing is not accessible.

[0837] "Yeah, I mean I think that there's definitely that piece around sharing information. I do think HIFIS has definitely helped with sharing that information. I think the problem is that we don't have– I know us as [agency], we have different kind of guidelines around privacy, but I just think that if there was a standard around information sharing and a practice that was consistent to benefit the individual and a dialogue with the individual about "here's your circle of care, here's what information would be shared"."

[0119] "We don't actually share information with other systems, like it's not our business to be sharing information of the people who walk through our doors. The times where we are providing support for housing it's often to hold systems accountable for the violence or harm that they're putting on our people and so those are the times when we have to come in and get the consent form signed to step in to support folks, but there's not a ton of sharing, "oh this time this person was at this place" and because what it feels like to me, is we're then holding people who are accessing services accountable. "Why didn't you go to that place? Remember when they offered you food, or a house, you didn't follow through." That's not what it's about."

Topic 5: Community locations where homelessness is observed.

- Highest Prob: sleep, encamp, tent, team, live, park, citi, hard, thing, area, rough, servic, come, time, two, need, take, coupl, car, stay

[2]HIFIS is the Homeless Individuals and Families Information System [61].

- FREX: encamp, sleep, tent, team, park, rough, hard, car, citi, live, vehicl, anywher

Topic 5 related to the geography and mobility of PEH. High probability words focused on sleeping arrangements and locations (sleep, encamp, tent, park, rough, area, citi, live, car, stay), and the outreach teams that interact with those sleeping in tents or in the rough (team, service). Vehicle homelessness and people who sleep in their cars are also discussed in documents with a high proportion of this topic.

[1103] "I mean it's a vast area. [. . .] like there's no mode of transportation, public transportation as of yet. How do people get from point A to B? Where are they staying? What are they staying in? Just the conditions and just even how far people have to travel to do things."

[1073] "Yeah, they'll sleep in stairwells, ATMs, there was a big spot under a bridge at one point this summer. Just kind of basic. They'll camp out in encampments."

[1233] "There's an outreach team [. . .] their sole focus is to be actively searching for people sleeping in the rough and connecting with them out in their camps and the community."

Topic 6: Comfort and safety of people with lived experience of homelessness.

- Highest Prob: live, feel, homeless, experi, definit, safe, talk, indigen, place, servic, differ, popul, thing, identifi, access, time, experienc, need, system, come
- FREX: feel, experi, live, safe, homeless, definit, indigen, talk, popul, experienc, place, identifi

Topic 6 is about the feelings, comfort, and safety of groups and individuals with lived experience of homelessness. Words relating to subjective experiences (feel, safe, experi, experienc, differ, identifi) and to subpopulations (indigen, popul) appeared with high probability in this topic. Documents with a high proportion of this topic encompass discussions of the definitions of homelessness and feelings of cultural appropriateness of services to populations, particularly Indigenous peoples.

[1286] "I think that that's probably one of the challenges a lot of folks have. If it's not necessarily

articulated that way I think that's at least a portion of why, you know it doesn't really feel accurate or like it's taking into account our specific experiences [. . .] I think it just goes against our way of thinking and way of being is just to quantify somebody as a number and I don't think that that feels good to anyone, but it's especially kind of triggering for a lot of folks in the Indigenous community."

[0124] "For a lot of the folks that we work with they would be 100% houseless but they would not be homeless. Because at the end of the day, they have a place where they can go where they feel loved, and they feel secure, and they have people they can trust, and they have a care system and a care network in place, as tenuous and as fragile as that might be in some situations [. . .] the definitions of houselessness and the way that we provide resources to people are based on very white colonial models of what is and what can be funded under houselessness and homelessness. So a lot of the times that doesn't reflect Indigenous realities and that does not reflect queer realities of community building."

Topic 12: Using databases to track and report on homelessness.

- Highest Prob: use, system, hifi, client, report, manag, case, track, agenc, access, inform, databas, shelter, hmis, address, servic, pull, intak, tri, time
- FREX: use, report, system, manag, hmis, case, hifi, track, client, address, databas, agenc

Topic 12 has words that relate to databases that services and stakeholders use to collect data on homelessness in the community (system, HIFIS, client, case, agenc, database, inform, HMIS[3], intake) and what is done with these databases (report, manag, track, access, pull). Documents with a high proportion of this topic mainly described the use and collection of data; documents generally did not attribute sentiment, and were purely descriptive.

[0922] "It's something that we've created, we don't currently have our own data management, case management software, I know like [agency] uses [software] and [agency] uses [software], I believe, but we are currently just like operating off our own little system."

[3]HMIS refers to Homelessness Management Information Systems [61].

[1117] "So we can use that to cross reference between the reports that we're pulling from HMIS and then our [software] to see, you know, what overlap exists, and then, who is not accessing or in HMIS and kind of who this population is what their needs are. [. . .] I kind of send it off to our data analyst, he kind of pulls all that data and builds those reports."

[0100] "So, and then we report on that as part of our annual report, which is to the public. But we don't use HIFIS."

Topic 13: By-name lists and coordination around these lists.

- Highest Prob: list, coordin, bynam, probabl, access, hifi, homeless, servic, polic, thing, agenc, now, two, time, need, program, way, engag, tri, year
- FREX: list, bynam, coordin, probabl, hifi, access, polic, homeless, agenc, engag, two, touch

Topic 13 discusses the presence or absence of by-name lists (list, bynam) often in conjunction with coordinated access through HIFIS (coordin, access, HIFIS, service, agenc). In some documents, other lists surrounding information on people experiencing homelessness are also discussed.

[0826] "I don't even know if they're doing the by-name lists now or it's essentially every single homeless person in the city of [city]. I think we had the by-name list and then HIFIS came in and the by-name list kind of was antiquated because HIFIS was so much more."

[1099] "With coordinated access, the ideal is that it's all of our agencies are going to be brought on to HIFIS eventually. It's been a process."

Topic 14: Addictions, health, and mental health issues contribute to houselessness.

- Highest Prob: health, mental, issu, addict, home, care, client, back, homeless, servic, come, trauma, thing, need, place, code, program, way, tri, year
- FREX: health, mental, addict, issu, care, home, client, back, trauma, homeless, code, place

Topic 14 relates to the intersection of (mental) health and addictions with homelessness

(health, mental, issu, addict, care, trauma). Documents in this topic discuss ways in which the healthcare system, mental health and addictions services, and the justice system contribute to homelessness in their communities, or when and how service providers in these sectors interact with individuals experiencing homelessness.

[0409] "Trauma, trauma, trauma. And trauma looks like addictions, trauma looks like mental health, trauma looks like not being able to maintain housing, trauma looks like the lack of social skills. Trauma looks like all these things that are acting as a barrier to people being able to attain or maintain housing."

[0553] "We do have a housing program that works with people with mental health and addictions issues in terms of housing and we have some private operators, community care homes, specific to mental health and addictions [...] and of course for more physical or intellectual needs we have personal care homes and long term care homes [...] sometimes unfortunately they leave those community care homes and go into an emergency shelter, because of issues they're struggling with."

Topic 16: Vulnerable populations' intersections with the shelter system.

- Highest Prob: women, shelter, indigen, popul, men, see, violenc, use, servic, place, come, now, homeless, time, thing, children, need, town, singl, drug
- FREX: women, shelter, men, indigen, violenc, popul, children, town, use, see, drug, singl

Topic 16 discusses major populations in the shelter system (women, indigen, popul, men, children, singl) as well as the services they tend to use (shelter, use, servic, place, come). The violence and abuse that vulnerable PEH face in and outside the system is also discussed in documents with a high probability of this topic.

[0833] "I think major issue in [city] is our Indigenous population is disproportionately represented in our shelter. We know about a third of shelter use in [city] is for individuals connected to Indigenous communities around [city], and they make up less than 3% of our city's population."

[0205] "I noticed that you know there's a lot of services for ladies and you know and children in

town and there's, you know the shelter basically is the only service for men. Right? Right. So, like I mean, like, you know, the ladies, in particular, go to the [Organization] I believe that's their probably their first stop in the transition itself."

7.3.3.2 Lower quality topics

These topics have some themes that could be identified using high probability words and documents but have word or document intrusion that makes their direct interpretation difficult (i.e., some high probability words and documents with high proportions of these topics did not appear to correspond with the otherwise apparently semantically coherent topics). We do not provide exemplary documents for these topics to emphasize the variety of themes discussed in talk turns with high proportions of these topics.

Topic 1: Collecting data and current systems.

- Highest Prob: actual, shelter, now, databas, come, number, use, hifi, abl, track, work, way, thing, emerg, program, time, need, servic, homeless, releas
- FREX: actual, databas, number, shelter, track, now, hifi, emerg, releas, use, feder, come

Topic 1 relates to current practices (actual, now) and how databases are used (databas, number, HIFIS, use, abl, track, work), particularly in shelters and the shelter systems (shelter, emerg). This theme overlapped somewhat with Topic 12, although generally documents with a high probability of this topic appeared less specific.

Topic 3: Evictions and tenant-landlord contacts; data issues when it comes to Indigenous peoples.

- Highest Prob: see, talk, evict, hear, number, thing, tenant, landlord, senior, notic, indigen, now, veteran, high, rent, spdat, client, homeless, move, time
- FREX: evict, talk, see, tenant, hear, landlord, notic, veteran, senior, number, spdat, rent

Topic 3 mainly relates to tenancy (evict, tenant, landlord, notic, rent, move), but has some

verbs related to observations by the stakeholders (see, talk, hear, notic) and also identifies some additional populations (veteran, senior, indigen). Some documents in this topic relate to Topic 6, about the cultural appropriateness of data collection about Indigenous peoples.

Topic 7: Case management for people experiencing homelessness.

- Highest Prob: work, connect, individu, program, access, care, adult, shelter, might, servic, thing, organ, name, time, young, case, age, need, street, agenc
- FREX: work, connect, adult, care, individu, access, young, name, program, case, age, shelter

Topic 7 has words that relate to typical service provision (work, connect, program, access, shelter, service, organ, agenc) and to the people they serve (individual, care, adult, name, age). Documents in this topic discuss the agencies' interactions with one another and with clients or cases. A subset of documents discuss youth transitioning to adulthood and adult services.

Topic 8: Making new contacts with people and organizations.

- Highest Prob: pretti, come, close, team, see, hotel, let, guy, servic, open, new, contact, thing, person, client, now, th, long, time, shelter
- FREX: pretti, close, hotel, guy, let, team, th, contact, come, open, busi, new

Topic 8 relates to making contact with organizations and people. Some key words relate to people (team, guy, person, client), to forming contacts (contact, come, see) and to openings/closures (open, close) some of which appear to be in relation to COVID-19. "New" in context in the documents is often used in conjunction with "contact", describing how a service provider or stakeholder might first become aware of someone newly experiencing homelessness in the community. More documents with a high proportion of this topic simply discuss common interactions and locations in the community.

Topic 9: Observations about service utilization

- Highest Prob: find, come, even, often, homeless, mani, servic, guess, better, thing, food, see, time, sometim, need, call, program, way, tri, worker
- FREX: find, even, often, come, better, food, guess, homeless, worker, understand, individu, agenc

Topic 9 has many qualifying words (even, often, many, better) and services (service, food). In conjunction with the verbs (find, come, see). Among the documents with a high proportion of this topic, some discuss meal programs but others relate to observations the stakeholders have made about PEH and relationships with other providers and PEH in the community.

Topic 10: Connections with invisible homelessness and access to services

- Highest Prob: mean, servic, may, invis, access, certain, homeless, probabl, within, come, thing, communiti, get, interview, time, need, program, abl, way, tri
- FREX: mean, invis, may, servic, certain, interview, within, access, probabl, communiti, appli, homeless

The most unique word present in Topic 10 is "invis", relating to the question of invisible and hidden homelessness, and what services people who would otherwise be invisible might access (may, probabl, certain, service, access, community). The high probability word "interview" also relates to this theme, as researchers were seeking out PEH who might be missed through typical data collection to interview. However, many of the documents with a high proportion of this topic do not fit the theme and relate instead to general observations about services in the community or the geography and mobility of PEH. This topic was low-quality.

Topic 11: Point-in-time counts and questions/answers from PEH.

- Highest Prob: question, day, now, ask, thing, year, count, servic, see, answer, run, time, last,

call, need, program, way, tri, said, collect

- FREX: question, day, ask, now, count, answer, year, run, thing, last, five, collect

Topic 11 contains words related to questions (question, ask, answer, count, collect) and to time (day, now, year, time). The co-occurrence of "time" and "count" leads to many of the high probability documents being about the point-in-time counts of the homeless population, which relates the question-asking component (for data collection) to the time component (when these things occur) in the topic.

Topic 15: Renting and system timings.

- Highest Prob: month, come, mayb, time, shelter, rent, thing, referr, everi, number, pay, end, two, move, three, need, last, now, servic, program
- FREX: month, rent, mayb, referr, come, pay, move, three, number, last, everi, two

Topic 15 has high probability words that relate to timing, including some about frequency (month, time, every, pay, end, last), as well as about rent (rent, pay, move) and referrals. The documents which have a high proportion of this topic discuss rentals from the tenant perspective including difficulty paying, but many also discuss the time constraints of the system of shelters and services.

Topic 17: The "soaker"

- Highest Prob: work, use, servic, come, shelter, thing, see, homeless, access, time, need, now
- FREX: use, work, shelter, system, access, client, now, homeless, agenc, individu, come, servic

Unlike the others, this topic had a maximum document-topic proportion of 0.12. Essentially, its place is in soaking up the remaining words that were not able to be meaningfully attributed to other topics.

7.3.4 Estimate Effect

We explored three topics with specific interest in whether we could demonstrate an association between the prevalence of topics and remoteness or population: Topic 2 (youth homelessness and school systems), Topic 3 (evictions/tenancy and other issues), and Topic 6 (comfort and safety of PEH). Although we specified this comparison a priori, our ability to estimate the association between remoteness and the topics was limited by the specificity of the topics at hand. For example, Topic 12 (using databases to track and report homelessness) was expected to be discussed in each community, no matter the remoteness, because it appeared to address a question from the focus group guide. On the other hand, more emergent topics like Topic 2 were not prompted by the guide and might vary in prevalence based on characteristics of the community and the representatives of in the community in the focus groups.

We used a post hoc regression to validate our interpretations of some topics which seemed to correspond to questions on the facilitator's guide: Topics 4 (sharing information), 12 (using databases), and 13 (by-name lists and coordinated access). We also performed a regression on some of the lower quality topics (Topics 1, 7, and 8) to attempt to obtain additional evidence for their interpretation in light of which questions they tended to follow.

We use the mean population (212077 residents) and mean remoteness index (0.245, an "accessible area" [112]) of our sample of communities to provide an interpretation of the baseline prevalence of each of these topics in "the average community". However, we note that no community corresponding to these exact values was considered in our sample.

Table 7.1: Covariate associations with subpopulations and causes of homelessness. Table 7.1 presents the results of regressions on the topic prevalence of Topic 2 (youth and family), Topic 3 (evictions and tenancy), and Topic 6 (comfort and safety of PEH). The

Table 7.1: Regression results on talk turn-topic proportions for Topic 2 (Youth/family), Topic 3 (Evictions and other issues), and Topic 6 (Safety and comfort of PEH).

	Topic 2		Topic 3		Topic 6	
	Coefficient (SD)	t	Coefficient (SD)	t	Coefficient (SD)	t
Intercept	.048 (.055)	.857	.099 (.060)	1.64	.035 (.063)	.551
Log popn	-.0007 (.0091)	-.076	-.0069 (.0098)	-.699	.0038 (.010)	.371
Remoteness	.0024 (.062)	.039	-.051 (.069)	-.803	.023 (.073)	.313
Question 1	-.0068 (.010)	-.671	-.011 (.011)	-1.01	-.014 (.011)	-1.20
Question 2	-.0091 (.017)	-.534	-.0053 (.018)	-.292	-.014 (.019)	-.761
Question 3	-.0021 (.013)	-.166	-.015 (.012)	-1.22	-.013 (.014)	-.924
Question 5	-.0030 (.023)	-.128	-.017 (.025)	-.696	-.021 (.027)	-.792
Question 6	.022 (.013)	1.75	-.00090 (.011)	-.080	.024 (.015)	1.66
Question 7	.015 (.019)	.770	.022 (.026)	.854	-.0028 (.024)	-.115
Question 9	-.018 (.019)	-.910	.021 (.021)	-1.01	.019 (.022)	.879
	Var components (SD scale)	% Explained	Var components (SD scale)	% Explained	Var components (SD scale)	% Explained
Intercept	.012 (.0078)	.95%	.019 (.0094)	1.98%	.017 (.0091)	1.26%
Residual	.13 (.0083)		.13 (.0058)		.15 (.0081)	

intercept of the regression on the document-topic proportions for youth and family homelessness was 0.048 (0.055), corresponding to a mean proportion of 0.044 for the topic of youth homelessness in a talk turn in an average community. Similarly, the mean document-topic proportion across all questions and focus groups in a community with an average population and remoteness index among our sample was 0.049 for Topic 3 and 0.061 for Topic 6. In each regression, differences between focus groups in the amount a given topic was discussed accounted for between 1 and 2% of the variance in the topic proportions, with the most variation in the amount Topic 3 was discussed (2.0%).

In these three regression analyses, we were unable to demonstrate an association with remoteness index or population as the variances around the coefficient estimates were quite large. Of the three, Topic 3 had the largest potential association relative to the amount of variance: there was a small negative trend in the amount evictions/tenancy was discussed as remoteness increased.

For Topic 2, participant responses following the facilitator asking about invisible home-

Table 7.2: Regression results on talk turn-topic proportions for Topic 4 (Sharing information), Topic 12 (Using databases), and Topic 13 (By-name lists and coordinated access).

	Topic 4		Topic 12		Topic 13	
	Coefficient (SD)	t	Coefficient (SD)	t	Coefficient (SD)	t
Intercept	-.029(.084)	-.340	.047 (.089)	.535	.090 (.053)	1.69
Log popn	.022 (.014)	1.56	.013 (.014)	.915	-.0056 (.0086)	-.645
Remoteness	-.00058 (.098)	-.006	-.085 (.11)	-.806	.0033 (.065)	.0500
Question 1	-.011 (.013)	-.866	.074 (.014)	5.40	-.012 (.010)	-1.17
Question 2	.0058 (.024)	-.242	-.0044 (.025)	-.178	-.014 (.017)	-.823
Question 3	.090 (.019)	4.79	.024 (.020)	1.22	-.0018 (.013)	-.134
Question 5	-.018 (.031)	-.583	.058 (.049)	1.19	.123 (.033)	3.76
Question 6	-.042 (.013)	-3.28	-.042 (.015)	-2.88	-.017 (.010)	-1.67
Question 7	-.046 (.021)	-2.24	-.038 (.025)	-1.54	-.024 (.018)	-1.36
Question 9	.055 (.028)	1.95	-.033 (.026)	-1.27	-.036 (.017)	-2.10
	Var components (SD scale)	% Explained	Var components (SD scale)	% Explained	Var components (SD scale)	% Explained
Intercept	.033 (.0096)	4.29%	.028 (.011)	2.41%	.012 (.0071)	.88%
Residual	.16 (.0065)		.18 (.0075)		.13 (.0062)	

lessness (Question 6) discussed youth/families 2.2% (SE: 0.013, t =1.75) more than the average document-topic proportion across all questions. For Topic 6, the mean (SD) topic proportion was 2.4% (1.5%, t=1.66) larger than the grand mean topic proportion in talk turns preceded by Question 6.

Table 7.2: Covariate associations with question-attributable topics. Table 7.2 presents the results of regressions on the topic prevalence of Topic 4 (sharing information)[4], Topic 12 (databases), and Topic 13 (by-name lists and coordinated access). In the average community, the mean topic proportion across all documents was 0.086 for Topic 4, 0.097 for Topic 12, and 0.061 for Topic 13. The standard deviation of the difference between focus groups in the average proportion that Topic 4 was discussed was 3.3%, which explained 4.3% of the variation in the document-topic proportions; in Topic 12 the between-focus

[4]Note on negative proportion and interpretation: Since we did not center our continuous variables, this intercept corresponds to the proportion of the topic for the average question and at the grand mean over focus groups when the log population is 0 (1 person) and the remoteness index is 0 (totally urban). Obviously, this is an impossible scenario (much like a negative proportion). However, this is a limitation of using a linear regression to model proportions.

group differences explained 2.4% of the variance, and in Topic 13 they only accounted for 0.88% of the variance. When the population increased by a factor of 10, the proportion that Topic 4 was discussed in any given talk turn (averaged across all questions) increased by 2.2% (SD: 1.4%, t=1.56), holding all other variables constant. The other two topics did not show strong associations with population size or remoteness index.

The mean document-topic proportion of Topic 4 increased by 9.0% (SD: 1.9%, t=4.79) following the facilitator asking Question 3 (data sharing), holding other variables constant. The mean proportion of a document devoted to Topic 4 also increased following Question 9 (opinions on current data systems), after which it was discussed 5.5% (SD: 2.8%, t=1.95) more on average. The amount this topic was discussed decreased by 4.2% (1.3%, t=-3.28) following Question 6 (invisible homelessness) and by 4.6% (2.1%, t=-2.24) following Question 7 (major subpopulations of PEH).

Topic 12 was strongly associated with Question 1: talk turns following Question 1 (data collection/management) discussed Topic 12 an average of 7.4% (SD: 1.4%, t=5.40) more than the global mean amount per talk turn. This topic also showed some evidence of an average increase in use per talk turn following Question 3 (data sharing) and Question 5 (by-name lists). In contrast, this topic was used a mean of 4.2% (SD: 1.5%, t=2.88) less on average following Question 6 (invisible homelessness) and showed some evidence of being discussed less following Question 7 (subpopulations) and Question 9 (opinions on current data systems).

Finally, Topic 13 was discussed more in talk turns following Question 5 (by-name lists): the average topic proportion of this topic increased by 12.3% (SD: 3.3%, t=3.76). Topic 13 was discussed somewhat less on average following all other questions, but especially Question 9 (lower by 3.6% (1.7%, t=-2.10)) and Question 6 (lower by 1.7% (1.0%, t=-1.67)).

Table 7.3: Covariate associations with lower quality topics. Table 7.3 presents the results of regressions on the topic prevalence of three lower quality topics: Topic 1 (current

Table 7.3: Regression results on talk turn-topic proportions for Topic 1 (Collecting data and current systems), Topic 7 (Case management for PEH), and Topic 8 (Making contact with PEH and organizations).

	Topic 1		Topic 7		Topic 8	
	Coefficient (SD)	t	Coefficient (SD)	t	Coefficient (SD)	t
Intercept	.042 (.054)	.785	.080 (.065)	1.22	.038 (.057)	.665
Log popn	.0031 (.0091)	.345	-.00077 (.011)	-.069	.0010 (.0091)	.115
Remoteness	.0038 (.066)	.059	-.016 (.073)	-.215	.035 (.070)	.505
Question 1	.027 (.012)	2.24	-.0052 (.011)	-.477	-.012 (.010)	-1.15
Question 2	-.0062 (-.021)	-.302	-.0028 (.020)	-.137	.027 (.020)	1.37
Question 3	-.010 (.015)	-.689	.021 (.014)	1.54	-.0048 (.013)	-.358
Question 5	.0079 (.031)	.253	.015 (.029)	.520	-.019 (.024)	-.765
Question 6	-.0085 (.011)	-.744	-.0077 (.012)	-.617	-.0016 (.011)	-.144
Question 7	-.021 (.020)	-1.02	-.017 (.020)	-.853	.0026 (.019)	.138
Question 9	.013 (.025)	.497	-.0043 (.024)	-.179	-.010 (.019)	-.527
	Var components (SD scale)	% Explained	Var components (SD scale)	% Explained	Var components (SD scale)	% Explained
Intercept	.016 (.0070)	.63%	.018 (.0089)	1.50%	.012 (.0080)	.83%
Residual	.15 (.0070)		.15 (.0070)		.14 (.0089)	

data systems), Topic 7 (case management for PEH), and Topic 8 (making contact with PEH and organizations). In the average community, Topic 1 had a mean document-topic proportion in each talk turn of 0.060, Topic 7 had a mean of 0.072, and Topic 8 had a mean of 0.052. The amount these topic proportions varied on average between focus groups was generally low (a standard deviation of topic proportions between 0.012 for Topic 8 and 0.018 for Topic 7) and did not explain very much of the overall variance in the proportions (between only 0.6% for Topic 1 and 1.5% for Topic 7).

Topic 1 had an average increase in its topic proportion in talk turns following Question 1 of 2.7% (SD; 1.2%, t=2.24); Topic 7 showed some evidence of an increase in average talk turn-topic proportions following Question 3 (increased by 2.1%, SD: 1.4%, t=1.54); and Topic 8 showed some evidence of an increase following Question 2 (new PEH) with a mean (SD) increase of 2.7% (2.0%, t=1.37).

7.4 Discussion

In a secondary topic modelling analysis of 29 focus group transcripts as part of the Homelessness Counts project, we modelled 17 topics, of which 8 were judged to be high-quality after looking at the representative words and documents. Among the 9 lower quality topics, some seemed to collect words and documents which were similar to those already addressed by high-quality topics (e.g., Topic 1) while others seemed to have split themes among themselves (e.g., Topic 3, 11, and 15). Broadly, like the questions in the focus group guide, the topics were divided into three groups: about the subpopulations of people experiencing homelessness, about data collection and its applications, and about local interactions with PEH.

7.4.1 Comparison to previous work

Two qualitative analyses of the present focus group data (one focusing on COVID-19 using an ethnographic approach [43], and one on rural and remote homelessness using a thematic analysis [42]) were previously conducted. Although different in their scope and research questions, we can contrast some of the topics against comparable qualitative results for validation.

7.4.1.1 Subpopulations and causes of homelessness

Topic 2 was devoted to youth homelessness and intersections with school systems. That youth homelessness emerged as a key topic in the focus groups is unsurprising: both qualitative analyses of the focus groups also identified youth as a key population [42, 43]. This is corroborated the government of Canada, which found 26% of PEH in the 2018 point-in-time counts were youth or children [60], and the majority of people currently experiencing homelessness first experienced homelessness as youth [62, 60]. In terms of the content of Topic 2, “LGBT” and “trans” were among words associated with this topic, echoing the In-

frastructure Canada result that nearly one quarter of homeless youth identified as LGBTQ+ [62]. Also in Topic 2 were words relating to school and education; in documents with a high proportion of Topic 2, these were discussed by stakeholders as points of contact with youth and families experiencing or at risk of homelessness. A post hoc regression on Topic 2's prevalence across documents suggested that this topic was discussed more than average in stakeholder talk turns following them being asked about hidden homelessness. This suggests that stakeholders may associate youth with under-counted populations of PEH. Interpreted along with the "school" content this topic, targeted counts that utilize the school systems may obtain even higher estimates of the prevalence of youth homelessness.

Our topic model also identified other vulnerable groups in homelessness in Topic 16, particularly women and Indigenous peoples. Tellingly, "violence" also arose as a high probability word in this topic. These results are corroborated by the government counts which identified an overrepresentation of Indigenous persons among PEH [62] and of women fleeing abuse [59] and is echoed by the thematic analysis of rural and remote homelessness, which identified women and Indigenous peoples as important populations [43]. Another high probability word, "single" gains additional context in light of the finding that homeless single-parent families with child dependants were common [62]. Topic 16 also included "men" as a high probability word, which from looking at high probability documents, was often used in either the context of gender-based violence against women, or the division and disparate supports available in communities for homeless women versus homeless men.

That Indigenous representation among homeless populations in Canada did not emerge as a topic on its own is unsurprising when the scope of the topics is examined: rather than dividing by population, looking at word co-occurrence placed "Indigenous" as a high probability word in three topics. First, the aforementioned vulnerable populations topic. Second, a topic which mostly contained documents relating to evictions, but had several

high probability documents relating to a subtopic of insensitive data collection on Indigenous PEH using specific tools. Finally, "Indigenous" appeared as a high probability word in Topic 6 which focused on safety and comfort of people with lived experience of homelessness. These last two results, particularly the higher quality topic, Topic 6, align well with the qualitative theme from the thematic analysis that pointed to cultural stigma and barriers to services as affecting service provision to specific groups [42]. In the thematic analysis, this quote was chosen as exemplary for this theme: "Well, none of the services are Indigenous-led. And most of the people that are homeless are First Nations, Inuit, and Métis. Therefore, they wouldn't access services because they're not culturally appropriate and they're rigid, and specifically people's lived experience have made it clear that they don't feel comfortable accessing those services quite often. At one point there was between three hundred and five hundred people that chose not to stay at shelters either, or not reach out to other people either." In our topic model, this document had an estimated proportion of 47% of Topic 6. It is then unsurprising that we observed an increase in document-topic$_6$ prevalence following Question 6, which asked about invisible homelessness. Stakeholders in focus groups appear to associate a need for services that are developed to ensure the safety and comfort of key populations, such as Indigenous peoples, with unmeasured homelessness in their communities.

A final population that the topic model captured were older adults: "seniors" arose as a high probability word associated with evictions and tenancy. This is corroborated by the government of Canada finding that the main cause of homelessness among older adults were evictions [59], and by a similar finding in the ethnographic qualitative analysis [43]. However, our topic model did not separately capture the final subpopulation from the thematic analysis: people with disabilities [42]. This is likely because most interactions between healthcare and homelessness were captured in Topic 14, which discussed health and addictions.

The lack of demonstrated association between topics about homeless subpopulations or events causing homelessness with the remoteness or population size in this topic modelling analysis is not altogether surprising. Although some populations may be particularly disadvantaged in small or rural communities, this does not necessarily correlate to a larger amount of time devoted to the subject in a given talk turn or among stakeholders in a particular community. Partly, this could be due to the division of talk turns which created a lot of variance within focus groups. However, it is important to remember that a topic model simply cannot distinguish between prevalence of discussion and prevalence of the populations. More than likely, specific subpopulations, such as youth, women, and members of Indigenous communities, are overrepresented across Canada in homelessness, and as such their presence was reflected across all focus group discussions.

7.4.1.2 Data collection and use

That there is a lack of literature on data systems for the homelessness sector in Canada motivated this focus group study, however it makes comparison of our topic models to external information challenging. We thus rely largely on the post hoc regressions to validate our topic interpretations for topics related to data collection and use from an external standpoint.

Topic 4 in our model concerned data sharing between service providers; a post hoc regression was able to confirm that this topic was discussed in higher proportion following the facilitator asking about data sharing in Question 3. It was also discussed relatively more in talk turns following Question 9 (opinions on current data systems). Both "challenge" and "barrier" were words associated with this topic, suggesting some difficulties with sharing of information. The thematic analysis identified that data sharing was done differently in rural and remote communities, where dedicated electronic systems were less frequently used than a word-of-mouth referral [42]. Our topic model did not distinguish between these two types of sharing, although the high probability words in Topic 4 seem to be largely

about computer-based information (e.g., "HIFIS"). Regression on the topic proportions of Topic 4 did not identify a change in the amount this topic was discussed by remoteness index, however it did show an increase in the average amount a talk turn was devoted to this topic in communities with larger populations. This may relate to the thematic analysis' observation that there are fewer resources in smaller and more remote communities [42]; larger cities may have more staff dedicated to data administration. Both the number of people available to implement and discuss data sharing systems and the variation in types of systems may explain why Topic 4 had the largest variation between focus groups in topic proportions.

Both Topics 1 and 12 showed an increased amount of discussion following Question 1 (about data collection), validating our interpretation. The relative strength of the association of Topic 12 with Question 1 when compared to Topic 1's association echoes our judgment that Topic 1 was more general and of lower quality than Topic 12, but still addressed data collection in the homelessness sector. Similarly to Topic 4, there was a relatively large amount of variation (2.4%) in the average document-topic proportion of Topic 12 attributable to between focus group differences. Nearly all focus groups were asked Question 1, so we hypothesize that this difference is related to variation in participant knowledge about the systems in use and that some of the responses to Question 1 were thus captured in a larger proportion by the less precise Topic 1. Topic 12 was particularly interesting for the number of high probability words in this topic that were verbs the stakeholders used when describing data utilization, including "report", "manage", "track", "access", and "pull".

Topic 13 (by-name lists and coordinated access) showed the expected increase in topic proportions following the asking of Question 5 (availability and accessibility of by-name lists). That Topic 13 had "HIFIS" as a high probability word is unsurprising: both the by-name list and coordinated access are features of HIFIS. This explains why Topic 12

(also with HIFIS as a high probability word) was discussed more in talk turns following Questions 3 (data sharing) and Question 5 (by-name lists).

Although we were able to validate the interpretation of these topics related to data collection and sharing in the homelessness sector, the topic model was not able to capture changes to these systems. For example, [43]'s ethnographic analysis identified changes to data systems with the onset of COVID-19. We had no topic which explicitly addressed the COVID-19 pandemic, nor did we have any covariates relating to the pandemic which could be used for external validation of this result. When looking at the topic-word proportions, "covid" was nearly equally distributed in probability across all topics and "pandemic" had only a slight increase for Topic 9. In this analysis, all topics should be interpreted in the context that these focus groups were conducted virtually during the first few years of the pandemic, which influenced the nature and content of the discussion.

7.4.1.3 Locations

Topic 5 and Topic 8 both contained words relating to locations in the community where homelessness was observed by stakeholders.

In looking at the documents with a high proportion of Topic 8 and its highest probability words, we noted that the presence of "new" and "contact" might suggest the topic was related to Question 2 (identifying a new PEH). This was corroborated by the regression analysis, which provided some evidence that the topic was present in higher proportions following Question 2. However, the variance around the estimate of this increase was large enough that we cannot be entirely confident in this association. This is likely due to the fact that Topic 8 was only of intermediate quality, meaning not all documents with a high proportion of this topic had a direct relationship to the assigned interpretation.

In the ethnographic analysis of the focus groups, the following quote was chosen as exemplary to illustrate an increase in people sleeping in the rough due to the COVID-19

pandemic [43]: "And you know, there's not a week goes by here that somebody doesn't come in and ask me for a tent and the sleeping bag and warm blankets and stuff. They have got no other place to go." In our topic model, the proportion of this document associated with Topic 5 was 58%. Like for data collection, the topic model was unable to distinguish changes in behaviour from the behaviour itself, but it was able to capture this document as belonging to a topic discussing sleeping arrangements and locations of PEH. Similarly, many of the same locations where PEH are observed were identified in Topic 5 as in the ethnographic analysis, without the dimension of identifying changes due to the pandemic [43]. We were also unable to distinguish migratory patterns using the topic model, nor explicitly identify the central theme of "displacement" from the thematic analysis looking at rural homelessness [42]. In practice, this is to be expected: although the mixSTM can distinguish groups of words based on co-occurrences, it has limited applicability to identifying a single unifying or overarching theme across the whole corpus.

7.4.2 Strengths and limitations of our topic model analysis

Many of the topics in our topic model were of high-quality, with high probability words forming semantically related sets and documents with high proportions of the topics discussing coherent themes. These topics aligned well with expected results, given external information, previous research, and the context of the focus groups. Even in lower quality topics, the word associations occasionally lined up with expected results although we necessarily had less confidence in these results due to high variability in the estimates.

The number of topics which were not easily interpretable or had intruding documents was likely due to the fact that we used a large number of topics relative to the mean stopworded document size of 30 words. When we chose the number of topics, we opted for 17 because the granularity of the high-quality topics was appealing relative to the 11-topic model. However, this led to several topics of middling quality which shared themes (e.g., Topic 3 and Topic 15 both discussed rentals, albeit with somewhat different perspectives). In

practice, we can be relatively confident in the high-quality topic results: many preliminary and unselected models shared similar topics to the ones identified here, at least in terms of high probability words.

Indeed, perhaps the most critical limitation of our analysis was that we chose to define documents as talk turns, meaning our documents were short. Using singular talk turns meant the regression results were easily interpretable at the level of the documents: we were able to make conclusions such as how much topics in talk turns tended to increase in prevalence following a question. To keep that interpretability, using bigrams could have increased our per-document vocabulary (as in [22] or [108]). Alternately, we could have modified our preprocessing strategy to eliminate all documents with 10 or fewer words after stopwording, although we may have then eliminated some documents with rich content. In practice, since we had a good number of longer documents and were able to incorporate additional information about the expected topic prevalence, our topic model was able to estimate semantically interesting topics.

The incorporation of covariates into the estimation of and post hoc inference on the topic proportions had some additional limitations. For one, incorporating the participants' roles in the homelessness sector into the model on topic prevalence could have led to additional insights. However, this information was redacted before the transcripts were cleaned for the present analysis. Another covariate of interest might have been the date of the focus group relative to COVID-19 waves, but the waves differed across Canada making this a challenge to model consistently. For another, we elected to include our two continuous covariates as having linear relationships with the topic prevalence. Roberts et al. suggest that splines, being more flexible, may result in better models [99]. In our analysis, we cannot exclude the possibility that there is a non-linear relationship between continuous covariates and the document-topic proportions. Finally, we elected not to incorporate covariates into the topic content model of the (mix)STM. In preliminary analyses, the addition of a content

covariate limited the interpretability of topics by providing summaries with mostly words distinct to the levels of the content covariate. Since along any divisions the participants in this study were relatively homogeneous in their language use and word choice, adding a content covariate masked the true most probable words (which were common to all groups). A more focused topic model study, looking specifically at word use differences between urban and rural settings in these focus groups, might be an interesting avenue for future work.

A particular strength of using a topic model in this context was its ability to analyze the corpus as a whole to identify patterns and assign words to topics in a global manner. However, we did exclude some transcripts from our analysis. This was to create a set of focus groups which resembled a typical semi-structured focus group corpus (multiple participants and multiple focus group guide questions asked) and to enable some of our post hoc analyses (such as validation by question). Had we chosen a less rigid approach to facilitator questions (like a fully qualitative analysis of the facilitator text or a preliminary topic model), more text may have been able to be incorporated into our model. Although our broad question categorization strategy appeared to align well with the conversation, other strategies would have also permitted more granular post hoc inference on topic prevalence. Unfortunately, these alternate approaches require significantly more time and resources, particularly as the number of focus groups and meetings increases. Extensions to this topic model analysis that incorporate all meetings, focus groups, and also interviews with PEH could elaborate on the themes here, potentially with more specificity.

7.4.3 Implications

A mixSTM applied to the focus group transcripts from the Homelessness Counts project was able to identify topics and collections of topics which aligned with previously identified themes from the corpus, and extracted some novel associations. This highlights the potential of topic models integrated into the analysis of focus groups: documents with a

high proportion of a certain topic could be used as a starting point for a qualitative analysis of a given topic/group of topics or as a complementary tool to supplement previous descriptions; previous topic model analyses agree [85, 84, 83, 25]. However, since topic models are but a method of language exploration, there can be lost nuance that could be (and was, in our case) better captured by qualitative approaches. Miyaoka et al. arrived at similar conclusions after their focus group analysis using topic models: although there was a high degree of agreement between the topic model and the qualitative approach in broad themes, the topic model was less nuanced and unable to identify differences by subpopulation or evaluate sentiment from the participants [84].

It was interesting that our topic model captured some topics that seemed to be directly answering a given focus group guide question (such as Topic 4 about data sharing), while other topics were not (such as Topic 14 about mental health and addictions). On the one hand, topics that relate to guide questions have access to a direct external validation strategy using the post hoc regression. These topics could be useful in highlighting words and documents that are likely associated with a facilitator's question. On the other hand, estimating relationships of other (not directly guide-related) topics with the questions can provide additional insight into the topic's context. An increase in prevalence of an emergent topic after a question might indicate that a relationship exists between the words in the topic and the subject of the question, opening avenues for further hypotheses. Furthermore, these unprompted topics emerged throughout the corpus, suggesting particular relevance to the participants.

A novel insight from the mixSTM is in the amount of variation in topic prevalence attributable to between-session differences. This quantity can be interpreted in terms of the context of the focus groups and their sampled participants.

Finally, we repeat that many of the choices made in estimating a topic model are subjective. Had we chosen a different number of topics, the resulting topics would not be the same.

For example, in both the 11- and 14-topic preliminary models, there were topics which had "couch" and "surfing" as high probability words, referring to an often-invisible population of PEH; these words did not appear with high probability in any topic in the chosen 17-topic model. Is this a limitation? Yes and no: that the chosen model cannot make conclusions or inference about couch surfing populations is unfortunate in that it might have been interesting to explore, but the other two models were not chosen because many of their topics were less precise. The motivation for choosing the topic modelling approach, the number of topics, and the post hoc estimation strategy is unique to the research question and to the analyst. Any topic model results should be interpreted in light of that.

7.4.4 Conclusions

This secondary analysis using a Mixed Structural Topic Model of focus group transcripts was able to describe the corpus in seventeen topics. We validated the results of our topic model from a semantic lens (looking at high probability words and documents with high proportions of the topics) and from an external lens (running post hoc regressions on topic prevalence and comparing the results to previous studies). The highest quality topics and some intermediate-quality topics aligned well with previous results. Future topic models of this corpus should consider an integrated analysis of all meetings, interviews with PEH, and the selected focus groups, for the potential of additional insights.

Chapter 8

8 Conclusion

Focus groups are one of many qualitative data collection methods used in health and social research which provide an interesting source of data for an analysis integrated with quantitative methods. Motivated by a desire to model focus group transcripts (and their multi-participant, semi-structured, conversational nature) using topic models, we have proposed modifications to the Structural Topic Model of Roberts et al. [96, 100, 97, 99]. Although these modifications were developed in concert to study the same grouped document data sources, they need not be used together, depending on the corpus and the goals of the analysis.

8.1 Discussion of contributions and future work

In this section, we will discuss some strengths and limitations of the three proposed extensions to the STM as well as additional avenues for future models of grouped documents using the STM and mixSTM.

8.1.1 Random effects in the mean topic prevalence model

In Chapter 4, we proposed using a mixed effects regression with regularized fixed effects in the update to the expected document-topic weights. This modification better accommodates the potential for relationships between covariates to differ between groups and costs fewer degrees of freedom in the estimation than the original fixed effects regression. We showed in simulation that cases exist where the mixSTM's estimation for μ results in document-topic proportions that are closer to the true values and leads to an improved held-out likelihood.

A limitation of this modification to the mean topic prevalence estimation is that the stream-

lined implementation of the penalized mixed effects regression can occasionally face issues apparently when there is insufficient variation in the dependent variable to accurately model random slopes. A less streamlined process might require more time but avert these issues; other penalized mixed models could also be implemented at this step, for example that of Yi and Tang [132]. In practice, the regression strategy already relies on introducing bias through regularization and is only expected to retrieve strong effects that actually exist. So, the consequences of an inaccurate regression should be relatively minimal, particularly for early or intermediate iterations of the EM algorithm. At worst, so long as the model does converge, the maximum obtained might be a relatively poor bound on the ELBO– but it will always be a maximum thanks to the guarantees of variational inference.

In some ways, the mixSTM is designed for maximalism: unless the number of groups is large relative to the number of documents and/or there are several covariates whose effect on topic prevalence is expected to vary by group, an approach to the data using the original STM may perform equally as well or better. Unfortunately, in naturally generated text corpora, there is no way to know which variables "truly" influence topic prevalence, nor is it possible to measure all possible quantities and incorporate them. We recommend using the mixSTM particularly if there is a desire to model many random slopes or use the random effects correlation structure for additional pooling of information for topic prevalence.

In our simulations, we generated corpora which had a similar size to a hypothetical qualitative corpus, that is, relatively small compared to the scale that topic modelling approaches can handle. Even when we further decreased the mean document size from 150 words to 30 words (coincidentally exactly the same mean size as the documents in our case study) and decreased the between-group variance, the mixSTM still performed well in retrieving MAP estimates closer to the true document-topic proportions than the STM. However, there are limitations to simulation studies for topic models, namely that it is challenging to simulate

text that resembles human-generated content. Applying the mixSTM to naturally generated data in additional contexts would help explore its utility.

The modification to the mean document-topic weight estimation is applicable to any corpus with some grouped structure of the documents. The regression framework is a familiar one, making this a very accessible approach for incorporating structural assumptions about the corpus into the topic prevalence model of the STM.

8.1.2 Group-specific topic covariance matrices

In Chapter 5, we proposed a modification to the generative model of logistic-normal topic models where one topic covariance matrix is estimated per group, rather than one overall. This modification represents what we realistically expect in the case of inherently grouped documents: the correlations between topics may depend on the groups. In simulation, we showed that this approach has potential to improve model fit when there is a simple or no model on the document-topic mean weights and when the number of topics is small.

A primary limitation of this modification is that it can over-complicate the model. In estimating G covariance matrices, we are introducing many more parameters to estimate at each iteration using the same number of words and documents. When K is large, the matrices grow in size: $(K-1)(K-2)/2$ (co)variances must be estimated for each of G covariance matrices. As we observed in our simulations on small datasets, in more complex models, providing additional information about the expected document-topic proportions did not improve the fit, likely because the degrees of freedom are consumed by these many large matrices. As such, it may not be desirable to use this proposed modification to the STM in concert with a complex mixed effects model on the mean document-topic weights. However, estimating groupwise covariances did improve the fit in models without any covariates on the mean document-topic proportions, given that there was a grouping structure to estimate, as we have assumed throughout.

Another limitation of this modification to the STM is that it restricts our ability to generalize inferences. In simulation, even when estimating grouped covariances improved the fit of the model, its ability to fit to held-out documents was poorer. Beyond that, by estimating a covariance matrix per group, we do not have any information that could be used to estimate quantities about documents that belong to an outside group not included in our model. This is in contrast to our modifications to the mean document-topic proportions above, which, through random effects, have baseline information about groups external to the fitted model.

Incorporating group-specific correlations between topics into the generative model of the STM or any logistic-normal topic model has promise for settings with grouped documents when detailed information on covariates is limited. In these cases, the groupwise covariance matrices may result in better fitted topic models by providing information about document-topic proportions in a group-specific manner.

8.1.3 Estimating the effect of covariates on topic prevalence

In Chapter 6, we proposed an extension of the regression of covariates on document-topic proportions which can accommodate random effects in addition to the fixed effects. We showed in an example that this method successfully propagated estimation uncertainty from the latent variables into subsequent post hoc regression estimates.

Throughout, we have designated this regression "post hoc". Although the researcher may specify a comparison of interest a priori (in our case study, remoteness), the regression estimation cannot occur until the topics have been estimated and interpreted. We use "post hoc" to emphasize that this approach has some limitations when the desired interpretations are causal. For one, since the topic search can take on a qualitative dimension and is an iterative process, we are, in many cases, "fishing" for the ideal topics [37] and then choosing which topics to use as an outcome in the regression. From a validation perspective,

performing an exploratory regression is not an incorrect approach, although we must be cautious of over-interpreting results that may be coincidental. For another, although we can attempt to adjust for some confounding effects in our regression (e.g., the inherent grouping of documents within focus group discussions), there is likely to be unmeasured confounding of the association of topics with external quantities. This is natural: the dynamics between human speech cannot be wholly captured with a regression, even a mixed effects one. Also, depending on the document definition chosen, the crucial temporal aspect may not be fulfilled for the post hoc regression analysis on topic prevalence, making causal inferences impossible. Finally, Egami et al. point out that overfitting to the sample at hand is highly likely in text-based inference, thanks to the ease with which humans associate patterns with text [37]. We may discover spurious associations that would not hold outside the corpus. In topic models applied to qualitative datasets, this is less of an issue, only because the limitations of the sampling process are already baked into the research design; we are well aware that the sample (of text or participants) will not be representative or large enough to be broadly generalizable. However, the concerns of spurious associations are valid: we reiterate that the conclusions obtained from this regression are 1) not generally able to be interpreted as causal associations; and 2) should only be used as one piece of evidence in a larger scheme. If there is a desire to perform causal inference on text, analysts should consider strategies such as those outlined in [37, 98, 129, 109].

We omit p-values from our regression results. This is a two-fold decision: for one, estimating the appropriate number of degrees of freedom to use in t-test on regression coefficients in a mixed effects regression has additional complications. One could implement an additional procedure capable of estimating p-values, but the relative cost in time is outweighed by our second consideration: p-values are often misinterpreted as a dividing line between important/unimportant results [128]. We recommend that coefficients be presented with their standard deviations and/or t-values to be interpreted in terms of relative amounts of evidence with respect to background noise. In regards to the previous paragraph, we also

acknowledge that spurious associations may arise and that there may be confounding by unmeasured variables, making a multi-dimensional exploration of the results important.

Finally, we have argued throughout that one benefit of using a mixed effects model for focus groups is the ability to describe variance between focus groups rather than particular focus group levels. This may not be a desired result in all studies: both [9] and [77] explicitly present results by focus group. In the case of [9], the comparisons are built in to the research questions (which are explicitly contrastive). An approach which described the variation between focus groups in topic proportions could have been applicable to [77], similar to in our case study, although we recognize that when there are few groups, variance estimates may not be precise. Whichever method is preferred in a given study, it is easy to retrieve per-focus group estimates using the mixSTM's strategy (they are in fact explicitly estimated in the model fit and in the post hoc inference)– but only the mixSTM can also describe inter-group differences using the proportion of variance explained.

Estimating a post hoc mixed effects regression on document-topic proportions is possible with any grouped corpus, although its relevance may depend on the specific corpus and research questions. Using random effects in a regression may be motivated in many ways, including those enumerated in the previous chapters, and we encourage all analysts to consider their justification before using this or any model.

8.1.4 Future work on the STM and mixSTM for grouped documents

In the following, we identify seven potential avenues for using the framework of the STM and mixSTM on grouped documents.

First, we considered only the case of two-level nested groupings in the present application. While Degani et al. also extend their implementation of streamlined inference for penalized mixed models to three-level nested models [34], no extension of this method yet exists for crossed random effects. Three-level nested models could be useful in the analysis of focus

groups (e.g., to model the variance of participants nested within sessions) and in many other corpora. An extension to crossed effects, which would be able to model repeat participants across multiple focus group sessions, could model nuance in analyst beliefs about specific grouping structures. Crossed random effects are challenging in their estimation, and no methods yet exist that we know of which can rapidly estimate crossed random effects with some degree of penalization on the fixed effect coefficients.

Second, the method of Degani et al. considers three global-local prior specifications on the fixed effects coefficients [34]; we used only the Horseshoe prior on the mean document-topic weights to obtain the results presented here. We found little difference between the three implemented priors when experimenting with simulation settings, however we were not explicitly considering the environments in which each was expected to perform differently. Further simulations that explore the differences, in the context of the iterative regressions in the mixSTM, could help to provide recommendations for the mixSTM's implementation. Furthermore, many other shrinkage priors exist in and outside of this family [68]. Extensions which incorporate other priors may also be an avenue for future work.

Third, we assume in our modelling of the mean topic weights that group size is non-informative. In other clustered-data settings, this is not always the case and can lead to confounding of the effects of interest [107]. Methods that explicitly consider this in the modelling may lead to more robust inference on topic prevalence, particularly in the post hoc regression we have suggested for use in the mixSTM.

Fourth, in our exploration of group-specific covariance matrices in the generative model of the mixSTM, we considered completely separate estimations of covariances per group. This contributed to the risks of overfitting and overparameterization since the information available to a single group is smaller than the amount in the whole corpus. Alternate approaches which could pool information across groups or otherwise limit the number of

parameters to be estimated should be considered. For one, we could impose explicit structural assumptions about the covariances. One such method is already provided in the `stm` package [99], whereby a user can tune the degree of diagonalization of the estimated Σ, although we did not explore the impact of applying this diagonalization in our simulations. Another method would be to set a prior on the covariance matrix with a degree of informativity in order to restrict the parameter space of the estimation, although this will have broader consequences for the variational inference algorithm. For another solution a degree of "pooling" of the covariances could be implemented: given an especially small or high-variance group, the group's covariance matrix could be pulled towards resembling the global covariance. These approaches may result in more stable inference when using grouped covariance estimation.

Fifth, we assume throughout that the document-topic weights are independent draws from a normal distribution with some mean and covariance. The relatedness of the topic prevalence between documents within groups is modelled through shared means and covariances in the M-step, but no explicit relationship between the documents within groups is modelled. As a future direction for the mixSTM, integrating a specific parameter which quantifies the relationship between every pair of documents could help describe a grouped document corpus. This could be done for example in the manner of the Relational Topic Model (which estimates a binary indicator for each pair and will be described further below) [27], or in a way which leverages the matrix normal structure in the topic prevalence model to incorporate correlation parameters.[1] However, there may be limitations to this approach: incorporating more parameters requires more estimated quantities and adds more complexity to an already complex model– without sufficient data, the estimates may be

[1]The topic weights η for a given document are independently drawn from a normal distribution with mean μ_d and covariance Σ. Across all documents, we have a matrix E: $\begin{bmatrix}\eta_1^\intercal & \dots & \eta_D^\intercal\end{bmatrix}^\intercal$ and a same-dimension matrix M of the per-document means. We know from identities of the matrix normal ($\mathcal{MN}$) distribution that $E \sim \mathcal{MN}(M, \mathbf{I}, \Sigma)$. If the identity matrix is replaced by a matrix with off-diagonal entries, the documents will be assumed to have some correlation.

inaccurate or imprecise.

Sixth, as briefly mentioned in previous sections, there are additional explorations of the mixSTM that merit consideration. In our simulations, we demonstrated the mixSTM on relatively small corpora. The applicability of the mixSTM to corpora larger than what could reasonably be expected to result from qualitative research should be explored in future work. Furthermore, more information is needed about the limits of the STM and mixSTM when applied to small corpora, for example in the manner of [114]. This is especially true in the context of generating further evidence (for example in the form of permutation tests [100]) for whether integrating covariate information into structural topic models can induce effects on the corpus. In our case study, even with short documents, we did not appear to have generated any group-related effects for the comparisons of interest. This also relates to a final investigation which should be performed in future work: semantic interpretability. We compared the mixSTM to the STM and CTM using primarily numerical measures of fit. However, topic models of text are primarily considered useful for creating *interpretable* lower dimensional representations of text. Evaluations of the mixSTM on real corpora of text in terms of its ability to generate semantically interpretable topics should be considered, particularly in comparison to simpler models like the CTM.

Finally, we discussed only the topic prevalence model in our extensions to the STM for grouped documents, but the topic content model also provides an interesting avenue in this setting. The topic content model accommodates one categorical covariate, which it incorporates as deviations from the baseline topic-word probabilities [97]. However, the computational cost of incorporating a large number of content covariate levels is high. Also, if too small of a vocabulary is used relative to the number of content covariate levels, the results may not be very interpretable, since the topics are presented by level rather than globally. Although focus group session could be incorporated as a content covariate, the topic content model of the STM cannot accommodate our motivating desire to describe dif-

ferences between focus groups using one measure of variance rather than estimates at each level. However, if the interest in topic models for focus groups is to directly compare the sessions on some higher-level covariate, e.g., if focus groups are homogeneously sampled from different population groups, the topic content model as integrated into the original STM is promising.

8.2 The mixSTM in a broader context

Throughout, we have discussed applying the mixSTM to focus group transcripts. Other topic models may be relevant for an analysis of this type of corpus and the mixSTM is also applicable to other corpora and qualitative concepts.

8.2.1 Alternative models of grouped documents

We cannot claim that every corpus of grouped documents necessitates the mixSTM. Instead, the mixSTM has its place in an incredibly vast network of topic modelling methods. For any given task, an analyst must choose a model that coincides with the goals of the analysis and the presumed characteristics of the corpus. We take the time here to highlight a few alternate approaches which may suit grouped document corpora like those motivating our development of extensions to the STM.

Mimno et al. propose graph-based priors which allow for complex relationships between groups of documents [81]. Similar to the STM, they propose shared mean topic weights within groups, but their approach additionally incorporates information about the relationships between the groups' means. The similarity to our motivation for the mixSTM, which models grouped documents and the relationships between topic prevalences within groups, is clear. However, in contrast to the mixSTM, the incorporation of additional covariates besides grouped structure into the model on topic prevalence is not explicitly available in [81].

Other methods of modelling relationships between documents have also been proposed. The Relational Topic Model samples, for each document pair, an indicator of whether the two documents are linked [27] making explicit the relationship between documents. This adds an appealing additional dimension to the interpretation of topic models, especially when the documents are expected to be related. Several extensions and alternative models which estimate a network-like relationship between documents exist, many of which are discussed in [118]. These approaches have a direct congruence to our desire to model relationships between documents which share some innate grouping, although the Relational Topic Model infers the structure from the data rather than imposing it on the model [27].

Not only the relationships between documents but between topics within groups of documents may be of interest. The Constrained Relational Topic Model [118] allows the analyst to incorporate additional prior information about the document relationships into the Relational Topic Model, by limiting the sets of topics that documents must or must not share given the constraint identity. In a similar vein, grouped LDA [76] and its related topic models also restrict the topic distribution between groups of documents. In the GLDA of [76], the generative model assumes that each document belongs to a group, and each group has certain "local" topics (belonging uniquely to that group) and some "global" topics (shared by the whole corpus) which will be sampled from for that document. In this approach, the number of groups is fixed but the members of the groups are inferred from the model. Similarly, in the Multi-Field CTM of [103], the number of groups (fields) is taken as fixed, and both the topic weights and word probabilities are defined on a per-group basis. This allows for explicit separation of the topics between disparate groups or text sources, and could be applied not just to a division of focus groups, but for jointly modelling facilitator and participant speech.

A final type of topic model that is particularly relevant to the question of grouped docu-

ments, are predecessor topic models to the STM that can typically incorporate only one "group-like" covariate. A prominent example would be the Dynamic Topic Model [15], which samples mean topic weights and the word probabilities such that they evolve from the previous (categorical) time slice. The extension to this, the Dynamic Correlated Topic Model, can model changes to both the mean topic prevalence and the covariance between topics over time [120]. These models have the appealing property of incorporating a single group-like dimension in the form of categorical time, although the relationship between the groups is unidirectional and sequential. This may be relevant when there is a longitudinal dimension to how the focus groups are conducted and/or the participant responses. Many other examples of single-covariate or single-prior structure topic models exist; Roberts et al. list several in their introduction to the STM package [99]. Depending on the groupings of documents and the structures and information that one wants to incorporate into the model, these might be of interest.

In the end, however, it is important to remember that complex models are not always better. Any potential analyst should weigh the limitations of these approaches (including the mixSTM) before applying them: in many cases, the LDA or CTM would produce acceptable results.

8.2.2 Applying the mixSTM to other corpora in health research

We motivated the development of the mixSTM using the case of focus group transcripts but these methods are also applicable to other collections of text data in health and social research.

A natural extension of our motivating case of focus groups is that of interviews. On the one hand, interview transcripts can be partitioned like focus group transcripts to form documents which are innately clustered by interview session (often participant) and could be analyzed in the same manner already elaborated on extensively. However, a more interest-

ing application of the mixSTM might be to longitudinal interviews. Multiple interviews with the same participants could be performed during the course of a clinical trial to obtain more information about patient experiences and opinions on the intervention. In that setting, interview text is grouped within the participants, but there is also a time dimension to consider along with multiple other covariates of interest (such as intervention arm). Analogously to how a longitudinal repeated measures regression would be performed to assess the impact of the intervention while accounting for the innate clustering within participants, the mixSTM could be used to compare longitudinal trends in the prevalence of topics that arise in the interviews. Very similarly, but without all the resource costs associated with performing and transcribing interviews, initial insights from longitudinally administered open-ended surveys could be obtained in this way.

For another example in health research, the mixSTM could be used to model patterns of co-occurrence in patient medical chart annotations by healthcare providers. Using a topic model on a corpus composed of the medical charts of several patients, we could obtain clusters of co-occurring notes or diagnosis codes. Documents– defined for example as collections of annotations by a single healthcare provider during one visit– are naturally clustered within patients. Using the mixSTM, additional covariates (e.g., provider type) could be modelled as having varying effects on topic prevalence between patients, and other covariates at the patient-level could be included without being perfectly collinear with the patient indicator variable. Post hoc inference on topic prevalence using the mixSTM could also be interpreted in light of the variation between patients in the sample, rather than on a per-patient level.

In these example corpora, like in focus group transcripts, the documents are nested hierarchically within groups. In reality, many potential groupings of documents may exist in the same corpus: subject- or domain-specific knowledge might be used to justify using one grouping over another. For example, we considered focus group transcripts "groups"

of documents at the level of the session, but we could have modelled them instead using the participants as groups. In many focus groups, a participant-level grouping structure neglects the fact that participants are jointly contributing to an overall discussion, which globally informs the document content. If the focus group study in question were stylistically more like a round table, a person-level grouping might be an appropriate choice. In this (like in many of the subjective choices about implementing a topic model) the analyst must make a decision.

8.2.3 Focus group analyses through a quantitative lens

We have suggested the use of structural topic models in focus groups primarily for topic extraction. However, there is potential for additional integration of topic modelling with qualitative concepts and goals.

For one, we could use topic models to infer the association between different focus group facilitators and the prevalence of topics. This is particularly interesting in the context of qualitative positionality [102], which must be explicitly considered in qualitative research for its effect on data collection: does the perception of the facilitator as an insider (or outsider) to the focus group participants affect the results? Topic models assign quantitative values (such as prevalence) to representations of text (topics), which could be leveraged as a source of evidence for an association via post hoc tests. Within the exploratory framework of the post hoc regression enabled by the mixSTM, other potential confounders of the association between the facilitator or their characteristics and topic prevalence (such as the variations between focus groups and the questions asked) could be incorporated. Our case study had only one facilitator for the vast majority of focus groups, which prevented us from making this comparison.

Another qualitative concept which may have a measurable counterpart in topic model analyses of focus group data is that of saturation. Oft striven for, saturation is the concept

that collecting more data on a certain subject or research question will not meaningfully change the results (whether those be additional codes generated or additional depth on existing codes) [104]. There have been attempts to quantify "saturation" in qualitative research [56]. We suggest that topic models could be used to study saturation in qualitative studies, using the topics' interpretations, topic proportions and word probabilities, or regression results on topic prevalence. This could involve an iterative process of estimating a topic model after each focus group and comparing the results on one or more of the above dimensions, or integrating a time dimension explicitly and in a way that could be tested (using the Dynamic Topic Model [15] or the mixSTM). For example, one could ask after how many focus groups the topics stay relatively stable in their interpretations; at what point the topic proportions and high probability words in each document no longer change significantly; or when the time dimension in the topic model no longer coincides with an evolution of topics. Clear limitations exist for this hypothetical: for one, choosing the number of topics is already difficult before we consider multiple re-estimations of a model; for another, defining "stability" for topics and their proportions will require extensive consideration. It is likely that word co-occurrence will never stabilize entirely. However, there may theoretically come a point where obtaining more word co-occurrences from a population that is similar enough in their positions and connections to the research question will no longer meaningfully change the topics. This might, in a broad sense, correspond to some forms of saturation [104].

8.3 Conclusion

Topic models are, without a doubt, useful. However, it is important to remember that topic models, including the mixSTM, are only models of *text*. Miner et al. arrived at a similar conclusion after comparing LDA-generated topics to thematic codes: "...computational approaches may be more useful when one is looking to use topic modelling to develop codes

that directly reflect the participant's language (i.e. an "emic" interpretation), as opposed to identifying more interpretive, "etic" constructs that rely on examining not only the language itself but also how the language is used to make meaning out of individual or social experiences" [83]. In any analysis using the mixSTM, we are limited to modelling what people articulate or what is recorded. Although the goal of the mixSTM can be seen as trying to represent some of the dynamics of conversation– in the relatedness between what one person says and the next and the ways in which people make connections between ideas across a whole discussion– it will always be limited, not only by the available text but by our model-based simplifications of the intricacies of human speech.

That being said, huge volumes of text data are generated in health research both in and outside of explicitly qualitative contexts. With advances in machine transcription, studying speech as text is more and more accessible. The ability to rapidly obtain reproducible insights from text through topic models like the mixSTM– even if it is just at the explicit, word-use level– is promising for contributing to a multi-layered understanding of health and social context by using all the information researchers have at their disposal.

Bibliography

[1] John Aitchison and Shir-ming Shen. “Logistic-Normal Distributions: Some Properties and Uses”. In: *Biometrika* 67.2 (Aug. 1980), p. 261. ISSN: 00063444. DOI: 10.2307/2335470.

[2] Alessandro Alasia et al. *Measuring remoteness and accessibility: A set of indices for Canadian communities*. Cat. no. 18-001-X. Statistics Canada, May 2017. URL: https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2017002-eng.htm.

[3] A. George Assaf and Mike Tsionas. “Testing for Collinearity using Bayesian Analysis”. In: *Journal of Hospitality & Tourism Research* 45.6 (Aug. 2021), pp. 1131–1141. ISSN: 1096-3480, 1557-7554. DOI: 10.1177/1096348021990841.

[4] David C. Atkins et al. “Topic models: A novel method for modeling couple and family text data.” In: *Journal of Family Psychology* 26.5 (2012), pp. 816–827. ISSN: 1939-1293, 0893-3200. DOI: 10.1037/a0029607.

[5] Xavier Basagaña et al. “Analysis of multicentre epidemiological studies: contrasting fixed or random effects modelling and meta-analysis”. In: *International Journal of Epidemiology* 47.4 (Aug. 2018), pp. 1343–1354. ISSN: 1464-3685. DOI: 10.1093/ije/dyy117.

[6] Douglas Bates et al. “Fitting Linear Mixed-Effects Models Using **lme4**”. In: *Journal of Statistical Software* 67.1 (2015). ISSN: 1548-7660. DOI: 10.18637/jss.v067.i01.

[7] Eric P. S. Baumer et al. “Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?” In: *Journal of the Association for Information Science and Technology* 68.6 (June 2017), pp. 1397–1410. ISSN: 2330-1635, 2330-1643. DOI: 10.1002/asi.23786.

[8] Kenneth Benoit et al. “quanteda: An R package for the quantitative analysis of textual data”. In: *Journal of Open Source Software* 3.30 (Oct. 2018), p. 774. ISSN: 2475-9066. DOI: 10.21105/joss.00774.

[9] Alec Biehl et al. “Where does active travel fit within local community narratives of mobility space and place?” In: *Transportation Research Part A: Policy and Practice* 123 (May 2019), pp. 269–287. ISSN: 09658564. DOI: 10.1016/j.tra.2018.10.023.

[10] Jonathan M. Bischof and Edoardo M. Airoldi. “Summarizing topical content with word frequency and exclusivity”. In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML’12. Madison, WI, USA: Omnipress, June 2012, pp. 9–16. ISBN: 9781450312851.

[11] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. New York: Springer, 2006. ISBN: 9780387310732.

[12] Keya Biswas. “Performances of different estimation methods for generalized linear mixed models”. Master of Science. Hamilton, Ontario, Canada: McMaster University, May 2015.

[13] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. “Variational Inference: A Review for Statisticians”. In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1285773.

[14] David M. Blei and John D. Lafferty. “Correlated topic models”. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS’05. Cambridge, MA, USA: MIT Press, Dec. 2005, pp. 147–154.

[15] David M. Blei and John D. Lafferty. “Dynamic topic models”. In: *Proceedings of the 23rd international conference on Machine learning - ICML ’06*. Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 113–120. ISBN: 9781595933836. DOI: 10.1145/1143844.1143859.

[16] David M. Blei and John D. Lafferty. “A correlated topic model of Science”. In: *The Annals of Applied Statistics* 1.1 (June 2007). ISSN: 1932-6157. DOI: 10.1214/07-AOAS114.

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent dirichlet allocation”. In: *The Journal of Machine Learning Research* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[18] Michael Bloor et al. *Focus Groups in Social Research*. 1 Oliver’s Yard, 55 City Road, London England EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2001. ISBN: 9780761957423. DOI: 10.4135/9781849209175. URL: https://methods.sagepub.com/book/focus-groups-in-social-research.

[19] Ben Bolker. *Post-hoc MCMC with glmmTMB*. Oct. 2023. URL: https://glmmtmb.github.io/glmmTMB/articles/mcmc.html.

[20] Mollie Brooks E. et al. “glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling”. In: *The R Journal* 9.2 (2017), p. 378. ISSN: 2073-4859. DOI: 10.32614/RJ-2017-066.

[21] Paul-Christian Bürkner. “**brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*”. In: *Journal of Statistical Software* 80.1 (2017). ISSN: 1548-7660. DOI: 10.18637/jss.v080.i01.

[22] Anthony L. Burrow et al. “Does Purpose Grow Here? Exploring 4-H as a Context for Cultivating Youth Purpose”. In: *Journal of Adolescent Research* 37.4 (2022), pp. 471–500. ISSN: 0743-5584, 1552-6895. DOI: 10.1177/0743558420942477.

[23] Peter Cahill. “Methods Content Analysis: A Role in Applied Health Research”. In: *McMaster University Journal of Public Health* 1.1 (Dec. 2022). ISSN: 2817-2701. DOI: 10.15173/mujph.v1i1.3073.

[24] Carlos M Carvalho, Nicholas G Polson, and James G Scott. “Handling Sparsity via the Horseshoe”. In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. Vol. 5. Artificial Intelligence and Statistics.

Clearwater Beach, Florida USA: PMLR, Apr. 2009, pp. 73–80. URL: http://proceedings.mlr.press/v5/carvalho09a/carvalho09a.pdf.

[25] Parijat Chakrabarti and Margaret Frye. "A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography". In: *Demographic Research* 37 (Nov. 2017), pp. 1351–1382. ISSN: 1435-9871. URL: https://www.demographic-research.org/articles/volume/37/42/.

[26] Jonathan Chang. *Collapsed Gibbs Sampling Methods for Topic Models*. Apr. 2024. URL: https://cran.r-project.org/web/packages/lda/lda.pdf.

[27] Jonathan Chang and David Blei. "Relational Topic Models for Document Networks". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2009, pp. 81–88. URL: https://proceedings.mlr.press/v5/chang09a.html.

[28] Jonathan Chang et al. "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Neural Information Processing Systems*. Vancouver, BC, 2009, pp. 1–9. URL: http://umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf.

[29] Jianfei Chen et al. "Scalable inference for logistic-normal topic models". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., Dec. 2013, pp. 2445–2453.

[30] Yingying Chen et al. "What We Can Do and Cannot Do with Topic Modeling: A Systematic Review". In: *Communication Methods and Measures* 17.2 (Apr. 2023), pp. 111–130. ISSN: 1931-2458, 1931-2466. DOI: 10.1080/19312458.2023.2167965.

[31] Zhen Chen and David B. Dunson. "Random Effects Selection in Linear Mixed Models". In: *Biometrics* 59.4 (2003), pp. 762–769. ISSN: 0006-341X. DOI: 10.1111/j.0006-341X.2003.00089.x.

[32] Tom S. Clark and Drew A. Linzer. "Should I Use Fixed or Random Effects?" In: *Political Science Research and Methods* 3.2 (May 2015), pp. 399–408. ISSN: 2049-8470, 2049-8489. DOI: 10.1017/psrm.2014.32.

[33] Ramit Debnath et al. "Grounded reality meets machine learning: A deep-narrative analysis framework for energy policy research". In: *Energy Research & Social Science* 69 (Nov. 2020), p. 101704. ISSN: 22146296. DOI: 10.1016/j.erss.2020.101704.

[34] Emanuele Degani et al. "Sparse linear mixed model selection via streamlined variational Bayes". In: *Electronic Journal of Statistics* 16.2 (Jan. 2022). ISSN: 1935-7524. DOI: 10.1214/22-EJS2063.

[35] Harm Deleu, Mieke Schrooten, and Koen Hermans. "Hidden Homelessness: A Scoping Review and Avenues for Further Inquiry". In: *Social Policy and Society* 22.2 (Apr. 2023), pp. 282–298. ISSN: 1474-7464, 1475-3073. DOI: 10.1017/S1474746421000476.

[36] Paul DiMaggio, Manish Nag, and David Blei. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding". In: *Poetics* 41.6 (Dec. 2013), pp. 570–606. ISSN: 0304422X. DOI: 10.1016/j.poetic.2013.08.004.

[37] Naoki Egami et al. "How to make causal inferences using texts". In: *Science Advances* 8.42 (Oct. 2022), eabg2652. ISSN: 2375-2548. DOI: 10.1126/sciadv.abg2652.

[38] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. "Sparse Additive Generative Models of Text". In: *Proceedings of the 28 th International Conference on Machine Learning*. ICML'11. Bellevue, WA, USA, June 2011. URL: http://www.icml-2011.org/papers/534_icmlpaper.pdf.

[39] Melissa Eliot et al. “Ridge regression for longitudinal biomarker data”. In: *The International Journal of Biostatistics* 7.1 (2011), Article 37. ISSN: 1557-4679. DOI: 10.2202/1557-4679.1353.

[40] Ingo Feinerer, Kurt Hornik, and David Meyer. “Text Mining Infrastructure in *R*”. In: *Journal of Statistical Software* 25.5 (2008). ISSN: 1548-7660. DOI: 10.18637/jss.v025.i05.

[41] Cheryl Forchuk. *Homelessness Counts Website*. URL: https://publish.uwo.ca/~cforchuk/homeless_counts/.

[42] Cheryl Forchuk, Richard Booth, and Sara Husni. “Rural and Remote Homelessness in Canada: Final Report”. internal report. London, ON, Canda, Aug. 2022.

[43] Cheryl Forchuk et al. “Community Stakeholders’ Perceptions of the Impact of the Coronavirus Pandemic on Homelessness in Canada”. In: *International Journal on Homelessness* (Aug. 2023), pp. 1–17. ISSN: 2564-310X. DOI: 10.5206/ijoh.2023.3.15051.

[44] Scott D. Foster, Arūnas P. Verbyla, and Wayne S. Pitchford. “Incorporating LASSO Effects into a Mixed Model for Quantitative Trait Loci Detection”. In: *Journal of Agricultural, Biological, and Environmental Statistics* 12.2 (2007), pp. 300–314. ISSN: 1085-7117. DOI: 10.1198/108571107X200396.

[45] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1 (2010). ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01.

[46] Baojun Gao et al. “Different voices between Airbnb and hotel customers: An integrated analysis of online reviews using structural topic model”. In: *Journal of Hospitality and Tourism Management* 51 (June 2022), pp. 119–131. ISSN: 14476770. DOI: 10.1016/j.jhtm.2022.03.004.

[47] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/ hierarchical models*. Analytical methods for social research. OCLC: ocm67375137. Cambridge ; New York: Cambridge University Press, 2007. ISBN: 9780521867061.

[48] Fatemeh Ghapani. "Stochastic restricted Liu estimator in linear mixed measurement error models". In: *Communications in Statistics - Simulation and Computation* 51.3 (Mar. 2022), pp. 1220–1233. ISSN: 0361-0918, 1532-4141. DOI: 10.1080/03610918.2019.1664581.

[49] Gene H. Golub, Michael Heath, and Grace Wahba. "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". In: *Technometrics* 21.2 (1979), pp. 215–223. ISSN: 0040-1706. DOI: 10.2307/1268518.

[50] Peter Grajzl and Peter Murrell. "Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas". In: *Journal of Comparative Economics* 47.1 (Mar. 2019), pp. 111–135. ISSN: 01475967. DOI: 10.1016/j.jce.2018.10.004.

[51] Justin Grimmer. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". In: *Political Analysis* 18.1 (2010), pp. 1–35. ISSN: 1047-1987. DOI: 10.1093/pan/mpp034.

[52] Justin Grimmer and Brandon M. Stewart. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21.3 (July 2013), pp. 267–297. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mps028.

[53] Andreas Groll and Gerhard Tutz. "Variable selection for generalized linear mixed models by L 1-penalized estimation". In: *Statistics and Computing* 24.2 (Mar. 2014), pp. 137–154. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-012-9359-z.

[54] Greg Guest et al. "Comparing focus groups and individual interviews: findings from a randomized study". In: *International Journal of Social Research Methodol-*

ogy 20.6 (Nov. 2017), pp. 693–708. ISSN: 1364-5579, 1464-5300. DOI: 10.1080/13645579.2017.1281601.

[55] Andrew Heiss. *A guide to modeling proportions with Bayesian beta and zero-inflated beta regression models*. Nov. 2021. URL: https://doi.org/10.59350/wbn93-edb02.

[56] Monique Hennink and Bonnie N. Kaiser. "Sample sizes for saturation in qualitative research: A systematic review of empirical tests". In: *Social Science & Medicine* 292 (Jan. 2022), p. 114523. ISSN: 02779536. DOI: 10.1016/j.socscimed.2021.114523.

[57] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1 (Feb. 1970), pp. 55–67. ISSN: 0040-1706, 1537-2723. DOI: 10.1080/00401706.1970.10488634.

[58] Nan Hu et al. "What do hotel customers complain about? Text analysis using structural topic model". In: *Tourism Management* 72 (June 2019), pp. 417–426. ISSN: 02615177. DOI: 10.1016/j.tourman.2019.01.002.

[59] Infrastructure Canada. *2016 coordinated point-in-time count of homelessness in Canadian communities.* Cat. No.: SSD-177-01-17E. 2017. URL: https://www.infrastructure.gc.ca/homelessness-sans-abri/reports-rapports/pit-counts-dp-eng.html.

[60] Infrastructure Canada. *Everyone Counts 2018: Highlights.* Cat. No.: Em12-25/2018E-PD. 2019. URL: https://www.infrastructure.gc.ca/homelessness-sans-abri/reports-rapports/pit-counts-dp-2018-highlights-eng.html.

[61] Infrastructure Canada. *Homeless Individuals and Families Information System (HIFIS)*. Mar. 2022. URL: https://www.infrastructure.gc.ca/homelessness-sans-abri/hifis-sisa/index-eng.html.

[62] Infrastructure Canada. *Everyone Counts 2020-2022: Preliminary Highlights Report*. Cat. No. T94-54/2024E-PDF. 2024. URL: https://www.infrastructure.gc.ca/homelessness-sans-abri/reports-rapports/pit-counts-dp-2020-2022-highlights-eng.html.

[63] Karoliina Isoaho, Daria Gritsenko, and Eetu Mäkelä. "Topic Modeling and Text Analysis for Qualitative Policy Research". In: *Policy Studies Journal* 49.1 (Feb. 2021), pp. 300–324. ISSN: 0190-292X, 1541-0072. DOI: 10.1111/psj.12343.

[64] Thomas Jacobs and Robin Tschötschel. "Topic models meet discourse analysis: a quantitative tool for a qualitative approach". In: *International Journal of Social Research Methodology* 22.5 (Sept. 2019), pp. 469–485. ISSN: 1364-5579, 1464-5300. DOI: 10.1080/13645579.2019.1576317.

[65] Jiming Jiang. *Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems*. 1st ed. Chapman and Hall/CRC, 2017. ISBN: 9781315119281. DOI: 10.1201/9781315119281.

[66] Meng Jiang et al. "Phrase-level pairwise topic modeling to uncover helpful peer responses to online suicidal crises". In: *Humanities and Social Sciences Communications* 7.1 (July 2020), p. 36. ISSN: 2662-9992. DOI: 10.1057/s41599-020-0513-5.

[67] Mohammad Emtiyaz Khan and Guillaume Bouchard. *Variational EM Algorithms for Correlated Topic Models*. Sept. 2009. URL: https://emtiyaz.github.io/Writings/ctmVarEm-1.pdf.

[68] Dimitris Korobilis and Kenichi Shimizu. *Bayesian Approaches to Shrinkage and Sparse Estimation*. Dec. 2021. DOI: 10.48550/arXiv.2112.11751.

[69] Louis J. Kruger et al. "Individual interviews or focus groups? Interview format and women's self-disclosure". In: *International Journal of Social Research Methodology* 22.3 (May 2019), pp. 245–255. ISSN: 1364-5579, 1464-5300. DOI: 10.1080/13645579.2018.1518857.

[70] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. DOI: 10.1214/aoms/1177729694.

[71] Özge Kuran and M. Revan Özkale. "Gilmour's approach to mixed and stochastic restricted ridge predictions in linear mixed models". In: *Linear Algebra and its Applications* 508 (Nov. 2016), pp. 22–47. ISSN: 0024-3795. DOI: 10.1016/j.laa.2016.06.040.

[72] Anselm Küsters and Elisa Garrido. "Mining PIGS. A structural topic model analysis of Southern Europe based on the German newspaper *Die Zeit* (1946-2009)". In: *Journal of Contemporary European Studies* 28.4 (Oct. 2020), pp. 477–493. ISSN: 1478-2804, 1478-2790. DOI: 10.1080/14782804.2020.1784112.

[73] Jey Han Lau, David Newman, and Timothy Baldwin. "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Shuly Wintner, Sharon Goldwater, and Stefan Riezler. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539. DOI: 10.3115/v1/E14-1056.

[74] Joie Le. "Profile Creation with Topic Modeling and Semantic Analysis from Conversations about COVID-19 among U.S. Older Adults". B.S. Computer Science and Engineering. Cambridge, Massachusetts, USA: Massachusetts Institute of Technology, Feb. 2023. URL: https://hdl.handle.net/1721.1/150291.

[75] Patricia Leavy. *Research design: quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. New York ; London: Guilford Press, 2017. ISBN: 9781462529995.

[76] Ximing Li et al. "Group topic model: organizing topics into groups". In: *Information Retrieval Journal* 18.1 (Feb. 2015), pp. 1–25. ISSN: 1386-4564, 1573-7659. DOI: 10.1007/s10791-014-9244-9.

[77] Dan Liang et al. “Examining Senior Drivers’ Attitudes Toward Advanced Driver Assistance Systems After Naturalistic Exposure”. In: *Innovation in Aging* 4.3 (May 2020). Ed. by Richard Pak, igaa017. ISSN: 2399-5300. DOI: 10.1093/geroni/igaa017.

[78] Xu-Qing Liu and Ping Hu. “General ridge predictors in a mixed linear model”. In: *Statistics* 47.2 (Apr. 2013), pp. 363–378. ISSN: 0233-1888, 1029-4910. DOI: 10.1080/02331888.2011.592190.

[79] Luca Maestrini. *DMTWpackage*. GitHub Repository. Jan. 2022. URL: https://github.com/lucamaestrini/DMTWcode.

[80] Niccolò Manych, Finn Müller-Hansen, and Jan Christoph Steckel. “The political economy of coal across 12 countries: Analysing qualitative interviews with topic models”. In: *Energy Research & Social Science* 101 (July 2023), p. 103137. ISSN: 22146296. DOI: 10.1016/j.erss.2023.103137.

[81] David Mimno, Hanna M. Wallach, and Andrew McCallum. *Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors*. 2008. URL: https://mimno.infosci.cornell.edu/papers/sampledlgstnorm.pdf.

[82] David Mimno et al. “Optimizing Semantic Coherence in Topic Models”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Ed. by Regina Barzilay and Mark Johnson. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 262–272. URL: https://aclanthology.org/D11-1024.

[83] Adam S Miner et al. “Formally comparing topic models and human-generated qualitative coding of physician mothers’ experiences of workplace discrimination”. In: *Big Data & Society* 10.1 (Jan. 2023). ISSN: 2053-9517, 2053-9517. DOI: 10.1177/20539517221149106.

[84] Atsushi Miyaoka et al. “Emergent Coding and Topic Modeling: A Comparison of Two Qualitative Analysis Methods on Teacher Focus Group Data”. In: *Interna-*

tional Journal of Qualitative Methods 22 (Jan. 2023), p. 160940692311659. ISSN: 1609-4069, 1609-4069. DOI: 10.1177/16094069231165950.

[85] Anouk Mols, Jorge Pereira Campos, and João Fernando Ferreira Gonçalves. ""Those blimmin Ts and Cs": a mixed methods analysis of how people manage personal information, privacy, and impressions". In: *Human–Computer Interaction* (Mar. 2024), pp. 1–18. ISSN: 0737-0024, 1532-7051. DOI: 10.1080/07370024.2024.2325340.

[86] Cole C. Monnahan and Kasper Kristensen. "No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages". In: *PLOS ONE* 13.5 (May 2018). Ed. by Yong Deng, e0197954. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0197954.

[87] David L Morgan. *Focus groups as qualitative research*. 1st ed. Qualitative research methods 16. California, USA: SAGE, 1988. ISBN: 0803932081.

[88] Rasmus Munksgaard and Jakob Demant. "Mixing politics and crime – The prevalence and decline of political discourse on the cryptomarket". In: *International Journal of Drug Policy*. Drug Cryptomarkets 35 (Sept. 2016), pp. 77–83. ISSN: 0955-3959. DOI: 10.1016/j.drugpo.2016.04.021.

[89] Laura K. Nelson. "Computational Grounded Theory: A Methodological Framework". In: *Sociological Methods & Research* 49.1 (Feb. 2020), pp. 3–42. ISSN: 0049-1241, 1552-8294. DOI: 10.1177/0049124117729703.

[90] Xiao Ni, Daowen Zhang, and Hao Helen Zhang. "Variable Selection for Semiparametric Mixed Models in Longitudinal Studies". In: *Biometrics* 66.1 (2010), pp. 79–88. ISSN: 0006-341X. DOI: 10.1111/j.1541-0420.2009.01240.x.

[91] Tui H. Nolan, Marianne Menictas, and Matt P. Wand. "Streamlined Computing for Variational Inference with Higher Level Random Effects". In: *Journal of Machine Learning Research* 21.157 (July 2020), pp. 1–62. DOI: 10.48550/arXiv.1903.06616.

[92] M. Revan Özkale and Funda Can. "An evaluation of ridge estimator in linear mixed models: an example from kidney failure data". In: *Journal of Applied Statistics* 44.12 (Sept. 2017), pp. 2251–2269. ISSN: 0266-4763, 1360-0532. DOI: 10.1080/02664763.2016.1252732.

[93] José Pinheiro, Douglas Bates, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. Nov. 2023. URL: https://cran.r-project.org/web/packages/nlme/index.html.

[94] José Pinheiro and Douglas M Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. New York: Springer-Verlag, 2000. ISBN: 9780387989570. DOI: 10.1007/b98882.

[95] Dimitris Rizopoulos. *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature*. Oct. 2023. URL: https://cran.r-project.org/web/packages/GLMMadaptive/index.html.

[96] Margaret Roberts et al. "The structural topic model and applied social science". In: Neural Information Processing Society., 2013.

[97] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. "A Model of Text for Experimentation in the Social Sciences". In: *Journal of the American Statistical Association* 111.515 (July 2016), pp. 988–1003. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2016.1141684.

[98] Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for Confounding with Text Matching". In: *American Journal of Political Science* 64.4 (Oct. 2020), pp. 887–903. ISSN: 0092-5853, 1540-5907. DOI: 10.1111/ajps.12526.

[99] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. "**stm** : An *R* Package for Structural Topic Models". In: *Journal of Statistical Software* 91.2 (2019). ISSN: 1548-7660. DOI: 10.18637/jss.v091.i02.

[100] Margaret E. Roberts et al. “Structural Topic Models for Open-Ended Survey Responses”. In: *American Journal of Political Science* 58.4 (Oct. 2014), pp. 1064–1082. ISSN: 0092-5853, 1540-5907. DOI: 10.1111/ajps.12103.

[101] Michael Röder, Andreas Both, and Alexander Hinneburg. “Exploring the Space of Topic Coherence Measures”. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM ’15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324.

[102] Wendy E Rowe. *The SAGE Encyclopedia of Action Research: Positionality*. Thousand Oaks, CA, 2014. DOI: 10.4135/9781446294406.

[103] Konstantin Salomatin, Yiming Yang, and Abhimanyu Lad. “Multi-field Correlated Topic Modeling”. In: *Proceedings of the 2009 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Apr. 2009, pp. 628–637. DOI: 10.1137/1.9781611972795.54.

[104] Benjamin Saunders et al. “Saturation in qualitative research: exploring its conceptualization and operationalization”. In: *Quality & Quantity* 52.4 (July 2018), pp. 1893–1907. ISSN: 0033-5177, 1573-7845. DOI: 10.1007/s11135-017-0574-8.

[105] Jürg Schelldorfer, Peter Bühlmann, and Sara Van De Geer. “Estimation for High-Dimensional Linear Mixed-Effects Models Using L $_1$ -Penalization”. In: *Scandinavian Journal of Statistics* 38.2 (June 2011), pp. 197–214. ISSN: 0303-6898, 1467-9469. DOI: 10.1111/j.1467-9469.2011.00740.x.

[106] Patrick Schulze et al. “A Bayesian approach to modeling topic-metadata relationships”. In: *AStA Advances in Statistical Analysis* (Nov. 2023). ISSN: 1863-8171, 1863-818X. DOI: 10.1007/s10182-023-00485-9.

[107] Shaun Seaman, Menelaos Pavlou, and Andrew Copas. “Review of methods for handling confounding by cluster and informative cluster size in clustered data”.

In: *Statistics in Medicine* 33.30 (Dec. 2014), pp. 5371–5387. ISSN: 0277-6715. DOI: 10.1002/sim.6277. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4320764/.

[108] Sujeong Seo and Ernest Fokoue. "Estimation of Community Views on Criminal Justice a Statistical Document Analysis Approach". In: *Journal of Advances in Mathematics and Computer Science* 25.6 (Jan. 2018), pp. 1–21. ISSN: 24569968. DOI: 10.9734/JAMCS/2017/38582.

[109] Dhanya Sridhar and Lise Getoor. *Estimating Causal Effects of Tone in Online Debates*. June 2019. DOI: 10.48550/arXiv.1906.04177.

[110] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.32.6. 2024. URL: https://mc-stan.org/.

[111] David W. Stewart and Prem N. Shamdasani. *Focus groups: theory and practice*. Third edition. Applied social research methods series. Los Angeles: SAGE, 2015. ISBN: 9781452270982.

[112] Rajendra Subedi, Shirin Roshanafshar, and T Lawson Greenberg. *Developing Meaningful Categories for Distinguishing Levels of Remoteness in Canada*. Cat. no. 11-633-X — No. 026. Centre for Population Health Data (CPHD), Aug. 2020, p. 22. URL: https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2020002-eng.htm.

[113] Matt Taddy. "On Estimation and Selection for Topic Models". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, Apr. 2012, pp. 1184–1193. URL: https://proceedings.mlr.press/v22/taddy12.html.

[114] Jian Tang et al. "Understanding the limiting factors of topic modeling via posterior contraction analysis". In: *Proceedings of the 31st International Conference on*

International Conference on Machine Learning - Volume 32. ICML'14. Beijing, China: JMLR.org, June 2014, pp. I–190–I–198.

[115] Xueying Tang et al. "Bayesian Variable Selection and Estimation Based on Global-Local Shrinkage Priors". In: *Sankhyā: The Indian Journal of Statistics, Series A (2008-)* 80.2 (2018), pp. 215–246. ISSN: 0976-836X. DOI: 10.1007/s13171-017-0118-2.

[116] Martin A Tanner. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag, 1993. ISBN: 0387940316.

[117] Rochelle Terman. "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage". In: *International Studies Quarterly* 61.3 (2017), pp. 489–502. ISSN: 0020-8833. DOI: 10.1093/isq/sqx051.

[118] Silvia Terragni, Elisabetta Fersini, and Enza Messina. "Constrained Relational Topic Models". In: *Information Sciences* 512 (Feb. 2020), pp. 581–594. ISSN: 00200255. DOI: 10.1016/j.ins.2019.09.039.

[119] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (Jan. 1996), pp. 267–288. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[120] Federico Tomasi et al. "Stochastic Variational Inference for Dynamic Correlated Topic Models". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, Aug. 2020, pp. 859–868. URL: https://proceedings.mlr.press/v124/tomasi20a.html.

[121] Shawn Treier and Simon Jackman. "Democracy as a Latent Variable". In: *American Journal of Political Science* 52.1 (Jan. 2008), pp. 201–217. ISSN: 0092-5853, 1540-5907. DOI: 10.1111/j.1540-5907.2007.00308.x.

[122] Sara Van Erp, Daniel L. Oberski, and Joris Mulder. "Shrinkage priors for Bayesian penalized regression". In: *Journal of Mathematical Psychology* 89 (Apr. 2019), pp. 31–50. ISSN: 00222496. DOI: 10.1016/j.jmp.2018.12.004.

[123] Erik Van Zwet. "A default prior for regression coefficients". In: *Statistical Methods in Medical Research* 28.12 (Dec. 2019), pp. 3799–3807. ISSN: 0962-2802, 1477-0334. DOI: 10.1177/0962280218817792.

[124] A. M. Walker. "On the Asymptotic Behaviour of Posterior Distributions". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 31.1 (Jan. 1969), pp. 80–88. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.2517-6161.1969.tb00767.x.

[125] Rebecca Wallace, Elizabeth Goodyear-Grant, and Amanda Bittner. "Harnessing Technologies in Focus Group Research". In: *Canadian Journal of Political Science* 54.2 (June 2021), pp. 335–355. ISSN: 0008-4239, 1744-9324. DOI: 10.1017/S0008423921000226. URL: https://www.cambridge.org/core/product/identifier/S0008423921000226/type/journal_article.

[126] Chong Wang and David M. Blei. "Variational inference in nonconjugate models". In: *The Journal of Machine Learning Research* 14.1 (Apr. 2013), pp. 1005–1031. ISSN: 1532-4435.

[127] Dong Wang, Kent M. Eskridge, and Jose Crossa. "Identifying QTLs and Epistasis in Structured Plant Populations Using Adaptive Mixed LASSO". In: *Journal of Agricultural, Biological, and Environmental Statistics* 16.2 (June 2011), pp. 170–184. ISSN: 1085-7117, 1537-2693. DOI: 10.1007/s13253-010-0046-2.

[128] Ronald L. Wasserstein and Nicole A. Lazar. "The ASA Statement on *p* -Values: Context, Process, and Purpose". In: *The American Statistician* 70.2 (Apr. 2016), pp. 129–133. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.2016.1154108.

[129] Galen Weld et al. *Adjusting for Confounders with Text: Challenges and an Empirical Evaluation Framework for Causal Inference*. May 2022. DOI: 10.48550/arXiv.2009.09961.

[130] Oliver Wieczorek et al. "Mapping the field of psychology: Trends in research topics 1995–2015". In: *Scientometrics* 126.12 (Dec. 2021), pp. 9699–9731. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-021-04069-9.

[131] Mingan Yang, Min Wang, and Guanghui Dong. "Bayesian variable selection for mixed effects model with shrinkage prior". In: *Computational Statistics* 35.1 (Mar. 2020), pp. 227–243. ISSN: 0943-4062, 1613-9658. DOI: 10.1007/s00180-019-00895-x.

[132] Jieyi Yi and Niansheng Tang. "Variational Bayesian Inference in High-Dimensional Linear Mixed Models". In: *Mathematics* 10.3 (Jan. 2022), p. 463. ISSN: 2227-7390. DOI: 10.3390/math10030463.

[133] Luwei Ying, Jacob M. Montgomery, and Brandon M. Stewart. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures". In: *Political Analysis* 30.4 (Oct. 2022), pp. 570–589. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2021.33.

Appendices

Appendices

A Content Appendices

A.1 Appendix to Chapter 4: Implementation of Degani et al.'s regression in mixSTM

The streamlined mixed effects regression was implemented in the mixSTM with use of the "StreamlinedMfvbLmmTwoLevelGlobLocPrior" function as provided in the DMTWPackage on GitHub [79]. A few notes about the implementation are below.

- Convergence: The original function by Degani et al. does not implement a method to check for convergence, and instead recommends the regression be run a set number of times that is sufficiently large to lead to small updates in the parameter estimates [34]. As alternatives, they suggest to run the regression until the relative absolute change in the parameter estimates or change in the ELBO is small [34]. When iterating per topic and per EM iteration, not checking for convergence 1) increases the overall runtime for each iteration, and 2) can lead to issues in poorly identifiable models (especially if the true random effect variances are close to zero) where a set-to-be-inverted matrix becomes singular. The C++ implementation will attempt to find an approximate solution to the inversion of this matrix, but this is best avoided. Instead, a rapid convergence check was implemented to match the STM's, which stops the regression when the sum of the magnitude of the changes in the coefficients is less than 0.0005. In this case, we take the sum of the absolute values of the difference in posterior mean, $\mu_{q(\gamma,u)}$, between iterations. Like the convergence check in the STM, this method is vulnerable to changes is scale: if the scale of all the covariate data is

very small, this threshold may be reached quickly and before true convergence. To avoid this, covariates should be scaled to similar magnitudes before being entered into the model. Taking the sum across all coefficients, however, means that this is only a concern when all coefficients are small. Furthermore, the maximum number of iterations for the mixed effects regression is set to 200. In practice, even in complex models, this iteration limit was rarely reached except in the initial iterations of the EM algorithm. Setting an iteration limit is helpful to ensure the regression runs quickly, particularly for the first few iterations, as the relationship between the initial estimates of η and the covariate data is unlikely to be meaningful.

- Since the regression is run $K - 1$ times per iteration of the EM algorithm, each time with different values for the outcome, hyperparameters of the regression are set to be relatively non-informative. Modelled after Degani et al.'s simulations [34], we set default values of $\nu_\beta = 1$ as the degrees of freedom on the Half-t distribution over the residual standard deviation, $\nu_\Sigma = 2$ as one of the parameters of the distribution of G, $s_\beta^2 = s_\Sigma^2 = s_\alpha^2 = 10^5$ to grant broad priors on the variances and minimal prior information about sparsity among the coefficients, and gave the random-effect associated coefficients γ^R a broad Normal prior with mean zero and diagonal covariance with diagonal entries equal to 10^{10}.

- Although there is no requirement for a model with a group-varying (random) slope to have a random intercept, in the case of modelling topical prevalence a random slope-only model is unlikely: it would suggest that in each group of documents there is necessarily a point where the topical prevalence is the same for each group, irrespective of the trend thereafter. As such, the implementation of the random effects model always incorporates at least a groupwise intercept for each model. This intercept is part of the γ^R coefficients and so is not directly penalized, and is shared across all groups so is not vulnerable to being a small reference category (for the groups

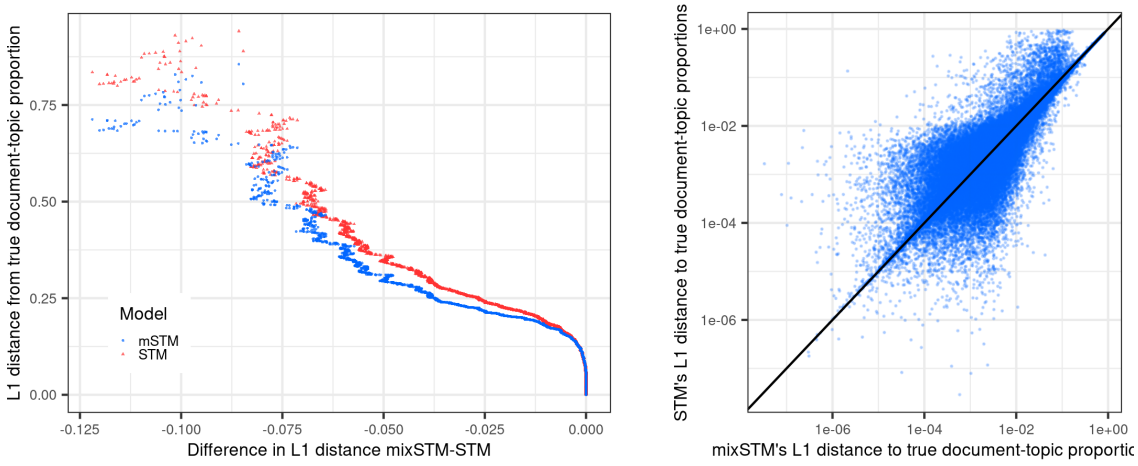
included in the random-effects model).

- The streamlined mixed effects regression as implemented by Degani et al. additionally has the capacity to include p^A fixed effect covariates which have as priors sparse Normal distributions, causing them to remain unpenalized during modelling. While this may be of interest in some applications, including as a manner to leave the intercept unpenalized in the aforementioned case where only random slopes are modelled, when used in mixSTM, $p^A = 0$. We desire all additional effects to be penalized in the regression, for reasons outlined in the text including reducing overfitting.

- When no additional covariates are of interest and the model exclusively has group-varying covariates, "StreamlinedMfvbLmmTwoLevelGaussPrior" is used to run the regression. This is the streamlined variational inference for mixed models implemented by Nolan [91] with sparse Normal priors on all coefficients.

A.2 Appendix to Chapter 4: Additional simulation graphs

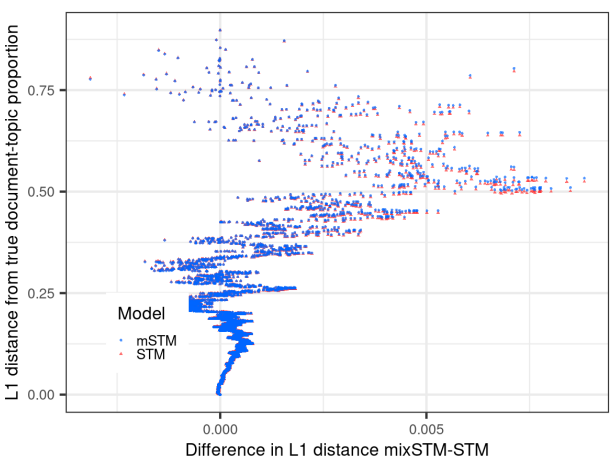
A.2.1 All Scenarios L1 Distance



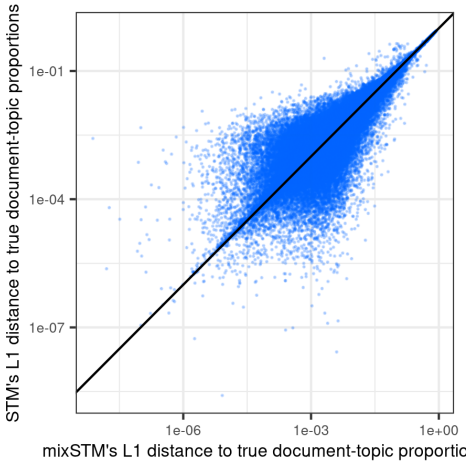
(a) Sorted L1 distances to truth for mixSTM and STM against the difference between the two L1 distances.

(b) L1 distance to true topic proportions for the mixSTM vs L1 distance to true topic proportions for STM.

Figure A.1: Scenario A1, L1 distance.

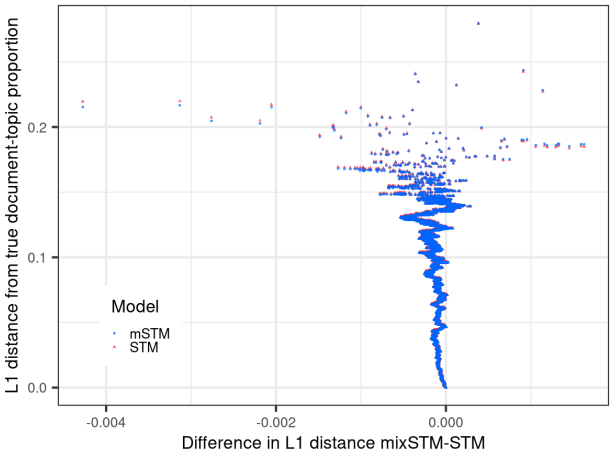


(a) Sorted L1 distances to truth for mixSTM and STM against the difference between the two L1 distances.

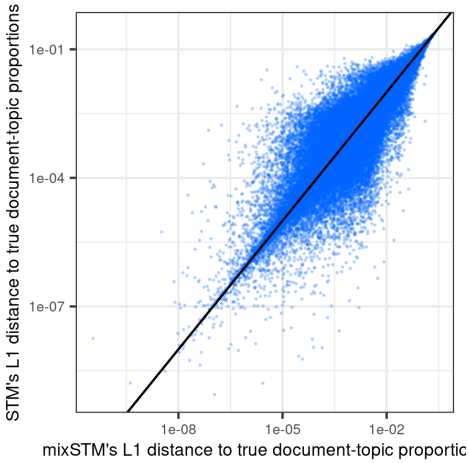


(b) L1 distance to true topic proportions for the mixSTM vs L1 distance to true topic proportions for STM.

Figure A.2: Scenario A2, L1 distance.



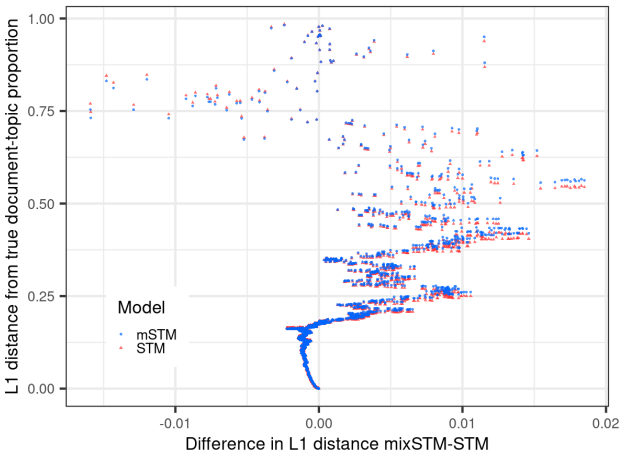
(a) Sorted L1 distances to truth for mixSTM and STM against the difference between the two L1 distances.



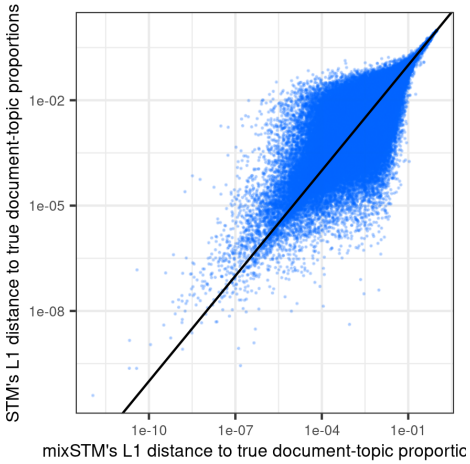
(b) L1 distance to true topic proportions for the mixSTM vs L1 distance to true topic proportions for STM.

Figure A.3: Scenario B1, L1 distance.

We can draw related conclusions to the ones presented in the main text (regarding scenario B2) for other scenarios. For scenario A1 (Fig. A.1), we can see that there is some density in favour of the STM for small L1 distances to the true topic proportions (which make up the majority of the points as seen in Chapter 4), however when the distances are larger, the mixSTM obtains universally closer L1 distances to the truth than the STM. For scenario A2



(a) Sorted L1 distances to truth for mixSTM and STM against the difference between the two L1 distances.

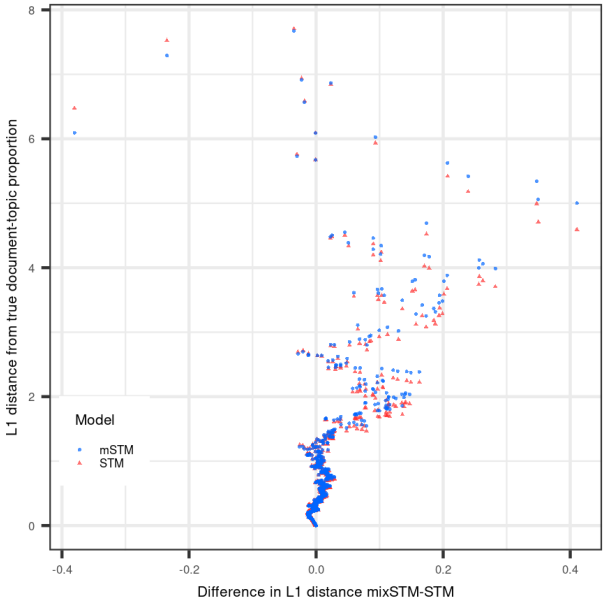


(b) L1 distance to true topic proportions for the mixSTM vs L1 distance to true topic proportions for STM.

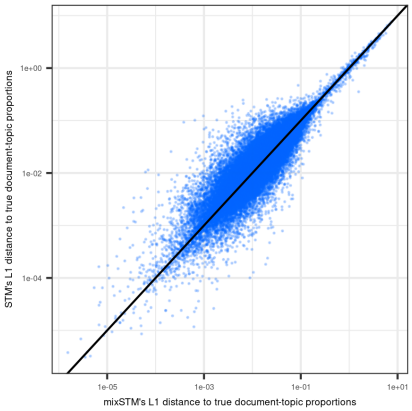
Figure A.4: Scenario B2, L1 distance.

(Fig. A.2), the mass of the distribution of differences seems similar, but the mixSTM still appears to achieve smaller values than the STM for the very smallest and the very largest L1 distances. For scenario B1 (Fig. A.3), the fit is good for both models with few extremely incorrect topic proportions, and the mass of the difference is in favour of the mixSTM for many of the estimates.

A.2.2 KL Divergence for Scenario B2



(a) Sorted KL divergences between truth and MAP estimates for mixSTM and STM against the difference between the two KL divergences.



(b) KL divergence between true and MAP estimates of topic proportions for the mixSTM vs for STM.

Figure A.5: Scenario B2, KL Divergence.

A trend is more difficult to see when looking at plotted KL divergences (Fig. A.5) than when looking at L1 distance (Fig. A.4). In general, there appears to be larger mass above a line going through (0,0) and (1,1) (so in favour of the mixSTM retrieving smaller KL divergences to the true values than the STM). Similarly to with the L1 distances, in Figure A.5a, we can see that the mixSTM retrieves smaller very small KL divergences (which make up the majority of the distances), with some alternation between the two models for larger KL divergences.

A.3 Appendix to Chapter 4: Example words from simulation B2.70

Table A.1: Summary of topic model output B2.70: A topic model with 5 topics, 1000 documents and a 599 word dictionary.

STM	mixSTM
Topic 1 Top Words: Prob: w90, w508, w179, w539, w481, w107, w598 FREX: w147, w90, w539, w395, w218, w527, w430 Lift: w550, w543, w370, w258, w305, w218, w398 Score: w539, w90, w508, w179, w481, w598, w218	Topic 1 Top Words: Prob: w90, w508, w179, w539, w481, w107, w598 FREX: w147, w395, w539, w90, w218, w527, w430 Lift: w550, w543, w370, w258, w398, w305, w218 Score: w539, w90, w508, w179, w481, w598, w218
Topic 2 Top Words: Prob: w102, w221, w54, w48, w16, w515, w136 FREX: w48, w102, w16, w221, w321, w254, w54 Lift: w541, w48, w375, w199, w287, w102, w16 Score: w102, w254, w16, w48, w54, w541, w221	Topic 2 Top Words: Prob: w102, w221, w54, w48, w16, w136, w515 FREX: w48, w102, w16, w221, w321, w254, w54 Lift: w541, w48, w287, w375, w102, w199, w16 Score: w102, w16, w254, w48, w54, w541, w221
Topic 3 Top Words: Prob: w255, w44, w122, w456, w311, w173, w46 FREX: w255, w327, w243, w215, w122, w311, w44 Lift: w222, w424, w354, w215, w309, w327, w193 Score: w255, w44, w222, w122, w327, w243, w215	Topic 3 Top Words: Prob: w255, w44, w122, w456, w311, w173, w46 FREX: w255, w327, w243, w215, w122, w311, w44 Lift: w222, w424, w354, w215, w309, w327, w193 Score: w255, w44, w222, w122, w327, w311, w243
Topic 4 Top Words: Prob: w107, w586, w345, w336, w101, w325, w558 FREX: w442, w101, w237, w558, w540, w345, w75 Lift: w153, w317, w442, w597, w40, w514, w451 Score: w101, w442, w153, w75, w237, w597, w540	Topic 4 Top Words: Prob: w107, w586, w345, w336, w101, w325, w558 FREX: w442, w101, w237, w558, w345, w75, w540 Lift: w153, w317, w442, w597, w40, w514, w451 Score: w101, w442, w153, w597, w237, w75, w540
Topic 5 Top Words: Prob: w551, w266, w249, w358, w159, w190, w455 FREX: w551, w159, w278, w455, w303, w570, w23 Lift: w457, w247, w278, w518, w313, w359, w203 Score: w551, w159, w457, w278, w358, w453, w266	Topic 5 Top Words: Prob: w551, w266, w249, w358, w159, w190, w455 FREX: w551, w159, w278, w455, w303, w23, w570 Lift: w457, w247, w278, w518, w359, w203, w313 Score: w159, w551, w457, w278, w358, w266, w453

Summary of topic model output in terms of words in each topic that have the highest probability β ("Prob"), that are the most frequent and exclusive ("FREX"), that have the highest ratio of the topic-word distribution to the overall word count distribution ("Lift" [113]), and that maximize the "score" of [26], for the STM and mixSTM for a randomly chosen simulation (number 70) in scenario B2. Topic interpretability is first assessed by looking at words; by comparing high probability and exclusive words, we can perform an initial comparison of the topic model outputs.

By default in the `stm` package in R, the `summary` of a fitted structural topic model presents the top seven most probable words for each topic. Table A.1 presents this summary output for a randomly selected simulation from scenario B2. The topics described by these summaries are composed of the same words for the STM and mixSTM models, with occasional order changes for relatively lower probability words (e.g., the sixth and seventh highest probability words for Topic 2 are w515 (6th) and w136 (7th) in the STM, but w136 (6th) and w515 (7th) in the mixSTM). In terms of interpretation, this would be unlikely to meaningfully change the assigned label for Topic 2, but it is also difficult to compare abstracted topics like those generated in simulation.

A.4 Appendix to Chapter 6: Implementations in R

Existing packages in R can be leveraged to perform the post hoc regression on topic prevalence.

For a frequentist-style mixed model fit, we used the `glmmTMB` package [20], which is very flexible. `glmmTMB` uses the Laplace approximation to the intractable integrals in the likelihoods of generalized linear models [20]; user-set priors on the parameters can be incorporated. We chose this package as both linear and beta regression models can be implemented by specifying either `gaussian(link = "identity")` or `glmmTMB::beta_family(link = "logit")` for the regression family. Furthermore, the associated package `tmbstan` [86] is capable of performing post hoc sampling on the posterior of the coefficients using the likelihood, accomodating Avenue (2).

For a Bayesian-style mixed model fit, we use the `brms` package [21], which is highly customizable. Both linear regressions and Bayesian beta regressions are possible through the specification of a family and link function: `family=Gaussian()` (which sets a default identity link) for the linear model, or `family=Beta(link="logit", link_phi="log")` for the beta regression. The priors on each parameter of the regression can be set by the user. Notably, the default priors for `brms` are in fact flat priors on the regression coefficients; in our implementation, we specify Gaussian priors on the fixed effect coefficients and intercept.

Although these were our choices for implementation, we would be remiss not to mention a couple of alternatives:

1. `lme4` [6] and `nlme` [94, 93] are two R packages commonly used for fitting linear and generalized linear regression models. For use in a linear model where none of the uncertainty over θ is going to be propagated into the coefficients, this is a rapid alternative that can be performed directly on the MAP estimates of θ extracted at

the end of the model. Neither `lme4` nor `nlme` can perform beta regression. For use in a "global" or "local" uncertainty propagation, they can be used easily in the case of the asymptotic distribution of the maximum likelihood estimators. To the best of our knowledge, there is no package which implements a post hoc MCMC on these models, so if Avenue (2) were desirable, the sampler would have to be developed from scratch.

2. `adaptiveGLMM` offers an alternative to the mixed beta regression implemented in `glmmTMB` which instead uses the adaptive Gauss-Hermite quadrature rule to approximate the likelihood [95]. To our knowledge, no post hoc MCMC sampler has been written for this package. Biswas compares `adaptiveGLMM` and `glmmTMB` in [12].

A.5 Appendix to Chapter 7 11-topic mixSTM results

Topic 1: Databases for homelessness.

- Highest Prob: hifi, use, list, program, case, manag, databas, track, work, shelter, year, access, thing, coordin, now
- FREX: manag, databas, hifi, case, list, track, coordin, program, use, bynam, intak, fund, year, process, report

Topic 2: Ways in which people are underhoused.

- Highest Prob: stay, couch, surf, live, famili, friend, referr, month, often, come, thing, sometim, need, hous, system
- FREX: couch, stay, surf, friend, referr, famili, live, month, often, address, hidden, sometim, may, rent, medic

Includes overcrowding, couch surfing, vehicle homelessness, rental discussions.

Topic 3: Keeping track of PEH.

- Highest Prob: see, hous, someon, name, access, hmis, hospit, assess, inform, individu, fall, famili, even, system, count
- FREX: hmis, name, hous, someon, assess, hospit, see, fall, access, inform, individu, pick, caa, famili, outreach

Includes discussion of hospital data, geography, point-in-time counts.

Topic 4: Sharing information between agencies.

- Highest Prob: inform, agenc, use, share, client, consent, hifi, servic, provid, system, access, individu, work, differ, report
- FREX: agenc, share, inform, consent, use, provid, hifi, client, report, sign, intern, challeng, system, pull, sort

Topic 5: Health, mental health and addictions.

- Highest Prob: health, issu, mental, care, addict, area, access, servic, individu, work, use, client, mayb, trauma, mean
- FREX: health, mental, issu, addict, care, area, access, individu, trauma, servic, client, use, drug, mayb, depend

Topic 6: Programs and locations.

- Highest Prob: come, place, work, street, outreach, back, find, shelter, program, reach, tri, good, sleep, safe, servic
- FREX: street, place, come, outreach, pay, back, reach, safe, sleep, find, work, hotel, space, good, program

Some documents with a high proportion of this topic relate to observed changes in services due to the COVID-19 pandemic.

Topic 7: (broad)

- Highest Prob: indigen, thing, popul, see, differ, talk, park, hard, still, much, camp, bit, live, look, need
- FREX: indigen, park, popul, camp, hard, everyth, much, white, notic, still, differ, thing, bit, talk, build

Topic 8: Definitions of homelessness and lived experience.

- Highest Prob: homeless, definit, person, work, time, support, point, servic, live, count, may, mani, way, feel, experi
- FREX: definit, homeless, person, point, support, time, experi, convers, feel, may, count, live, mani, pretti, work

Topic 9: Evictions/tenancy and question-asking.

- Highest Prob: call, question, connect, talk, team, see, hear, now, servic, thing, evict, put, ask, need, tenant
- FREX: call, question, hear, team, connect, talk, tenant, evict, senior, see, clinic, put, now, four, direct

Topic 10: Populations in homlessness.

- Highest Prob: youth, women, home, shelter, famili, live, children, now, young, see, popul, work, age, men, kid
- FREX: youth, women, children, young, home, age, famili, men, live, adult, violenc, kid, school, shelter, singl

Topic 11: Locations to visit for interviews and services.

- Highest Prob: day, shelter, open, work, now, past, releas, home, servic, abl, come, thing, see, challeng, time
- FREX: day, open, past, shelter, releas, now, welcom, work, challeng, temporari, institut, home, abl, count, meal

B Ethics Approval

Ethics approval for the Homelessness Counts project was obtained from the Western Research Ethics Board, WREM number 116555.

Curriculum Vitae

Name:	Pascale Nevins
Post-Secondary Education and Degrees:	University of Ottawa Ottawa, Ontario, Canada 2017 - 2022 Honours B.Sc. Major Biochemistry, Minor Statistics
Honours and Awards:	Canada Graduate Scholarship (Master's) NSERC 2022-2023
Related Work Experience:	Research Assistant Ottawa Hospital Research Institute 2020 - 2024

Selected Publications:

Nevins P, Ryan M, Davis-Plourde K, et al. Adherence to key recommendations for design and analysis of stepped-wedge cluster randomized trials: A review of trials published 2016-2022. Clin Trials. 2024 Apr;21(2):199-210. doi: 10.1177/17407745231208397.

Vanderhout S, Nevins P, Nicholls SG, et al. Patient and public involvement in pragmatic trials: online survey of corresponding authors of published trials. CMAJ Open. 2023 Sep 19;11(5):E826-E837. doi: 10.9778/cmajo.20220198.

Nevins P, Davis-Plourde K, Pereira Macedo JA, et al. A scoping review described diversity in methods of randomization and reporting of baseline balance in stepped-wedge cluster randomized trials. J Clin Epidemiol. 2023 May;157:134-145. doi: 10.1016/j.jclinepi.2023.03.010.

Nevins P, Nicholls SG, Ouyang Y, et al. Reporting of and explanations for under-recruitment and over-recruitment in pragmatic trials: a secondary analysis of a database of primary trial reports published from 2014 to 2019. BMJ Open. 2022 Dec 9;12(12):e067656. doi: 10.1136/bmjopen-2022-067656.

Nevins P, Vanderhout S, Carroll K, et al. Review of pragmatic trials found that multiple primary outcomes are common but so too are discrepancies between protocols and final reports. J Clin Epidemiol. 2022 Mar;143:149-158. doi: 10.1016/j.jclinepi.2021.12.006.