New Perspectives on Non-negative Matrix Factorization for Grouped Topic Models

by

Gabriel C. Phelan

B.Sc., Rochester Institute of Technology, 2017

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2020

Approval

Name:	**Gabriel C. Phelan**
Degree:	**Master of Science (Statistics)**
Title:	**New Perspectives on Non-negative Matrix Factorization for Grouped Topic Models**
Examining Committee:	**Chair:** Jinko Graham Professor
	David Campbell Senior Supervisor Professor School of Mathematics and Statistics Carleton University
	Lloyd T. Elliott Supervisor Assistant Professor
	Thomas M. Loughin Internal Examiner Professor
Date Defended:	**August 18, 2020**

Abstract

Probabilistic topic models (PTM's) have become a ubiquitous approach for finding a set of latent themes ("topics") in collections of unstructured text. A simpler, linear algebraic technique for the same problem is non-negative matrix factorization: we are given a matrix with non-negative entries and asked to find a pair of low-rank matrices, also non-negative, whose product is approximately the original matrix. A drawback of NMF is the non-convex nature of the optimization problem it poses. Recent work by the theoretical computer science community addresses this issue, utilizing NMF's inherent structure to find conditions under which the objective function admits convexity. With convexity comes tractability, and the central theme of this thesis is the exploitation of this tractability to ally NMF with resampling-based nonparametrics. Our motivating example is one in which a document collection exhibits some kind of partitioning according to a discrete, indexical covariate, and the goal is to assess the influence of this partitioning on document content; we call this scenario a grouped topic model. Computation relies on several well-studied tools from numerical linear algebra and convex programming which are especially well suited for synthesis with permutation tests and the bootstrap. The result is a set of simple, fast, and easily implementable methodologies for performing inference in grouped topic models. This is contrast to parallel developments in PTM's where ever-more cumbersome inference schemes are required to fit complex graphical models.

Keywords: Non-negative Matrix Factorization, Linear Algebra, Topic Models, Convex Programming, Resampling, Matrix Balancing

Acknowledgements

Let's begin with the obvious. This work would have never happened without the watchful eye of my wonderful supervisor, Dave Campbell. Upon arrival at SFU, many faculty members waxed poetic about Dave's mentorship skills and my luck in having secured him as my advisor. What Dave saw in me, I'll never be quite sure. He is fond of calling mathematics "science fiction," while I will talk your ear off about geometry. Was there some kind of mistake? Jokes aside, everything Dave's colleagues said of him were dead-on correct. Foremost, Dave is an exemplary human being. He continues to look out for me in ways that extend far beyond academia and my completion of this program. We've joked together, shared meals, and spoken with an honesty and openness that is often lost in the power dynamics of student-teacher relationships. In short, he always addressed me as Gabe the person, not Gabe the student. This was and remains incredibly touching. The wisdom of his guidance was most evident in the scarcely believable amount of leeway I was given. As in, "have you ever seen a leash this long?" We began work together in close collaboration, sharing ideas and checking in at fairly regular intervals. This was fine, but I think Dave perceptively noticed that I'm the type who needs space, time, and reams of blank paper to let ideas stew. He let me forge my own path, and this thesis is a direct result of his trust in that process. Week after week he'd listen attentively as I gesticulated[1] my way through ideas that took us far afield from the existing plan. His support never wavered. I suspect other supervisors would have gone the safe route, suppressing my incessant desire to tear up the playbook and go searching for better alternatives. Instead, he's let me create a work I'm truly proud of. He even let me include proofs! Go figure. Thanks Dave. I raise my glass to you.

I would also like to thank Lloyd, Tom, and Jinko for their roles on my committee. Talking with Lloyd was a constant source of comfort; he entertained me with discussions of everything from Radiohead to Turing machines. Tom will be glad to know that some of his dedicated frequentism might have seeped into my brain; this project started very Bayesian and emerged nothing of the sort. Jinko should appreciate the role that computation plays in this thesis, especially in relation to linear algebra. Her course emphasized the importance

[1]His move to Carleton meant we had plenty of time to practice videoconferencing before the arrival of COVID.

of this beautiful subject, so often relegated to a handy bookkeeping device. I thank the rest of the SFU faculty as well, but give special mention to Rachel, Richard, and Marie – Rachel for her masterful pedagogy, Richard for his ever-present love of mathematics, Marie for coordinating swaths of newly minted TA's, and the three of them for their kindness. There are also all those who have gone before: John T. Whelan, who taught me that intuition is more important than rigour; Charles Hamperian, who made me fall in love with randomness; Sara y Godys Armengot, quienes me regalaron esta habilidad preciosa.

I saved the best for last. My family has been a constant source of love over the course of my life, and despite the invaluable contributions that everyone above has made, I'm afraid they'll have to settle for second place. When I think of my **Dad** I think of a bastion of humanity and humility; unbreakable principles and unflappable serenity; a pinch of salt, maybe some lemon, and if it ain't DOP, it ain't for me; Jeter, the 'Mick, the ballpark in the Bronx; *Liriodendron* – one of the few that I *can* identify; serpents, turtles, and the baron of the Pine Barrens; the books I haven't had time to finish, and the books I haven't had time to start. Your contribution to this thesis is in teaching me the value of technique. When I think of my **Mom** I think of endless generosity; empathy, selflessness, finding the good in us all; porches, cats, and steaming cups of joe; thinking outside the box, but not outside a good bottle of red; northern Maine, with its fresh air and your dear friend Claire; the strokes of Rembrandt, the *clicks* of Ellen Mark, Goldsworthy, Maier, the power of art; the engineer you've always been; knowing what's important in life. Your contribution to this thesis is in teaching me how to think like a mathematician. When I think of **Cecile** I think of a better world; "We have to build the Republic of Heaven where we are, because for us there is no elsewhere."; food grown by you, cooked by me, and enjoyed together; communing with nature, the wind in a tree, your bountiful forest – our balcony!; Te lo digo en español, tu réponds en français; mounds of books, from François Cheng to bell hooks; our favourite places: upstate New York, Aubenas, the great(est) state of Massachusetts; the simple beauty of your deeds, the intricate beauty of your piano strokes; resisting the "madness of modern life"; love as a verb. Your contribution to this thesis is in teaching me to take nothing for granted.

Dedication

To Sadika, Charlene, Kelly, and Jay, whose work is vital. They are the stars of our department. Let us never forget that, and show them all the appreciation they deserve. Thank you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent decades have witnessed a paradigm shift in statistics in which the central notion of data has expanded to include such diverse objects as images, audio recordings, and unstructured text. Not incidentally, this shift has fostered an increasing overlap between statistics and neighbouring fields like computer science and electrical engineering (EECS), the traditional homes of these non-standard data types. The relation is symbiotic; EECS researchers have discovered the power of statistical inference in their engineering pursuits, while statisticians uncover insights from previously untapped data sources. In this thesis we will assume the latter role, using inference to gain insights into the messy world of natural language. A popular approach for doing so asserts that a collection of documents can be reduced to a low-dimensional "topic" representation; we can broadly define such approaches as constituting the field of *topic modelling*. Loosely, this can be thought of as unveiling the central themes that permeate a collection of documents. Once compressed in this way, documents can be sorted, searched, and further manipulated for downstream information processing tasks. They can also be mined for underlying regularities that provide statistical insights into the nature of the content.

The goals of this thesis are twofold. First, we explore how the above can be achieved via relatively new perspectives on the famous problem of *non-negative matrix factorization* (NMF). This lays the groundwork for our main contribution, which is the marriage of simulation-based inference with NMF in order to analyse possible grouping structure over the documents; we call this setting *grouped topic models*. Our archetype for this problem is a collection of texts from a common source whose content varies geographically. We might intuit, for example, that documents written in different locations exhibit some sort of statistical variation. This concrete perspective will inform our methods and examples, but we always keep the general picture in mind so as to underscore the broad applicability of the results. Our methods are inferentially justified via a shift in perspective in which we treat the output to NMF as a *statistic* which can then be passed as input to statistical resampling algorithms. This mimics the spirit of bootstrap approaches to inference (in which one

computes a useful summary statistic and then resamples the original data to approximate its sampling distribution) with which the reader is likely familiar. Crucially, this style of nonparametric inference is valid under weak assumptions on the data-generating mechanism – we need not specify a complicated probabilistic graphical model. Finally, we show these methods in action by applying them to a data set of online beer reviews from across Canada. NMF reveals the underlying topics that recur in Canadians' beer reviews, while simulation-based inference shows how these topics (and by extension flavours, preferences, etc.) differ geographically.

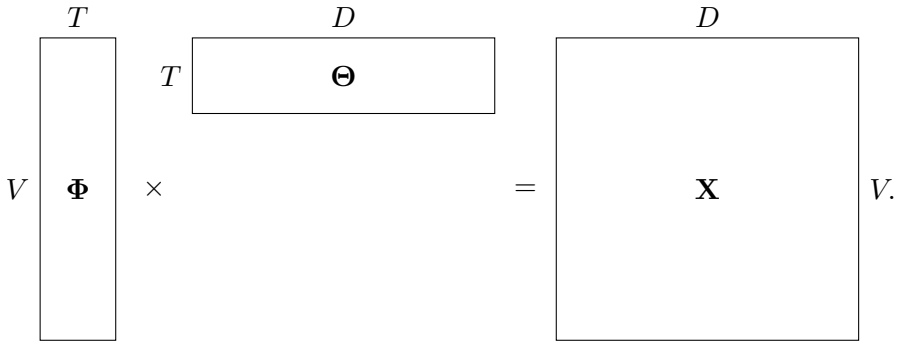
The thesis is organized as follows. Chapter 2 gives a brief account of non-negative matrix factorization and its ability to find latent structure in text data. We draw connections to probabilistic topic models and explain some of the difficulties that arise therein when trying to account for group structure. Chapter 3 discusses recent work by the theoretical computer science community in which certain natural conditions on the NMF solution lead to tractable algorithms that find unique global minima. The highlight of this chapter is the realization that a key step in this procedure can be computed and explained with a standard matrix decomposition from numerical linear algebra. Chapter 4 treats inference, giving a brief introduction to the relevant simulation-based inference methodologies and formalizing the notion of a grouped topic model. We then develop a Monte Carlo permutation test and bootstrapping algorithm to complement our purely descriptive treatment of NMF. Chapter 5 applies our methodology to "the beer data." This is also the chapter in which we address some of the more practical concerns users of our methods may face. Conclusions and future research directions follow in chapter 6.

Chapter 2

Non-negative Matrix Factorization and Related Ideas

2.1 NMF and its Uses in Text Analysis

An important problem in modern computational statistics is the factorization of a matrix $\mathbf{X}$ with non-negative entries into two "simpler" matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$, also containing non-negative entries; standard references include the work of Lee and Seung [LS96, LS99, LS01], or any number of statistical learning texts. "Simpler" is generally taken to mean low-rank; this condition is enforced by seeking a factorization whose inner dimension is much smaller than its outer dimension. Pictorially, *non-negative matrix factorization* (NMF) can be represented as follows:



The reason this abstract linear algebra problem has become a standard tool for practitioners is that it naturally represents non-negative data matrices as linear combinations of a small number of representative vectors. These representative vectors constitute low-rank structure that provide useful summaries of the original, high-dimensional data. We illustrate this fact with an example from the field of text analysis. While NMF has been applied in fields like astronomy [BHPJ10], genetics [SOAC+18], and computational vision [DLP+16], its role in text analysis is particularly illuminating. Furthermore, this will ultimately be our domain of application, thus framing things as such provides more than an informative example – it

Notation	Description
V	Number of words in a vocabulary.
D	Number of documents in a corpus.
T	Number of topics.
R	Number of regions.
$\mathbf{X}, \mathbf{Y}, \ldots$	Matrices.
$\mathbf{y}^i$	The ith row of $\mathbf{Y}$ (taken to be a column vector).
$\mathbf{y}_i$	The ith column of $\mathbf{Y}$ (a column vector).
y_{ij}	The (i,j)th entry of $\mathbf{Y}$ (a scalar).
$\mathbf{x}, \mathbf{y}, \ldots$	Vectors.
x_i	The ith entry of $\mathbf{x}$ (a scalar).
$\alpha, \beta, \ldots$	Scalars.
$\succ, \prec, \succeq, \preceq$	Element-wise inequalities for multivariate objects.
$\textsc{Proj}_{\mathbf{Y}}\mathbf{v}$	The projection of $\mathbf{v}$ onto the column space of $\mathbf{Y}$.
$\textsc{Proj}_{\mathcal{Y}}\mathbf{v}$	The projection of $\mathbf{v}$ onto the subspace $\mathcal{Y}$.
$\mathbb{Z}_+, \mathbb{Z}_-, \mathbb{R}_+, \mathbb{R}_-$	Non-negative/non-positive integers and real numbers.
$(\quad)^\star$	The solution to an optimization problem.

Figure 2.1: Common notational conventions. Departure from these conventions will be explicitly mentioned in the text.

sets the stage for our development at large. Consider a collection of D documents (henceforth referred to as a *corpus*) in which V total words are used. These V words make up a *vocabulary* which can formally be viewed as the set $\{1, 2, \ldots, V\}$ whose members stand in for the words proper. We then construct $\mathbf{X}$ as a so-called *term-document matrix* (TDM). That is, the columns of $\mathbf{X}$ represent documents by specifying each entry x_{ij} as the number of times word i appears in document j (see figure (2.1) for a summary of commonly used notation). This is often called a "bag of words" representation for it reduces a corpus to a series of in-document word counts. Evidently, much information is destroyed in this process and using the TDM would be wholly inappropriate for many language processing tasks. Nonetheless, much insight can be extracted from it and it demonstrates the remarkable ubiquity with which data can be encoded using the objects of linear algebra.

Now, suppose $\mathbf{X}$ (whose entries are not only non-negative but integral) can be factored into the product of two non-negative matrices $\mathbf{\Phi}, \mathbf{\Theta}$ with inner dimension $T \ll V$. Mathematically, we have

$$\begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_D \\ | & | & & | \end{bmatrix}_{V \times D} = \begin{bmatrix} | & | & & | \\ \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \ldots & \boldsymbol{\phi}_T \\ | & | & & | \end{bmatrix}_{V \times T} \begin{bmatrix} | & | & & | \\ \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \ldots & \boldsymbol{\theta}_D \\ | & | & & | \end{bmatrix}_{T \times D}$$

where we have emphasized the column-wise view of these matrices. A single document $\mathbf{x}_i$ can then be written as $\mathbf{x}_i = \mathbf{\Phi}\boldsymbol{\theta}_i$, or equivalently

$$\mathbf{x}_i = \theta_{i1}\boldsymbol{\phi}_1 + \theta_{i2}\boldsymbol{\phi}_2 + \ldots + \theta_{iT}\boldsymbol{\phi}_T.$$

This reveals the motivation for such a factorization: the documents $\{\mathbf{x}_i\}$ can be written as linear combinations of the columns of $\mathbf{\Phi}$. In the context of text analysis, these columns are called *topics*, a formalization of the intuitive notion that documents can be decomposed into a set of underlying ideas or themes. Because each $\boldsymbol{\phi}_i$ contains V entries, all of which are non-negative, $\boldsymbol{\phi}_i$ can be thought of as a vector of pseudo-counts for each word in topic i ($x_{ij} \in \mathbb{Z}_+$, but we usually have $\phi_{ij} \in \mathbb{R}_+$). Similarly, $\boldsymbol{\theta}_i$ encodes how much of each topic comprises document i – a set of topic weights. A concrete example is helpful in clarifying these ideas. Suppose a topic is described by the following table:

word	pseudo-count within topic
equation	10.11
linear	6.05
Fourier	1.02
⋮	⋮
spaghetti	.0000001
⋮	⋮

We might call this the "mathematics" topic; a document which gives this topic high weight is likely to concern mathematics. Such a document might ultimately be composed of topics like[1]

$$1.5 \times \text{mathematics} + .2 \times \text{physics} + .2 \times \text{France} + \ldots + 0 \times \text{Italian food},$$

perhaps indicating that it is about the famous French mathematician Joseph Fourier. Every document in the corpus can be written as a similar linear combination, their differences deriving from the different topic weights as encoded in $\mathbf{\Theta}$. Many interesting engineering and inference tasks rely on finding topics given only access to $\mathbf{X}$; we describe this next.

2.2 NMF as Optimization and Connections to Probabilistic Topic Models

So far we have supposed the existence of an exact factorization $\mathbf{X} = \mathbf{\Phi\Theta}$, $\mathbf{X}$ being generated in the forward direction in terms of $\mathbf{\Phi}$ and $\mathbf{\Theta}$. In practice, we would like to solve the inverse

[1]Note explicitly writing a topic name like "physics" here is used for clarity. No such labels are encountered in practice and one must look at the words within a topic to determine (subjectively) its significance.

problem. We observe $\mathbf{X}$ and seek to reconstruct the non-negative factors, thus revealing the thematic structure of the corpus. An exact solution is too much to hope for, so we instead solve the optimization problem

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{X} - \mathbf{\Phi\Theta}\|_{\mathrm{F}} \\ &\text{subject to} \quad \mathbf{\Phi}, \mathbf{\Theta} \succeq 0 \end{aligned} \tag{2.1}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm[2]. This gives an approximate factorization $\mathbf{X} \approx \mathbf{\Phi}^\star\mathbf{\Theta}^\star$. This mathematical program is non-convex and thus NP-hard in general [Vav10]. One strategy for dealing with this is to use the singular value decomposition (SVD) instead [DDF$^+$90], a historically older and successful approach but one which lacks the concrete interpretation of NMF. Another is in keeping with modern practices in machine learning where heuristic algorithms are applied with disregard for the formal hardness results, often finding "good" local minima anyway. One such scheme for NMF is a simple alternating minimization algorithm in which we iteratively optimize over one factor while holding the other fixed until convergence to a stationary point [LS01]. While giving sensible solutions in practice, this algorithm suffers from all the typical trappings of applying heuristics in the non-convex setting. Most notably, local minima are not guaranteed to be globally optimal. One of the main thrusts of the current work is to build on recent results showing that NMF can be cast as a convex program under certain conditions. These results are crucial in that NMF becomes tractable, meaning a global minimum can be found efficiently. Embedding NMF in the framework of non-parametric inference requires this tractability for it allows the output of an algorithm to be viewed as a statistic under a suitable sampling model. The multiplicity of local minima exhibited by non-convex problems prohibit such an interpretation since the associated optimization algorithms must inevitably employ randomness (either in the initialization or the optimization itself). Thus running the algorithm multiple times may well produce different results (i.e., different local minima are reached), and so there is no reasonable sense in which the output can be considered a statistic.

Supposing recovery of $\mathbf{\Phi}^\star$ and $\boldsymbol{\theta}^\star$, a useful probabilistic interpretation is obtained by normalizing the columns of these two matrices so that they sum to 1. Column normalization allows us to discard the somewhat unintuitive notions of pseudo-counts and topic weights in favour of probabilities. In this way, a document $\mathbf{x}_i$ can be thought of as a random sample from a mixture distribution of the T topics $\{\boldsymbol{\phi}_i\}$ (themselves now regarded as distributions over words). We emphasize that this is an interpretative device only – so far no stochasticity of any kind has been assumed. *Probabilistic topic models* (PTM) are a class of models which take this probabilistic interpretation quite literally, considering documents as generated ac-

[2]This is essentially an element-wise ℓ_2 norm for matrices.

cording to mixture distributions of the aforementioned kind. In this setup, $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are taken to be parameters or latent variables in a statistical model; the specific distributional assumptions involved give rise to various well-known approaches such as Probabilistic Latent Semantic Indexing [Hof99], Latent Dirichlet Allocation [BNJ03] and Correlated Topic Models [BL05]. Though these models are popular and effective, computational drawbacks abound for general PTM's. Approximate inference techniques like Markov chain Monte Carlo and variational inference are standard tools in this regime, but both are notoriously difficult to implement successfully in practice. Black box variational inference (BBVI) is a promising line of work for automating inference in PTM's (and large-scale probabilistic models more generally) [RGB14], but remains heuristic in nature and is currently only implemented in deep learning libraries that are often cumbersome for statisticians to use. Even when computation is not an issue, these techniques exhibit a host of difficulties. The accuracy of variational methods can be questionable, and is generally unknown to the user. While theoretically accurate in the limit of infinite computation, sampling-based inference struggles to handle multi-modal posteriors – a hallmark of PTM's. See for example Griffiths and Steyvers [GS04], who in reference to using Gibbs sampling for inference in Latent Dirichlet Allocation (LDA) assert:

> These estimates cannot be combined across samples for any analysis that relies on the content of specific topics. This issue arises because of a lack of identifiability. Because mixtures of topics are used to form documents, the probability distribution over words implied by the model is unaffected by permutations of the indices of the topics. Consequently, no correspondence is needed between individual topics across samples; just because two topics have index j in two samples is no reason to expect that similar words were assigned to those topics in those samples. However, statistics insensitive to permutation of the underlying topics can be computed by aggregating across samples.

The main consequence of these drawbacks is that PTM's are difficult to flexibly extend. Clever sampling algorithms that work for a particular model rely heavily on that model's structure, and break down with the introduction of new variables and/or dependencies. Likewise, deriving bespoke variational methods for new models is difficult and error-prone. This rigidity is what black-box approaches were designed to address, but experience suggests that the term "black box" is as yet unwarranted.

Despite these barriers, recent progress has been made in the development and implementation of large-scale PTM's which incorporate extra structure (like the document groupings that are of particular interest to us). Building on previous work [MM08, EAX11], *Structural Topic Models* (STM) provide a unified approach to PTM's with covariate effects modifying both the topics themselves and their allocation within the corpus [RSA16]. This is in contrast with the reverse construction, exemplified by *Supervised Topic Models* [BM07], in

which topics are considered covariates for some external response of interest. The appeal of STM's is their generality, adherence to the probabilistic framework, and accompanying R package which eliminates the practitioner's worrying about implementation details [RST19]. However, all this comes at a cost; one need only study the inference algorithms used to fit STM's to realize that the aforementioned problems persist. From Margaret et. al. [RSA16],

> As with other topic models, the exact posterior for the proposed model is intractable, and suffers from identifiability issues in theory (Airoldi et al. 2014a). Inference is further complicated in our setting by the nonconjugacy of the logistic Normal with the multinomial likelihood. We develop a partially collapsed variational expectation-maximization algorithm that uses a Laplace approximation to the nonconjugate portion of the model (Dempster, Laird, and Rubin 1977; Liu 1994; Meng and Van Dyk 1997; Blei and Lafferty 2007; Wang and Blei 2013).

Though this complexity is hidden from the analyst (by way of the R implementation), it is nonetheless daunting and involves multiple layers of approximation, all in order to maintain strict adherence to the probabilistic framework. The merits of this strategy are likely problem-specific, but we feel that the a priori rejection of simpler tools might constitute a case of "throwing the baby out with the bathwater." Given this and the nature of our goals, we elect to step outside the PTM framework, opting instead for methods that are fast, exact, and amenable to traditional nonparametric inference.

The key insight and contribution of this thesis is that new trends in NMF research play particularly nicely with simulation-based inference. A main advantage of this perspective is its simplicity, both in implementation and interpretation. Another is its speed, deriving from our reliance on several well-studied techniques from numerical linear algebra and optimization. While the resulting inferences are in principal more limited than those from a full-blown PTM, the previous sections call into question whether the added flexibility is worth the computational burden and excessive approximation. Our approach is part of a broader paradigm in which one reposes resampling-based nonparametrics on top of statistical learning techniques, an idea we think deserves more attention. An exemplary use and discussion of this perspective is offered by Taddy [TGCD16].

Chapter 3

Tractable NMF

3.1 Anchor Words and Their Role in Tractability

We have seen that NMF is, in general, NP-hard. However, high-quality solutions are often found in practice by employing greedy algorithms that converge to a local minimum. This is in keeping with recent trends in machine learning, where hard problems (in a complexity theoretic sense) are solved to great effect using algorithms that lack theoretical guarantees. Such developments have spawned a new sub-area of theoretical computer science called *beyond worst-case analysis* (BWCA) in which researchers look for special problem instances that *do* succumb to efficient computation, the idea being that some problems may be hard only on pathological instances which do not arise in practice [Rou19]. BWCA strikes a balance between between the extremes of average and worst-case analysis of algorithms, the former assuming a distribution over inputs and the latter allowing inputs that could be adversarial. In BWCA, no input distribution is assumed but structural properties of the problem are studied to identify where the hardness emanates from. An example is clustering, where there are results showing that "clustering is hard only when it doesn't matter" [Rou19]. The ultimate goal is a theory which distinguishes between natural and contrived problem instances, which might reveal why certain algorithms are phenomenally successful despite worst-case pessimism.

In the spirit of BWCA, Arora et. al. showed in a series of seminal works that NMF can be solved efficiently and practically under certain restrictions on the factorization and gave algorithms to do so [AGKM12, AGM12, AGH$^+$18]. Gillis provides a nice overview of the history of such approaches [Gil14] (and NMF more broadly), which span multiple research communities and over a decade of research. The key notion these results rely on is that of separability, first studied by Donoho and Stodden [DS04].

Definition 3.1.1. *A non-negative matrix $\mathbf{\Phi}$ is called separable if for all columns j there exists a row i such that $\phi_{ij} > 0$ but $\phi_{ik} = 0$ for all $k \neq j$.*

In the context of topic models, the columns of $\mathbf{\Phi}$ represent topics and the rows words, thus this condition states that for all topics there exists a word which only appears in that topic. Separability then essentially restricts what constitutes a legal topic; each topic must contain a word which is a perfect indicator of that topic's presence. For example, "Curie" might be such a word for a topic about chemistry. In the literature, words that have this property are often called *anchor words* or *anchors* for they "tie down" an associated topic. We too will adopt this terminology, but warn the reader that our notion of anchor words will be slightly different than existing definitions. These subtleties and their implications are discussed later in the chapter.

As in chapter 2, assume for the moment that an exact factorization $\mathbf{X} = \mathbf{\Phi\Theta}$ exists. To see why the existence of anchor words leads to tractable algorithms, first note that an appropriate re-ordering of the rows of $\mathbf{\Phi}$ gives a 2×1 block matrix whose upper block is diagonal:

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{\Gamma} \end{bmatrix}$$

where

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_T \end{bmatrix},$$

$\mathbf{\Gamma}$ is the remaining block, and T is the number of topics. This implies that $\mathbf{X}$ contains the rows of $\mathbf{\Theta}$ among its own rows, but scaled by a constant. This follows directly from the block form introduced above, since

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} \mathbf{\Lambda} \\ \mathbf{\Gamma} \end{bmatrix} \begin{bmatrix} \mathbf{\Theta} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{\Lambda\Theta} \\ \mathbf{\Gamma\Theta} \end{bmatrix} \end{aligned}$$

and left multiplication by a diagonal matrix has the effect of scaling rows. Thus we can say that $\mathbf{X}_\dagger = \mathbf{\Lambda\Theta}$ where $\mathbf{X}_\dagger$ signifies the block of the TDM corresponding to the anchor words. We can then rewrite the factorization in a form suggestive of the optimization problem to be posed shortly:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_\dagger \\ \mathbf{\Gamma\Lambda}^{-1}\mathbf{X}_\dagger \end{bmatrix}.$$

Herein lies the key to tractability, for one factor is predetermined up to a diagonal scaling. Assuming knowledge that the anchor words appear as the first T rows of $\mathbf{X}$, we can simply extract $\mathbf{X}_\dagger$ by inspection. As we'll see, the scaling factor $\mathbf{\Lambda}$ can be determined a posteriori by imposing an additional constraint on $\mathbf{\Phi}$, so that $\mathbf{\Theta} = \mathbf{\Lambda}^{-1}\mathbf{X}_\dagger$ can be recovered. This translates seamlessly to the optimization perspective in which we only observe $\mathbf{X}$ and seek a solution to (2.1). If $\mathbf{\Phi}$ is supposed separable, we can find the optimal factor $(\mathbf{\Gamma}\mathbf{\Lambda}^{-1})^\star$ by solving

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{X}_\ddagger - \mathbf{Y}\mathbf{X}_\dagger\|_{\mathrm{F}} \\ &\text{subject to} \quad \mathbf{Y} \succeq 0 \end{aligned} \tag{3.1}$$

where $\mathbf{X}_\ddagger = \mathbf{\Gamma}\mathbf{\Theta}$ is defined analogously to $\mathbf{X}_\dagger$ but for non-anchor rows and $\mathbf{Y}$ is the variable to be optimized over. Written this way, an interesting dual perspective emerges: (3.1) attempts to write the non-anchor words as linear combinations of the anchor words. More importantly, there is only one optimization variable in this problem; (2.1) has two. This is makes its solution dramatically easier. More concretely, (3.1) is nothing more than a series of non-negative least squares (NNLS) problems:

$$\begin{aligned} &\text{minimize} \quad \left\|\mathbf{x}_\ddagger^i - (\mathbf{X}_\dagger)^T \mathbf{y}^i\right\|_2 \\ &\text{subject to} \quad \mathbf{y}^i \succeq 0 \end{aligned}$$

for all rows i of $\mathbf{X}_\ddagger$. NNLS is well-studied and crucially, convex [CP]; in continuous optimization, convexity is synonymous with efficient algorithms that always reach global optima [BV04]. Combining each solution into a matrix $\mathbf{Y}^\star$, we conclude $\mathbf{Y}^\star = (\mathbf{\Gamma}\mathbf{\Lambda}^{-1})^\star$. The trick now is to determine the scale factor so that we can realize our original goal of deducing $\mathbf{\Phi}^\star$ and $\mathbf{\Theta}^\star$. This cannot be done without endowing $\mathbf{\Phi}$ with additional structure. A natural choice is that it have unit-norm columns. This has interpretational consequences as well – it essentially means that all topics are a priori weighted equally ($\mathbf{\Theta}$ tells us how each topic is weighted within a particular document, but topics should not have an inherent weight). To see how this helps, write

$$\left\|\boldsymbol{\phi}_j\right\|_2 = \sqrt{\lambda_j^2 + \sum_{i=1}^{V-1} \gamma_{ij}^2} = 1 \tag{3.2}$$

whence we find

$$\lambda_j = \sqrt{1 - \sum_{i=1}^{V-1} \gamma_{ij}^2} = \sqrt{1 - \lambda_j^2 \sum_{i=1}^{V-1} \left(\frac{\gamma_{ij}}{\lambda_j}\right)^2}.$$

Elementary algebra then gives

$$\lambda_j = \frac{1}{\sqrt{1 + \sum_{i=1}^{V-1} \left(\frac{\gamma_{ij}}{\lambda_j}\right)^2}}$$

and we notice the (i, j)th element of $\mathbf{\Gamma\Lambda}^{-1}$ in the denominator of the right-hand side. This quantity is solved for when computing (3.1), since $y_{ij}^\star = (\gamma_{ij}/\lambda_j)^\star$. Thus,

$$\lambda_j^\star = \frac{1}{\sqrt{1 + \left\|\mathbf{y}_j^\star\right\|_2^2}}.$$

This illustrates that under the unit-norm condition, (3.1) solves a reparameterized optimization problem whose desired parameterization can be found after the fact. The $\{\lambda_j\}$ have an interesting interpretation; first, observe that $0 < \lambda_j \leq 1$ for all j. If $\lambda_j = 1$, it must be that $\phi_{ij} = 0$ for all i. That is, topic j only includes the anchor word. Similarly, as $\lambda_j \to 0$ non-zero elements of $\boldsymbol{\phi}_j$ are necessarily introduced. We read this as saying that the $\{\lambda_j\}$ control the "diffuseness" of their respective topics. This is similar to some of the hyperparameters that must be set in PTM's like LDA, but with the advantage of being automatically "learned" from the data.

All told, we can use our findings to write $\mathbf{X} \approx \mathbf{\Phi}^\star\mathbf{\Theta}^\star$ purely in terms of the observed TDM and the solution to (3.1):

$$\mathbf{\Phi}^\star = \left[\begin{array}{cccc} \frac{1}{\sqrt{1+\left\|\mathbf{y}_1^\star\right\|_2^2}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{1+\left\|\mathbf{y}_2^\star\right\|_2^2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{1+\left\|\mathbf{y}_T^\star\right\|_2^2}} \\ \hline \\ \left| \right. & \left| \right. & & \left| \right. \\ \frac{\mathbf{y}_1^\star}{\sqrt{1+\left\|\mathbf{y}_1^\star\right\|_2^2}} & \frac{\mathbf{y}_2^\star}{\sqrt{1+\left\|\mathbf{y}_2^\star\right\|_2^2}} & \cdots & \frac{\mathbf{y}_T^\star}{\sqrt{1+\left\|\mathbf{y}_T^\star\right\|_2^2}} \\ \left| \right. & \left| \right. & & \left| \right. \end{array}\right]$$

and

$$\boldsymbol{\Theta}^{\star} = \begin{bmatrix} \sqrt{1 + \|\mathbf{y}_1^{\star}\|_2^2} & 0 & \dots & 0 \\ 0 & \sqrt{1 + \|\mathbf{y}_2^{\star}\|_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{1 + \|\mathbf{y}_T^{\star}\|_2^2} \end{bmatrix} \mathbf{X}_{\dagger}.$$

To summarize, anchor words ensure NMF can be solved efficiently and globally because they allow us to find certain pieces of the factorization by inspection, leading to a far simpler optimization problem. This is obviously advantageous from a computational perspective, but is essential to our inferential approach as well. Of course, all of the preceding discussion presumes the anchor words are known in advance. In practice we have to choose them, which is the subject of the next section.

3.2 Choosing Anchors

Recall that anchor words allow us to directly identify the optimal matrix of topic weights $\boldsymbol{\Theta}^{\star}$, gifting the remaining optimization problem the crucial property of convexity. The notion of anchor words can be perceived in two ways, and the distinction between them highlights the difference between our approach to topic models and that of the existing literature. In the first (exemplified by the previously cited work of Arora et. al.), we posit it that there exists a set of T true anchor words and the goal is to *find* them. This is in keeping with the PTM perspective in which an underlying generative mechanism is assumed. By contrast, a pure optimization perspective might consider anchor words to arise out of some pre-processing step in which we are free to *choose* any anchor words we like. This is eminently valid, since the factorization inherent to NMF is not considered part of any data-generating mechanism – it is simply a useful summary of the data. This idea extends to a weakly stochastic setting in which we assume the observed data to be generated by some unspecified distribution and the output to NMF is a statistic; the selection of anchor words is simply absorbed into the definition of this statistic. We adopt the latter approach given our rejection of PTM's in this work, the anchor words defined pragmatically as a set of representative words which are indicative of the corpus' overall structure. In this way, they can profitably be thought of as topic labels (in some sense nullifying our footnote from earlier).

In a PTM-based worldview, anchor words admit an interesting geometric perspective. It is not hard to show that anchor words are necessarily convex hull extreme points of the rows of an appropriately normalized word-word joint probability matrix [AGH$^+$18]. This suggests a geometric algorithm for estimating the anchors from data. Namely, empirical convex hull extreme points are identified and used in the subsequent factorization. While

this point of view and the ensuing theory relies on the assumed PTM, the anchor estimation algorithm of Arora et. al. is also useful in our context under slight modifications. In particular, we employ a simplification of their geometric algorithm but applied directly to the words – that is, the *unnormalized* rows of $\mathbf{X}$. In so doing, it firmly plants our approach to anchor words as one of *subset selection*.

Recall that words correspond to rows of $\mathbf{X}$, and imagine plotting these rows as vectors in $\mathbb{R}^D$. Then our first anchor selection procedure reads like:

1. The first anchor is defined as the point farthest from the origin; by the definition of the ℓ_2 norm, this is just the word that is used most frequently in the corpus[1].
2. The next anchor is the word farthest from the line spanned by the first.
3. The third is the word farthest from the plane spanned by the first two.
4. And so on.

This process repeats until T anchors are found. This is a well-known algorithm called the *Successive Projection Algorithm* (SPA) [Gil14, BGY$^+$01]. Luckily, this algorithm relies only on standard numerical linear algebra computations, and so can be carried out efficiently in numerous programming languages. In fact, it can even be computed exploiting existing implementations for OLS regression. To see this, note that at each step we must find the vector farthest from its projection onto the span of the anchor set thus far. Define $\mathbf{a}_k$ to be the row of $\mathbf{X}$ corresponding to the kth anchor word, and $\mathbf{A}_k$ as the matrix with $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\}$ as its columns. Then SPA finds

$$\begin{aligned}\mathbf{a}_{k+1} &= \arg\max_{\mathbf{v}} \|\mathbf{v} - \textsc{Proj}_{\mathbf{A}_k}\mathbf{v}\|_2 \\ &= \arg\max_{\mathbf{v}} \left\|\mathbf{v} - \mathbf{A}_k\left(\mathbf{A}_k^T\mathbf{A}_k\right)^{-1}\mathbf{A}_k^T\mathbf{v}\right\|_2\end{aligned}$$

where the maximum is taken over the all non-anchor words. Projection into the column space of $\mathbf{A}_k$ can be handled by functions like R's `lm.fit()`, and so is extremely easy to implement. Taking seriously SPA's original intent as a forward-mode variable selection method [BGY$^+$01], it can it be understood here as a trade-off between finding anchor words which are diverse yet relevant. The distance from a vector to a subspace is controlled by two factors: the vector's norm, and the angle it makes with that subspace. If we fix said angle, a longer vector will be farther from the subspace of interest. Conversely, distance increases with rotation through a positive angle when holding length fixed (reaching a maximum at

[1]Note that non-informative stopwords like "the" will be removed in pre-processing.

Algorithm 1: Choosing Anchor Words (Successive Projection Algorithm)

- Find the first anchor $\mathbf{a}_1$, the row of $\mathbf{X}$ with the largest norm.
- Initialize $k = 1$.
- While $k \neq T$ do:
  - Construct $\mathbf{A}_k$ with columns $\{\mathbf{a}_i\}_{i=1}^{k}$.
  - Find $\mathbf{a}_{k+1} = \arg\max_{\mathbf{v}} \left\| \mathbf{v} - \mathbf{A}_k \left(\mathbf{A}_k^T \mathbf{A}_k \right)^{-1} \mathbf{A}_k^T \mathbf{v} \right\|_2$ for $\mathbf{v}$ not among the anchors found thus far.
  - Iterate $k \leftarrow k + 1$.
- Output $\mathbf{A}$.

$\pi/2$ radians). This trade-off is formalized in algorithm (1) by virtue of the expression

$$\left\| \underbrace{\mathbf{v}}_{\text{length}} - \underbrace{\mathbf{A}_k \left(\mathbf{A}_k^T \mathbf{A}_k \right)^{-1} \mathbf{A}_k^T \mathbf{v}}_{\text{orthogonality}} \right\|_2,$$

which can be made large by increasing the norm of the first term or decreasing the norm of the second term; the latter is equivalent to increasing the angle between $\mathbf{v}$ and its projection onto the column space of $\mathbf{A}_k$. In conclusion, algorithm (1) picks anchors which maximize orthogonality to the span of the previous anchors, subject to appearing with some appreciable frequency. Orthogonality enforces diversity, while norm prevents obscure words from entering the set.

Algorithm (1) suggests a connection to the QR decomposition of numerical linear algebra. Making this connection explicit will lead to a simpler, faster algorithm with a well-understood interpretation.

3.3 The Gram-Schmidt Process and QR Decompositions

Suppose we wish to determine an orthonormal basis for the column space of some matrix $\mathbf{P}$. This can be accomplished sequentially using the Gram-Schmidt process, depicted below:

$$\begin{aligned}
\mathbf{q}_1 &= \frac{\mathbf{p}_1}{\|\mathbf{p}_1\|_2}, & \mathbf{q}_1 &\leftarrow \mathbf{q}_1 \\
\mathbf{q}_2 &= \mathbf{p}_2 - \langle \mathbf{q}_1, \mathbf{p}_2 \rangle \mathbf{q}_1, & \mathbf{q}_2 &\leftarrow \frac{\mathbf{q}_2}{\|\mathbf{q}_2\|_2} \\
\mathbf{q}_3 &= \mathbf{p}_3 - \left(\langle \mathbf{q}_2, \mathbf{p}_3 \rangle \mathbf{q}_2 + \langle \mathbf{q}_1, \mathbf{p}_3 \rangle \mathbf{q}_1 \right), & \mathbf{q}_3 &\leftarrow \frac{\mathbf{q}_3}{\|\mathbf{q}_3\|_2}
\end{aligned}$$

$\vdots$ $\vdots$

or in matrix form

$$\mathbf{q}_i = \mathbf{p}_i - \mathbf{Q}_{i-1}\mathbf{Q}_{i-1}^T\mathbf{p}_i$$

where $\mathbf{Q}_{i-1}$ has columns $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{i-1}\}$. At each step, $\mathbf{p}_i$ is projected into the column space of $\mathbf{Q}_{i-1}$ and $\mathbf{q}_i$ is taken to be the third side of the resulting right triangle, which is necessarily orthogonal to the column space of $\mathbf{Q}_{i-1}$. Finally, we scale $\mathbf{q}_i$ to have unit norm. This can be captured as a matrix product by the QR decomposition [TB97, GVL83]: $\mathbf{P} = \mathbf{QR}$ where $\mathbf{Q}$ contains the found orthonormal basis as its columns and $\mathbf{R}$ is an upper triangular matrix. More explicitly,

$$\mathbf{P} = \begin{bmatrix} | & | & & | \\ \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \langle \mathbf{q}_1, \mathbf{p}_1 \rangle & \langle \mathbf{q}_1, \mathbf{p}_2 \rangle & \cdots & \langle \mathbf{q}_1, \mathbf{p}_n \rangle \\ 0 & \langle \mathbf{q}_2, \mathbf{p}_2 \rangle & \cdots & \langle \mathbf{q}_2, \mathbf{p}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \mathbf{q}_n, \mathbf{p}_n \rangle \end{bmatrix}.$$

$\mathbf{P}$ can be of any shape. $\mathbf{Q}$ is of the same shape, except possibly with fewer columns in the case of rank deficiency. $\mathbf{R}$ is always square by our convections[2]. Algorithm (1) is intimately related to the QR decomposition, essentially performing a kind truncated version with the columns of $\mathbf{Q}$ constructed in very a particular order. Indeed for some anchor $\mathbf{a}_k$,

$$\mathbf{b}_k = \mathbf{a}_k - \textsc{Proj}_{\mathbf{A}_{k-1}}\mathbf{a}_k$$

is always orthogonal to the column space of $\mathbf{A}_{k-1}$. Furthermore, it is maximally far from the column space of $\mathbf{A}_{k-1}$ (by construction). Upon reflection, it appears as though algorithm (1) is reordering the rows of $\mathbf{X}$ (i.e. the words) so as to produce some sort of "optimal" QR decomposition on $\mathbf{X}^T$ (we transpose $\mathbf{X}$ in order to retain the column-wise picture described above). This is formalized in the following proposition.

Proposition 3.3.1. *Run algorithm 1 with $T = V$, and construct a permutation matrix $\mathbf{\Pi}$ whose columns have been permuted to agree with the order of the anchor words. Then the*

[2]Much like the SVD, one can give valid factorizations of different shapes for rank-deficient matrices. These often include rows of 0's that can optionally be removed.

following QR decomposition holds:

$$\mathbf{X}^T\mathbf{\Pi} = \begin{bmatrix} | & | & & | \\ \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_V \\ | & | & & | \end{bmatrix} \begin{bmatrix} \langle \mathbf{c}_1, \mathbf{a}_1 \rangle & \langle \mathbf{c}_1, \mathbf{a}_2 \rangle & \cdots & \langle \mathbf{c}_1, \mathbf{a}_V \rangle \\ 0 & \langle \mathbf{c}_2, \mathbf{a}_2 \rangle & \cdots & \langle \mathbf{c}_2, \mathbf{a}_V \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \mathbf{c}_V, \mathbf{a}_V \rangle \end{bmatrix}$$

where each $\mathbf{c}_k = \mathbf{b}_k / \|\mathbf{b}_k\|_2$, $\mathbf{b}_k = \mathbf{a}_k - \textsc{Proj}_{\mathbf{A}_{k-1}} \mathbf{a}_k$. *Furthermore, we have the following properties:*

$$|\langle \mathbf{c}_i, \mathbf{a}_i \rangle| \geq |\langle \mathbf{c}_j, \mathbf{a}_j \rangle| \text{ for all } i < j \quad \text{and} \quad |\langle \mathbf{c}_i, \mathbf{a}_j \rangle| \geq |\langle \mathbf{c}_k, \mathbf{a}_j \rangle| \text{ for all } i < k \leq j.$$

Before giving the proof, we need three facts. All are simple, but may be hard to parse within the argument below.

Claim 3.3.1. *The following hold:*

1. *Suppose that* $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathbb{R}^m$ *where* $\mathcal{A}, \mathcal{B}$ *are both linear susbspaces. Then for any* $\mathbf{v} \in \mathbb{R}^m$

$$\|\mathbf{v} - \textsc{Proj}_{\mathcal{A}} \mathbf{v}\|_2 \geq \|\mathbf{v} - \textsc{Proj}_{\mathcal{B}} \mathbf{v}\|_2.$$

2. *With* $\mathbf{v}$ *and* $\mathcal{A}$ *defined as in the previous item, we have*

$$\langle \mathbf{v}, \textsc{Proj}_{\mathcal{A}} \mathbf{v} \rangle = \|\textsc{Proj}_{\mathcal{A}} \mathbf{v}\|_2^2.$$

3. *Suppose* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^m$ *and* $\mathcal{A}$ *defined as in the previous items. Then*

$$\left\langle \mathbf{w}, \frac{\mathbf{v} - \textsc{Proj}_{\mathcal{A}} \mathbf{v}}{\|\mathbf{v} - \textsc{Proj}_{\mathcal{A}} \mathbf{v}\|_2} \right\rangle = \left\langle \mathbf{w}, \frac{\mathbf{w} - \textsc{Proj}_{\mathcal{A}} \mathbf{w}}{\|\mathbf{w} - \textsc{Proj}_{\mathcal{A}} \mathbf{w}\|_2} \right\rangle.$$

The first point above is a consequence of the fact that orthogonal projections necessarily produce the smallest residual vector of any such projection. Orthogonal projection onto the embedded subspace $\mathcal{B}$ gives a residual vector which is in general *not* orthogonal to $\mathcal{A}$, and thus longer by the properties of right triangles. The second point asserts that the inner product between a vector and its projection is simply the squared length of the projection itself – clearly true when considered geometrically. Finally, the last point is true because all residual vectors are necessarily collinear – projection onto one produces the same vector as projection onto another when normalized appropriately.

Proof of Proposition 3.3.1. Without loss of generality, we may assume that $\mathbf{X}^T$ is a tall rectangular matrix. If it is wide, it must be rank deficient. In this case, the trailing linearly dependent columns of the first factor can be arbitrary vectors in the column space of $\mathbf{X}^T$

and the corresponding block of the triangular factor is filled with zeros. The argument that follows then goes through as before. Now, it is a straightforward exercise in matrix algebra to check the correctness of the factorization, so it remains to show that the upper triangular factor obeys the required properties:

$$\begin{aligned}
|\langle \mathbf{c}_i, \mathbf{a}_i \rangle| &= \left| \left\langle \frac{\mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i}{\left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2}, \mathbf{a}_i \right\rangle \right| \\
&= \frac{\left| \langle \mathbf{a}_i, \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \rangle \right|}{\left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2} \\
&= \frac{\left| \langle \mathbf{a}_i, \mathbf{a}_i \rangle - \langle \mathbf{a}_i, \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \rangle \right|}{\left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2} \\
&= \frac{\left| \|\mathbf{a}_i\|_2^2 - \left\| \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2^2 \right|}{\left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2} && \text{(claim 3.3.1, point 2)} \\
&= \frac{\left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2^2}{\left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2} && \text{(Pythagoras)} \\
&= \left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2 \\
&\geq \left\| \mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_j \right\|_2 && \text{(def. of algorithm 1)} \\
&\geq \left\| \mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{j-1}} \mathbf{a}_j \right\|_2 && \text{(claim 3.3.1, point 1)} \\
&= |\langle \mathbf{c}_j, \mathbf{a}_j \rangle|
\end{aligned}$$

where the last line follows by the identity established earlier in the chain of deductions. The proof of the second property is similar:

$$\begin{aligned}
|\langle \mathbf{c}_i, \mathbf{a}_j \rangle| &= \left| \left\langle \frac{\mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i}{\left\| \mathbf{a}_i - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_i \right\|_2}, \mathbf{a}_j \right\rangle \right| \\
&= \left| \left\langle \frac{\mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_j}{\left\| \mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_j \right\|_2}, \mathbf{a}_j \right\rangle \right| && \text{(claim 3.3.1, point 3)} \\
&= \frac{\left| \langle \mathbf{a}_j, \mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_j \rangle \right|}{\left\| \mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_j \right\|_2} \\
&= \left\| \mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{i-1}} \mathbf{a}_j \right\|_2 && \text{(similar to last argument)} \\
&\geq \left\| \mathbf{a}_j - \textsc{Proj}_{\mathbf{A}_{k-1}} \mathbf{a}_j \right\|_2 && \text{(claim 3.3.1, point 1)} \\
&= |\langle \mathbf{c}_k, \mathbf{a}_j \rangle| \,.
\end{aligned}$$

□

This proposition reveals two beautiful properties. First, the residuals (i.e., the distances between each anchor and the subspace of previous anchors) decrease monotonically as the algorithm runs. Relatedly, projecting an anchor onto each of the subspaces from previous iterations gives a monotonically decreasing sequence of residuals (in the iterations of the

Algorithm 2: Successive Projection Algorithm by QR with Column Pivots

- Compute QR decomposition of $\mathbf{X}^T$ (with column pivoting).
- $\sigma :=$ the permutation of the document indices corresponding to the pivoting strategy.
- For $k = 1, 2, \ldots, T$ do:
  - $\mathbf{a}_k = \mathbf{x}^{\sigma(k)}$, the row of $\mathbf{X}$ corresponding the kth pivot.
- Output $\mathbf{A}$.

algorithm). Together, these say that anchor selection picks points that "fill up" the ambient vector space in a sequentially optimal way. While we arrived at this result and its proof independently, it is no surprise that this "discovery" is well-known. In fact, what we have just established is a link between algorithm (1) and the QR decomposition *with column pivoting* [Eng97, BG65, Gil14]. Anchor words are revealed by the pivoting strategy, which is just the particular permutation applied to the columns of $\mathbf{X}^T$. From this new perspective, the decomposition picks a set of "very linearly independent" columns [BBP$^+$10]. Much work has been done on extending this basic algorithm and giving bounds on its success in terms of the singular values of $\mathbf{X}^T$ [GE96, Cha87, BBP$^+$10], but the vanilla version is known to work well in practice [Gil14]. Ultimately, the QR connection makes explicit the idea that anchors capture most of the information in the TDM. It also affords a computational advantage in that fast QR implementations (like that of `LAPACK` [ABB$^+$99]) can be used in anchor selection. This change in viewpoint is reflected by algorithm (2).

Finally, we provide a word of caution about our usage of the term "anchor" in subsequent chapters. Sometimes it is more convenient to think in terms of a set of *indices* giving the rows of $\mathbf{X}$ at which anchor words occur. Other times (in particular when carrying out computations, as in this chapter) it is profitable to think of the *rows themselves*. We will move freely between these perspectives, but attempt to provide clarification in cases of possible confusion.

3.4 Speed-ups by Random Projections

We conclude this chapter by noting that both anchor selection and the solving of NMF itself are amenable to exciting results in the flourishing field of *randomized numerical linear algebra* (RandNLA). RandNLA is an interdisciplinary field populated by theoretical computer scientists, statisticians, numerical analysts, and others, whose goal is to find faster algorithms for massive matrix computations by employing a certain degree of randomness [DM16]. The hallmark result in this arena is the Johnson-Lindenstrauss (JL) lemma

[JL84, DG03], which states that a collection of points in high-dimensional Euclidean space can be projected onto a (suitably chosen) random subspace such that all pairwise distances are nearly preserved (with high probability). The intent is to construct a subspace whose dimension is small relative to the initial space, thereby speeding up many algorithms; an original application was high-dimensional nearest-neighbour search [IM98].

Let us first study why this should work in the context of anchor selection. Let $\mathbf{H} \in \mathbb{R}^{m \times n}$ be the projection matrix, whose entries are random according to some user-specified distribution. Here we are thinking of $m \ll n$. The JL lemma states that for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, we have $\|\mathbf{Hv} - \mathbf{Hw}\|_2 \approx \|\mathbf{v} - \mathbf{w}\|_2$. Taking $\mathbf{w} = \mathbf{0}$, we see that this implies norm preservation. In this way, $\mathbf{H}$ is acting like an orthonormal matrix, "rotating" points down to a lower dimension with minimal distortion. Continuing this analogy, consider the central computation in step $k+1$ of algorithm (1) after applying $\mathbf{H}$ to the rows of $\mathbf{X}$:

$$\left\| \mathbf{Hv} - \mathbf{HA}_k \left((\mathbf{HA}_k)^T \mathbf{HA}_k \right)^{-1} (\mathbf{HA}_k)^T \mathbf{Hv} \right\|_2 .$$

If we are willing to assert $\mathbf{H}^T\mathbf{H} \approx \mathbf{I}$ (as would be true exactly for genuine orthonormal matrices), this is approximately equal to

$$\left\| \mathbf{Hv} - \mathbf{HA}_k \left(\mathbf{A}_k^T \mathbf{A}_k \right)^{-1} \mathbf{A}_k^T \mathbf{v} \right\|_2$$

which is in turn approximately equal to the original distance calculation (by the JL lemma). Remarkably, the construction of $\mathbf{H}$ is very simple – an appropriately scaled matrix filled with i.i.d. standard Gaussian variates suffices. For our purposes, this lemma means that anchor selection can occur in a space of much lower dimension, giving potentially large speed-ups. Worst-case bounds provide guidance as to the choice of subspace dimension (typically on the order of $\log m/\epsilon^2$ for ϵ the multiplicative distortion), but our experience suggests that dimension can be reduced drastically (by an order of magnitude or more) while approximately preserving the selected anchor words.

NMF too can benefit from the random projection treatment, though it is not so straightforward as for purely distance-based calculations. Using the same notation as above, Boutsidis and Drineas proved that the following scheme achieves the desired result [BD08]:

1. Construct $\mathbf{H}_n$, an $n \times n$ *Haddamard-Walsh* matrix [HW78]. Likely a new encounter for statisticians, this family of matrices is defined recursively for powers of 2 as

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{H}_{2^n} = \begin{bmatrix} \mathbf{H}_{2^{n-1}} & \mathbf{H}_{2^{n-1}} \\ \mathbf{H}_{2^{n-1}} & -\mathbf{H}_{2^{n-1}} \end{bmatrix}.$$

Because it is only defined for powers of 2, zero padding is employed to make the data matrix and response of coincident dimension. The motivation for using such a matrix is that it has properties reminiscent of the finite Fourier transform and can be applied to vectors (and by extension matrices) very quickly, much faster than by naive matrix multiplication.

2. Sample the rows of $\mathbf{H}_n$ with probability m/n and scale the selected rows by a factor of $\sqrt{n/m}$; call the resulting matrix $\widetilde{\mathbf{H}}$.

3. Construct $\mathbf{D}$, an $n \times n$ diagonal matrix whose entries are selected uniformly from $\{\pm 1\}$.

Then left multiplication of both the data matrix and response by $\widetilde{\mathbf{H}}\mathbf{D}$ minimally distorts the solution to a non-negative least-squares problem (which now takes place in $m \ll n$ dimensions) with high probability. Thus, plugging this method into our usual NMF pipeline is an attractive idea, particularly for massive TDM's. Empirically, we have observed that pre-processing NMF in this way permits dramatic dimensionality reduction and a resulting drop in computational time – topic models with thousands of documents can be fit in a matter of minutes.

Chapter 4

Inference: Permutation Tests, Matrix Balancing, and Beyond

4.1 Grouped Topic Models

Recall from chapter 2 that the term document matrix $\mathbf{X}$ contains as its entries the number of times word i appears in document j. This matrix has thus far been treated as a deterministic object, and we gave an algorithm for approximately factoring it into two non-negative matrices $\mathbf{\Phi}^\star$ and $\mathbf{\Theta}^\star$. In this sense our development has kept within the realm of descriptive statistics, in which we carry out procedures that suggest the data's underlying structure without any notion of a sample or a population from which the data arose. In this section we move beyond the deterministic with what might be termed a "weakly stochastic model" in order to facilitate the inferential procedures proposed in this chapter.

An extremely simple probabilistic model (too simple to constitute a PTM in our opinion) for $\mathbf{X}$ would be

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_D \overset{\text{i.i.d}}{\sim} F$$

for some distribution F. That is, we consider each document (column of $\mathbf{X}$) to be drawn i.i.d from an unspecified distribution. Such a model implicitly underlies the validity of most resampling methods, and encompasses popular PTM's like LDA (these model word and topic associations with complex dependencies but the marginal distribution over documents can indeed be framed as above). Unlike in most PTM's, document length is here considered stochastic which validates the "identical distribution" portion of the i.i.d. assumption. Additionally, it may be that the documents exhibit some sort of partitioning. For example, articles may be classified according to the news outlet in which they appeared, or by their authors. As mentioned in the introduction, our canonical scenario is one in which documents can be divided regionally. Formally, the columns of $\mathbf{X}$ are partitioned so as to come from a

set of distributions:

$$\{\mathbf{x}_i\}_{i\in\mathcal{R}_j} \overset{\text{i.i.d.}}{\sim} F_j$$

where $\{\mathcal{R}_j\}_{j=1}^R$ forms a partition on the documents. In other words, the documents within a specific region are drawn i.i.d from their own, regional distribution (of which there are R). We call this setup a *grouped topic model*, our goal being to make inference about regional variation in the corpus. Geography-informed topic models are not new [EOSX10, YCH+11], but to the best of our knowledge this weakly stochastic perspective is; recall our underlying philosophy from chapter 1. The approaches referenced above rely on full-scale PTM's fit using variational inference, or solve a different problem entirely (such as fitting a different set of topics within each region).

We would like to know which regions differ, if any, and how these differences are reflected in the shared topic structure of the corpus. In fact, all inferences will ultimately be based on topic structure and so it is advantageous to work with a restricted model in which

$$\{\boldsymbol{\theta}_i^\star\}_{i\in\mathcal{R}_j} \overset{\text{i.i.d.}}{\sim} G_j.$$

That is, we rely on the same stochastic assumptions but applied directly to the vectors of topic weights. In words, this says that the topic makeup of each document is i.i.d. within each region, capturing our intuition that the allocation of topics in the corpus will depend on geography. Dealing directly with $\boldsymbol{\Theta}^\star$ in this way is well motivated. First, it directly addresses our primary interest in the topics. The original documents are high-dimensional, sparse, and noisy – we care much more about their essential *content* as expressed in the columns of $\boldsymbol{\Theta}^\star$. What's more, inferences in this model implicitly condition on the uncovered set of anchor words. Were this not the case, inferences with respect to particular topics would be unjustified. This is undesirable, for the analyst likely cares about the particular set of topics in front of them as indicated by the anchor words, and not a set of unlabelled, abstract topics. Finally, the compressed nature of a document's topic representation leads to fast and simple computations.

We address two central inferential questions. First, we consider the hypothesis test:

$$\mathcal{H}_0: \quad G_1 = G_2 = \ldots = G_R \tag{4.1}$$

$$\mathcal{H}_1: \quad G_i \neq G_j \text{ for some } i, j \tag{4.2}$$

where we can restrict the test to two regions for direct region-region comparisons

$$\mathcal{H}_0: \quad G_j = G_k \quad \text{and} \quad \mathcal{H}_1: \quad G_j \neq G_k \tag{4.3}$$

if desired. Second, we seek to estimate a measure of association between topics and regions. This complements the above hypothesis test by telling us how regions and topics interact in the presence of a rejected null. Both of these goals are realized using a common statistic

$$\widehat{\boldsymbol{\Theta}}^{\star}(\boldsymbol{\theta}_1^{\star}, \boldsymbol{\theta}_2^{\star}, \ldots, \boldsymbol{\theta}_D^{\star})$$

which reflects the within-region allocation of topic weight to documents. Specifically, we take $\widehat{\boldsymbol{\Theta}}^{\star}$ to be the $T \times R$ matrix found by summing $\boldsymbol{\Theta}^{\star}$ row-wise within each region so that its (i, j)th element is given by $\sum_{k \in \mathcal{R}_j} \theta_{ik}^{\star}$. We may sometimes call this the topic-region matrix, and refer to its columns as $\left\{\boldsymbol{\theta}_{\mathcal{R}_j}^{\star}\right\}$. Basing tests and estimation procedures on $\widehat{\boldsymbol{\Theta}}^{\star}$ is a direct reflection of our interest in the relationship between topics and regions, but is further motivated by its facilitation of feasible computations. This is in large part due to its compact form: generally $R \ll D$, and this will be true to several orders of magnitude in the applied example of chapter 5.

Testing and estimation theory based on $\widehat{\boldsymbol{\Theta}}^{\star}$ is developed in the subsequent sections, relying respectively on permutation tests and the bootstrap. We remind the reader that this perspective and the resulting inferences are only possible because of the recent advances in NMF outlined in this thesis. Because solving the NMF problem as described in chapter 3 gives a unique matrix $\boldsymbol{\Theta}^{\star}$ for any input matrix $\mathbf{X}$, $\widehat{\boldsymbol{\Theta}}^{\star}$ can be considered a statistic. This would not be true without anchor words and separability; the same $\mathbf{X}$ could map to different $\boldsymbol{\Theta}^{\star}$'s due to the randomization needed to handle nonconvex loss surfaces.

4.2 Introduction to Permutation Tests

Permutation tests are an ingenious suite of methods for testing hypotheses under a bare minimum of assumptions. For this reason, they are especially attractive for use with modern, high-dimensional data for which familiar parametric assumptions may be misguided. Much like the bootstrap, permutation tests replace tedious calculations under brittle model assumptions with randomized computation; indeed, they are often lumped together with the jackknife, bootstrap, cross validation, etc. as constituting *statistical resampling methods*. Another advantage to permutation tests (and resampling techniques at large) is that they can often repose on top of statistical learning methods which are highly algorithmic in nature and difficult to frame in the traditional language of inferential statistics. As alluded to in the introduction, this is our primary motivation for using them. Our treatment thus far has centred around optimization and numerical linear algebra; permutation tests now allow us to bring inferential thinking to bear without compromising the tools developed in previous chapters.

Suppose we would like to test whether a treatment group differs from a control group.

Plenty of parametric approaches exist to solve this problem, as do many nonparametric tests based on ranks (not incidentally, ranks also play an important role in the theory justifying permutation tests). A particularly elegant idea is the following [MB01, Goo94]: under the null hypothesis that the groups do not differ, the labels indicating which group an individual belongs to are meaningless. Thus, for some test statistic T we can evaluate T under all permitted permutations of the labels and obtain a p-value as the fraction of times T exceeds its original, observed value. Amazingly, this test is valid under very broad circumstances. T can be any sensible function of the data (so long as it is not symmetric in its arguments), the groups need not be drawn from specific parametric families, and the test is exact in the sense that type I error is controlled as advertised. While the simple two-sample scenario aptly conveys the essential reasoning, extensions to more general settings are straightforward. As such, permutation tests provide robust and distribution-free analogues to classical t-tests, F-tests, and the like.

While intuitively appealing, a more formal exposition is required to put permutation tests on solid mathematical grounds. In particular, the fact that type I error is controlled is non-trivial. We will illustrate the theory for the two-sample problem described above, as it keeps the level of obfuscatory mathematics to a minimum; the additional machinery required for more general scenarios does not aid in understanding the main ideas. Consider two random samples

$$X_1, X_2, \ldots, X_{n_1} \overset{\text{i.i.d.}}{\sim} F \quad \text{and} \quad Y_1, Y_2, \ldots, Y_{n_2} \overset{\text{i.i.d}}{\sim} G$$

from distributions F, G and we would like to test

$$\begin{aligned} \mathcal{H}_0 &: \quad F = G \\ \mathcal{H}_1 &: \quad F \neq G. \end{aligned}$$

F and G may admit densities, but this need not be the case. Under the null, our two samples are equivalent to one large random sample from a common distribution $H = F = G$. That is,

$$\underbrace{Z_1, Z_2, \ldots, Z_{n_1}}_{X_1, X_2, \ldots, X_{n_1}}, \underbrace{Z_{n_1+1}, Z_{n+2}, \ldots, Z_{n_1+n_2}}_{Y_1, Y_2, \ldots, Y_{n_2}} \overset{\text{i.i.d}}{\sim} H.$$

Now, imagine a test statistic T meant to discern between $\mathcal{H}_0$ and $\mathcal{H}_1$. A particularly simple example relevant in the two-sample testing problem is the absolute difference in means, or more appropriately, the function which computes the difference of means between its first n_1 and last n_2 arguments. When evaluated on the original data, this gives $|\overline{X} - \overline{Y}|$ under the alternative but $|\overline{Z}_1 - \overline{Z}_2|$ under the null (the difference in means between two samples from

the same population). Recomputing T using all $N = (n_1+n_2)!/n_1!n_2!$ allocations of n_1+n_2 items to groups of sizes $n_1!$ and $n_2!$ provides a means of testing $\mathcal{H}_0$, for the behaviour of this schema varies markedly depending on whether or not the null holds. Since permuting the data effectively shuffles the groupings by which we calculate means, these T's should not vary much from their original, unpermuted value $T^{\text{obs.}}$ if it is true that the samples come from the same distribution. Conversely, $\mathcal{H}_1$ implies samples from F and G will be mixed among T's arguments, thus producing a sequence that looks "unusual" in the context of the initial $T^{\text{obs.}}$. This motivates the following two-sided testing procedure:

1. Observe $T^{\text{obs.}}$, the value of the test statistic on the original data.
2. Construct $T_1, T_2, \ldots, T_N$, the test statistic evaluated on all permitted permutations of the data.
3. Let p be the fraction of $\{T_i\}$ which exceed $T^{\text{obs.}}$ in absolute value.
4. Reject $\mathcal{H}_0$ if $p < \alpha$ for some pre-specified level of significance α.

While intuitively clear, showing that this strategy controls type I error requires a definition and some analysis. We introduce the concept of exchangeability and provide evidence for the validity of permutation tests by showing error control in a restricted setting.

Definition 4.2.1. *A random vector $(Z_1, Z_2, \ldots, Z_n)$ is said the be exchangeable if for all permutations π of $\{1, 2, \ldots, n\}$, we have $(Z_1, Z_2, \ldots, Z_n) \stackrel{d}{=} \left(Z_{\pi(1)}, Z_{\pi(2)}, \ldots, Z_{\pi(n)}\right)$.*

This definition says that a sequence of random variables is exchangeable if and only if permuting its indices does not change the joint distribution. This condition often holds in situations statisticians care about, but certainly not always. A Markov process, for example, poses a clear violation. While exchangeability is a deep concept worthy of study in its own right, for the current work it is sufficient to notice that i.i.d sequences are exchangeable – a simple consequence of the fact that their joint densities factor into a product of marginals. Exchangeability is a weaker concept than i.i.d. however, and the proof given below goes through under exchangeability alone – note that nowhere do we rely exclusively on the i.i.d. nature of the data.

Proposition 4.2.1. *The simple permutation test outlined above achieves type I error control.*

Proof Sketch. This argument is based on the comments in [Was]; it is surprisingly difficult to find concise statements of this kind about permutation tests. Much of the literature is either intuitive in nature or highly mathematical. Because notation can become cumbersome rather quickly, the first part of the argument restricts its attention to the case $n_1 = n_2 = 1$. That is, the random sample under the null is just the pair (Z_1, Z_2). This simplifies notation

while maintaining the argument's essentials. Furthermore, we assume a one-sided test. First, put

$$T_1 = T(Z_1, Z_2) \quad \text{and} \quad T_2 = T(Z_2, Z_1).$$

We now invoke the properties of rank statistics. Consider the probability that T_1, T_2 exhibit a particular ordering:

$$\Pr_{\mathcal{H}_0}(T_1 < T_2).$$

This event depends solely on the random sample (Z_1, Z_2) because it can be mapped uniquely to an ordering on T_1, T_2. In other words,

$$\Pr_{\mathcal{H}_0}(T_1 < T_2) = \Pr_{\mathcal{H}_0}\{(Z_1, Z_2) \in \mathcal{Z}\}$$

where

$$\mathcal{Z} = \{(z_1, z_2) \text{ such that } T(z_1, z_2) < T(z_2, z_1)\}$$

is the subset of the sample space inducing the desired event. Consider next the alternative ordering in which $T_2 < T_1$. The probability of this event occurring is

$$\Pr_{\mathcal{H}_0}(T_2 < T_1) = \Pr_{\mathcal{H}_0}(T(Z_2, Z_1) < T(Z_1, Z_2)).$$

In words, this event occurs when the *permuted sample* (Z_2, Z_1) lies in $\mathcal{Z}$. Thus,

$$\begin{aligned}\Pr_{\mathcal{H}_0}(T_2 < T_1) &= \Pr_{\mathcal{H}_0}\{(Z_2, Z_1) \in \mathcal{Z}\} \\ &= \Pr_{\mathcal{H}_0}\{(Z_1, Z_2) \in \mathcal{Z}\}\end{aligned}$$

where the last line follows from exhangeability. In summary, the two possible orderings are equiprobable and so must occur with probability $1/2$. The exact same arguments apply in the general case and we conclude that orderings of the statistics $\{T_1, T_2, \ldots, T_N\}$ are uniformly distributed on their $N!$ possible arrangements. An equivalent statement is that the set of rank statistics $\{\Psi_i\}$ is uniformly distributed on permutations of the integers $\{1, 2, \ldots, N\}$, and an easy corollary is that a particular rank Ψ_i is uniformly distributed on the set $\{1, 2, \ldots, N\}$ itself. Now, fix some $\alpha \in (0, 1)$ and consider the statistic

$$\frac{\#\{T_i \text{ such that } T_i \geq T^{\text{obs.}}\}}{N}.$$

Ranking all the T's (with 1 taken to mean largest) so that Ψ_i denotes the rank of T_i, this can be written as

$$\frac{\Psi^{\text{obs.}}}{N}$$

where $\Psi^{\text{obs.}}$ is the rank of $T^{\text{obs.}}$ among the T_i's. Then

$$\begin{aligned}\Pr_{\mathcal{H}_0}\left(\frac{\Psi^{\text{obs.}}}{N} < \alpha\right) = \Pr_{\mathcal{H}_0}\left(\Psi^{\text{obs.}} \leq \lfloor N\alpha \rfloor\right) &= \sum_{j=1}^{\lfloor N\alpha \rfloor} \Pr_{\mathcal{H}_0}(\Psi^{\text{obs.}} = j) \\ &= \frac{\lfloor N\alpha \rfloor}{N} \\ &< \alpha\end{aligned}$$

as desired. Type I error control is established. □

4.3 Choosing a Test Statistic

Equipped with the theory of permutation tests, we now propose suitable test statistics for (4.2) and (4.3). Recall that we work with the topic-region matrix $\widehat{\boldsymbol{\Theta}}^\star$ which will lead to fast, simple tests. First, observe that under the null we have

$$\mathop{\mathrm{E}}_{\mathcal{H}_0} \boldsymbol{\theta}_i^\star = \boldsymbol{\eta}$$

for all i and some unknown $\boldsymbol{\eta}$. Assuming that each region contains $d_1, d_2, \ldots, d_R$ documents respectively, linearity of expectation means

$$\mathop{\mathrm{E}}_{\mathcal{H}_0} \boldsymbol{\theta}_{\mathcal{R}_j}^\star = d_j \boldsymbol{\eta}.$$

This implies that in expectation, all columns of $\widehat{\boldsymbol{\Theta}}^\star$ will be collinear. Deviation from collinearity provides evidence against the null. A natural "distance" from which to build a test statistic is then the cosine distance between two columns of $\widehat{\boldsymbol{\Theta}}^\star$:

$$\Delta(\boldsymbol{\theta}_{\mathcal{R}_j}^\star, \boldsymbol{\theta}_{\mathcal{R}_k}^\star) = \frac{\langle \boldsymbol{\theta}_{\mathcal{R}_j}^\star, \boldsymbol{\theta}_{\mathcal{R}_k}^\star \rangle}{\left\|\boldsymbol{\theta}_{\mathcal{R}_j}^\star\right\|_2 \left\|\boldsymbol{\theta}_{\mathcal{R}_k}^\star\right\|_2}.$$

The cosine distance is a well-established measure of similarity in the text analysis and information retrieval communities [JM09], and can be understood statistically as an uncentred version of Pearson's correlation coefficient. Geometrically, it gives the cosine of the angle between two vectors of topic weights and thus is perfectly suited to detect departure from collinearity. Note that it is invariant to the length of either of its arguments, which is important since we know the columns of $\widehat{\boldsymbol{\Theta}}^\star$ could have very different lengths (emanating from

the different within-region sample sizes). This leads easily to two test statistics, depending on whether we are testing (4.2) in full or a pairwise comparison between regions as given in (4.3):

$$T = \frac{1}{R}\sum_{j,k} \Delta(\boldsymbol{\theta}^\star_{\mathcal{R}_j}, \boldsymbol{\theta}^\star_{\mathcal{R}_k}) \qquad \text{or} \qquad T = \Delta(\boldsymbol{\theta}^\star_{\mathcal{R}_j}, \boldsymbol{\theta}^\star_{\mathcal{R}_k}).$$

These are not symmetric in their arguments when considered functions of $\{\boldsymbol{\theta}^\star_1, \boldsymbol{\theta}^\star_2, \ldots, \boldsymbol{\theta}^\star_D\}$ – as is required of the permutation technology – since they depend on the aggregation of topic weight within each region. A permutation test is thus enacted by shuffling the region labels in all possible ways, recording the value of T at each step. The simplicity of T means it can be computed rapidly; this is in contrast with many popular ANOVA-like statistics (see [MB01]) which, while potentially more powerful, are expensive to evaluate.

This gives rise to algorithms (3) and (4) – the somewhat abstruse notation $\boldsymbol{\theta}^\star_{\pi_s(\mathcal{R}_j)}$ signifies the vector of topic weights for region j under a permutation of the region labels. These tests are left-tailed only, reflecting the fact that collinearity is further violated as the cosine distance decreases to 0. Note that we must concede to approximating the p-value using a Monte Carlo estimate, as in most problems the number of possible permutations is vast. However, this is a small price to pay and is similar to the approximation inherent in bootstrapping, say. In principal, the p-value is exact given that we have access to the requisite computational resources. We proceed by taking a large number s of all possible permutations on $\{1, 2, \ldots, D\}$, which we hope is large enough to approximate the p-value arbitrarily well. Since the details of this randomized algorithm are known exactly, we can quantify its error using standard statistical techniques (confidence intervals are most popular) but pragmatism dictates that a large number of permutations will suffice. In practice, this intuition is born out. The Monte Carlo approach has an advantage as well in that it naturally handles the fact that not all $D!$ permutations are allowable, since $d_1! \cdots d_R!$ of those permutations do not change the overall assignment of documents to regions. The probability of seeing such a permutation is $s/d_1!d_2! \cdots d_R!$ – negligibly small in most cases – and so the randomization saves us from having to explicitly enumerate those permutations which are allowed.

4.4 Matrix Balancing and Sinkhorn-Knopp

We now turn to estimation. The preceding sections have taught us how to detect a statistically significant difference in the topic structure of each region – that is, a departure from the null hypothesis that the distribution "generating" topic allocation is the same in each region. Of course, this is only a partial answer to the questions an analyst will likely be interested in. A complete analysis would surely investigate just how the topics differ

Algorithm 3: Permutation test, overall.

- Use algorithm (1) or (2) to find the anchor words (rows of $\mathbf{X}$).
- Solve (3.1) once to find the optimal $\mathbf{\Lambda}^\star$.
- Construct $\mathbf{\Theta}^\star$, $\widehat{\mathbf{\Theta}}^\star$, and the observed test statistic $T^{\text{obs.}} = \frac{1}{R}\sum_{j,k}\Delta(\boldsymbol{\theta}^\star_{\mathcal{R}_j}, \boldsymbol{\theta}^\star_{\mathcal{R}_k})$.
- For $i = 1, 2, \ldots, s$ do:
  - Construct a random permutation π_s on $\{1, 2, \ldots, D\}$.
  - $T_{\pi_s} = \frac{1}{R}\sum_{j,k}\Delta(\boldsymbol{\theta}^\star_{\pi_s(\mathcal{R}_j)}, \boldsymbol{\theta}^\star_{\pi_s(\mathcal{R}_k)})$.
- $p :=$ fraction of $T_{\pi_s} \leq T^{\text{obs.}}$
- Reject $\mathcal{H}_0$ if $p < \alpha$ for some specified α.

Algorithm 4: Permutation test, region-region comparison.

- Use algorithm (1) or (2) to find the anchor words (rows of $\mathbf{X}$).
- Solve (3.1) once to find the optimal $\mathbf{\Lambda}^\star$.
- Construct $\mathbf{\Theta}^\star$, $\widehat{\mathbf{\Theta}}^\star$, and the observed test statistic $T^{\text{obs.}} = \Delta(\boldsymbol{\theta}^\star_{\mathcal{R}_j}, \boldsymbol{\theta}^\star_{\mathcal{R}_k})$ for a pair of desired regions $\mathcal{R}_j, \mathcal{R}_k$.
- For $i = 1, 2, \ldots, s$ do:
  - Construct a random permutation π_s on $\{1, 2, \ldots, D\}$.
  - $T_{\pi_s} = \Delta(\boldsymbol{\theta}^\star_{\pi_s(\mathcal{R}_j)}, \boldsymbol{\theta}^\star_{\pi_s(\mathcal{R}_k)})$.
- $p :=$ fraction of $T_{\pi_s} \leq T^{\text{obs.}}$
- Reject $\mathcal{H}_0$ if $p < \alpha$ for some specified α.

between regions, and thus the likely cause of a rejected null. We are also led naturally to the reverse construction – what we might deem the region structure of each topic. These dual perspectives answer related, yet distinct questions, and together complete the picture initiated by permutation tests. Concisely then, this section explores the following two questions: "which topics are important to each region?", and "which regions are important for a particular topic?". Ultimately, we demonstrate the shortcomings of considering each of these questions on their own terms. The solution will instead address both simultaneously via a single construction with clear geometric and problem-specific interpretations.

Naively, we might expect to answer our questions of interest by simply inspecting $\widehat{\boldsymbol{\Theta}}^\star$. Recall that each column of this matrix has the following structure:

$$\boldsymbol{\theta}^\star_{\mathcal{R}_j} = \begin{bmatrix} \sum_{i\in\mathcal{R}_j}\theta_{1i} \\ \sum_{i\in\mathcal{R}_j}\theta_{2i} \\ \vdots \\ \sum_{i\in\mathcal{R}_j}\theta_{Ti} \end{bmatrix} \begin{matrix} \text{total wt. for topic 1 in region } j \\ \text{total wt. for topic 2 in region } j \\ \vdots \\ \text{total wt. for topic } T \text{ in region } j. \end{matrix}$$

Thus, we could assert that large coordinates in the above vector indicate that the corresponding topics are important to region j. The problem with this idea is that it fails to account for the fact that topics are distributed unevenly throughout the corpus. Certain topics will appear often, and consistently across all regions. These will be given high weight in the above vectors due to their ubiquity, but are uninformative for distinguishing one region from another. To alleviate this, we need a way of downweighting topics that are given high weight across the entire corpus. This is accomplished by simply normalizing the rows of $\widehat{\boldsymbol{\Theta}}^\star$ via the transformation

$$\widehat{\boldsymbol{\Theta}}^\star \mapsto \begin{bmatrix} 1/\left\|\boldsymbol{\theta}^\star_{\mathcal{T}_1}\right\|_2 & 0 & \cdots & 0 \\ 0 & 1/\left\|\boldsymbol{\theta}^\star_{\mathcal{T}_2}\right\|_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\left\|\boldsymbol{\theta}^\star_{\mathcal{T}_T}\right\|_2 \end{bmatrix} \begin{bmatrix} — & \boldsymbol{\theta}^\star_{\mathcal{T}_1} & — \\ — & \boldsymbol{\theta}^\star_{\mathcal{T}_2} & — \\ & \vdots & \\ — & \boldsymbol{\theta}^\star_{\mathcal{T}_T} & — \end{bmatrix}. \tag{4.4}$$

where $\left\{\boldsymbol{\theta}^\star_{\mathcal{T}_i}\right\}$ give the rows of $\widehat{\boldsymbol{\Theta}}^\star$. Normalizing like this effectively ensures that all topics are on equal footing by forcing them to have total weight 1 across the corpus. This means that post-transformation, the columns of $\widehat{\boldsymbol{\Theta}}^\star$ indeed give a meaningful indication of which topics are of particular importance to a given region. Translating to more statistical language, there is large variance among the usage of each topic, and so this process can be seen as a form of topic-wise standardization.

As alluded to in the first paragraph of this section, we are also interested in investigat-

ing which *regions* are of import to a particular *topic*. This inverts our perspective, and not incidentally we must exploit a certain duality to make sense of it. Recall that the TDM is a matrix of counts, with columns representing documents and rows representing words. In this way, a document is a just a vector of counts across the entire vocabulary; a word is just a vector of counts across all documents. We can make a similar observation with $\widehat{\boldsymbol{\Theta}}^\star$: regions are vectors of topic weight across all topics, while topics are analogously defined across regions. The key difference here as opposed to the sort of data matrices typically encountered in regression settings, say, is that the units are consistent between rows and columns. Pictorially,

$$\widehat{\boldsymbol{\Theta}}^\star = \begin{array}{c} \\ \text{Topic 1} \\ \text{Topic 2} \\ \vdots \\ \text{Topic } T \end{array} \begin{array}{c} \begin{array}{cccc} \text{Region 1} & \text{Region 2} & \cdots & \text{Region } R \end{array} \\ \begin{bmatrix} * & * & \cdots & * \\ * & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \end{bmatrix} \end{array}.$$

An obvious extension of our idea is to now consider the rows of $\widehat{\boldsymbol{\Theta}}^\star$, in which topics are written in terms of regions. The same issue applies however, as certain regions might be much "larger" than others. This means more topic weight is assigned in those regions in general, stemming originally from the disparity in sample size (i.e., number of documents) across regions. Not surprisingly, we need a column-normalizing transformation:

$$\widehat{\boldsymbol{\Theta}}^\star \mapsto \begin{bmatrix} | & | & & | \\ \boldsymbol{\theta}^\star_{\mathcal{R}_1} & \boldsymbol{\theta}^\star_{\mathcal{R}_2} & \cdots & \boldsymbol{\theta}^\star_{\mathcal{R}_R} \\ | & | & & | \end{bmatrix} \begin{bmatrix} 1/\|\boldsymbol{\theta}^\star_{\mathcal{R}_1}\|_2 & 0 & \cdots & 0 \\ 0 & 1/\|\boldsymbol{\theta}^\star_{\mathcal{R}_2}\|_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\left\|\boldsymbol{\theta}^\star_{\mathcal{R}_R}\right\|_2 \end{bmatrix} \tag{4.5}$$

with $\left\{\boldsymbol{\theta}^\star_{\mathcal{R}_j}\right\}$ the columns of $\widehat{\boldsymbol{\Theta}}^\star$ as before. Whereas the row-normalizing transformation ensures topic weight is unity *per topic*, we now require the same *per region*. This allows us to meaningfully inspect rows and discern which regions achieve more or less topic weight relative to a given topic. Unfortunately, in normalizing columns we destroy the row normalization achieved by the first transformation – these two views appear to be at odds with each other. One remedy is to simply apply whichever transformation is relevant to the question of interest, but the broken symmetry of this idea immediately suggests a more elegant solution: find a normalization which respects both rows and columns.

Were total topic weight distributed evenly amongst both regions *and* topics, we could move

seamlessly between the column and row-centric views, simultaneously answering both questions with which we began this section. Given that row and column normalization give the "correct" perspectives in turn, an obvious algorithm to try to force simultaneous row/column homogeneity would be to normalize both rows and columns in an alternating fashion, iterating until convergence (assuming the algorithm indeed converges). This method of so-called matrix balancing is referred to as the *Sinkhorn-Knopp algorithm* [SK67, Kni07], and is often used to produce a doubly-stochastic scaling of a given matrix; see algorithm (5). Under mild conditions, a doubly-stochastic scaling is possible if the matrix is square (and the norm used is ℓ_1). Our case is a sort of geometric analogue in which the rows and columns are both required to lie on a sphere in their respective spaces. Since $\widehat{\mathbf{\Theta}}^\star$ is rectangular in general, these spheres will be of different size. Empirically, we observe that for tall matrices (with $T > R$) columns will have unit norm (as they must, since they are normalized last before the algorithm's termination) and rows will have norm $\sqrt{\frac{R}{T}}$. This discrepancy is unimportant, because the algorithm still manages to balance the topic weight across both topics and regions.

An illuminating and compact formulation of Sinkhorn-Knopp is found by expressing the cumulative results of its iterations algebraically. Recalling (4.4) and (4.5), we note that the algorithm consists of repeated application of diagonal matrices to the left and right of $\widehat{\mathbf{\Theta}}^\star$:

$$\widetilde{\mathbf{\Theta}}^\star = \underbrace{\mathbf{D}_{\mathcal{T}}^{(m)} \cdots \mathbf{D}_{\mathcal{T}}^{(2)} \mathbf{D}_{\mathcal{T}}^{(1)}}_{\text{row normalization}} \widehat{\mathbf{\Theta}}^\star \underbrace{\mathbf{D}_{\mathcal{R}}^{(1)} \mathbf{D}_{\mathcal{R}}^{(2)} \cdots \mathbf{D}_{\mathcal{R}}^{(m)}}_{\text{column normalization}}$$

where we suppose convergence in m iterations. Combining the diagonal matrices from each iterate, we obtain a final pair $\mathbf{D}_{\mathcal{T}}$, $\mathbf{D}_{\mathcal{R}}$ so that

$$\widetilde{\mathbf{\Theta}}^\star = \mathbf{D}_{\mathcal{T}} \widehat{\mathbf{\Theta}}^\star \mathbf{D}_{\mathcal{R}}$$

has rows and columns each of equal ℓ_2 norm. Taking inverses and rearranging, we find

$$\widehat{\mathbf{\Theta}}^\star = \mathbf{D}_{\mathcal{T}}^{-1} \widetilde{\mathbf{\Theta}}^\star \mathbf{D}_{\mathcal{R}}^{-1}$$

which has a concrete interpretation, solidly grounding our methodology. Intuitively, every topic-region matrix is equivalent to a balanced counterpart plus "distortions" based on region size and topic importance. More formally, $\widehat{\mathbf{\Theta}}^\star$ is equivalent to $\widetilde{\mathbf{\Theta}}^\star$ under diagonal coordinate changes in both the row and column spaces. These coordinate changes scale the standard bases in each space to account for the unequal weight apportioned to both topics and regions. Finding such a coordinate change in one direction is not hard; the remarkable feature of Sinkhorn-Knopp is that it does so simultaneously in both directions. To summarize then, Sinkhorn-Knopp answers the question, "what would our topic-region matrix

look like if all regions were the same size and all topics were equally important?" Using a single matrix, this lets us read off answers to our two motivating questions, finding the topics which are given high (or low) weight in a given region and vice versa. We call these the topic-within-region (TwR) and region-within-topic (RwT) weights; see algorithm (5).

As statisticians, we would of course like to perform genuine inference on the TwR and RwT weights. For this, we turn away from the permutation methodology and towards the familiar nonparametric bootstrap which overlays seamlessly onto Sinkhorn-Knopp. Algorithm (6) describes bootstrapping formally, where the output is a histogram estimating the sampling distribution of the desired entries in $\widetilde{\mathbf{\Theta}}^\star$. We demonstrate this method in chapter 6, where the number of regions is small enough to allow for a single plot containing all such histograms for a given topic; this graphical RwT inference is not unlike plots of marginal posterior distributions that have become commonplace in visualization for Bayesian inference.

We conclude this chapter by drawing an interesting connection to the permutation tests of the previous section. Since Sinkhorn-Knopp restricts $\widetilde{\mathbf{\Theta}}^\star$'s row and column norms to $\sqrt{\frac{R}{T}}$ and 1 respectively, a "neutral" RwT/TwR weight will be equal to $1/\sqrt{T}$. In other words, an entry of $1/\sqrt{T}$ indicates no association between that particular region/topic pair. It is natural then to ask what happens when the entirety of $\widetilde{\mathbf{\Theta}}^\star$ is filled with such entries. Intuitively, this should mean that there is no association between regions and topics, which is essentially what we test using algorithms (3) and (4). Letting $\mathbf{J}$ be a matrix of ones, direct calculation gives

$$\begin{aligned}
\widehat{\mathbf{\Theta}}^\star &= \mathbf{D}_{\mathcal{T}}^{-1}\widetilde{\mathbf{\Theta}}^\star\mathbf{D}_{\mathcal{R}}^{-1} \\
&= \frac{1}{\sqrt{T}}\mathbf{D}_{\mathcal{T}}^{-1}\mathbf{J}\mathbf{D}_{\mathcal{R}}^{-1} \\
&= \frac{1}{\sqrt{T}}\begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_T \end{bmatrix}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_R \end{bmatrix} \\
&= \begin{bmatrix} | & | & & | \\ \frac{\beta_1}{\sqrt{T}}\boldsymbol{\alpha} & \frac{\beta_2}{\sqrt{T}}\boldsymbol{\alpha} & \cdots & \frac{\beta_R}{\sqrt{T}}\boldsymbol{\alpha} \\ | & | & & | \end{bmatrix}
\end{aligned}$$

where the α's and β's are general scalars and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_T)$. Since Sinkhorn-Knopp provides a unique decomposition [SK67], this shows that a homogeneous matrix is returned if and only if the input matrix has columns which are scalar multiples of each other. In our context, we have that $\widetilde{\mathbf{\Theta}}^\star$ is filled with $1/\sqrt{T}$'s if and only if $\widehat{\mathbf{\Theta}}^\star$ has collinear columns –

Algorithm 5: Sinkhorn-Knopp for TwR and RwT weights

- Construct $\widehat{\mathbf{\Theta}}^\star$.
- While not converged do:
  - Normalize the rows of $\widehat{\mathbf{\Theta}}^\star$.
  - Normalize the columns of $\widehat{\mathbf{\Theta}}^\star$.
- Output $\widetilde{\mathbf{\Theta}}^\star$, the balanced matrix.

Algorithm 6: Bootstrapping for TwR and RwT inference.

- Use algorithm (1) or (2) to find the anchor words (rows of $\mathbf{X}$).
- Solve (3.1) once to find the optimal $\mathbf{\Lambda}^\star$.
- Construct $\mathbf{\Theta}^\star$.
- For $b = 1, 2, \ldots, B$ do:
  - Bootstrap the columns of $\mathbf{\Theta}^\star$.
  - Construct $\widehat{\mathbf{\Theta}}^\star$ on the boostrapped sample by summing the entries row-wise within each region.
  - Apply algorithm (5) to $\widehat{\mathbf{\Theta}}^\star$, obtaining $\widetilde{\mathbf{\Theta}}^\star$ for bootstrap iterate b. Store the desired entries.
- Output histograms over the desired entries of each boostrapped $\widetilde{\mathbf{\Theta}}^\star$.

precisely the condition our test statistics are based on! Thus, the sense in which Sinkhorn-Knopp implies a lack of region-topic association is the same as that of our permutation tests. This unifies our two inferential programs, and suggests a new test statistic that may be more powerful. Namely, we could devise a χ^2-type test based on the deviation between $\widetilde{\mathbf{\Theta}}^\star$ and its expectation under the null, $1/\sqrt{T}\mathbf{J}$.

Chapter 5

An Application: "The Beer Data"

In this chapter we demonstrate the central methodologies of the thesis through a real-world example. Summarizing chapters 3 and 4, our approach to grouped topic modelling rests on four steps, each of which produces output that is of potential use to the analyst. These steps are:

1. The identification of anchor words. Anchor words give a quick summary of the corpus' thematic structure and are specifically designed to honour this principal. Recall that each anchor is associated with a particular topic and is defined so that it only appears within that topic.

2. The identification of topics (that is, $\mathbf{\Phi}^\star$). These give a richer description of the corpus' structure than do the anchor words alone by detailing how the entire vocabulary behaves within the given topic. It is common practice to rank words in each topic by pseudo-count and print the top fraction.

3. The testing of regional differences. We perform both overall and region-region tests, packaging the results as a series of plots depicting the histogram of permutation statistics under the null.

4. Computation of TwR and RwT weights with bootstrap estimates of their sampling distributions. We emphasize RwT inference especially, which give the extent to which regions are associated with particular topics. The bootstrap provides a fast means of estimating uncertainty in these weights.

The output associated with each of these steps is demonstrated in the applied analysis that follows.

Our data consists of online beer reviews from across Canada, obtained from a popular website while respecting their `robots.txt` policy. Reviews contain considerable descriptive language about a beer's flavour characteristics and the reviewer's overall drinking experience. The data underwent extensive pre-processing, as is almost always required in text analysis.

Punctuation and stopwords[1] – those containing little statistical signal such as "that" – were removed and reviews combined within beers. By this last point we mean that "documents" are taken to be the combined reviews of a particular beer. Words occurring less than 10 times in the corpus (after stopword removal) were omitted entirely. Ultimately, the data consists of $V = 9567$ unique words and $D = 6427$ documents. Algorithms were implemented in R and combined via a simple command line interface which was further developed into a fully-fledged software package, freely available at `https://gitlab.com/gabrielp19/infernmf`. We rely crucially on the aforementioned `LAPACK` [ABB+99] package for numerical linear algebra routines and `nnls` [Mv12] for solving the non-negative least squares problems required of tractable NMF.

5.1 Anchor Words and Recovered Topics

Table (5.1) gives $T = 100$ anchor words as found by algorithm (2), which does seem to capture a highly informative set of words from which to form topics. Looking forward to inference by permutation tests, a natural question of interest is whether Canadians' beer or flavour preferences vary regionally. To that end, these anchor words are promising in that they appear to primarily be about tastes, flavours, and beer styles (along with some other descriptors and sentiment words). A curious feature of anchor words as a way of determining topics is that they do not depend on the choice of T. Had we specified $T = 200$, the first 100 of those anchors would be precisely the words presented in table (5.1). This has important ramifications for selecting the number of topics. For one, it implies a sort of robustness that is not shared by other methods like LDA. There, topic makeup itself can be sensitive to the choice of T whereas in this regime picking the number of topics is similar to choosing the number of components in PCA. We may "over-pick", but the optimal topics remain intact as a subset of those chosen. One idea is to examine the upper triangular factor produced by the QR decomposition with column pivots to determine when, if ever, the numerical rank of $\mathbf{X}$ becomes saturated. In practice we found this unhelpful, much in the same way that scree plots are often inconclusive. In the end, we adopt a pragmatic "topics by eye" approach in which we look for the point at which anchors appear to add little new subject matter. We can then trim the number of topics appropriately. Furthermore, this can all be done before solving the optimization problem posed by NMF. In this way, we can intervene in model selection *before* a model is fully specified and fit. This is in keeping with our quest for computational efficiency.

Table (5.2) shows the 10 most prevalent words within 25 randomly chosen topics, labelled by their anchor word in the far left column. Topics appear to be genuinely meaningful,

[1] Various R packages give lists of English stopwords – we have used the `stopwords` package [BMW20].

1	coffee	**26**	sweet	**51**	corn	**76**	smooth
2	beer	**27**	oak	**52**	hop	**77**	lime
3	dark	**28**	lager	**53**	stout	**78**	flavour
4	apple	**29**	nice	**54**	dry	**79**	smell
5	vanilla	**30**	taste	**55**	bottle	**80**	white
6	alcohol	**31**	honey	**56**	carbonation	**81**	quad
7	hops	**32**	fruit	**57**	amber	**82**	creamy
8	yeast	**33**	porter	**58**	malts	**83**	pretty
9	pumpkin	**34**	orange	**59**	saison	**84**	tripel
10	light	**35**	brown	**60**	notes	**85**	mild
11	chocolate	**36**	black	**61**	bitterness	**86**	earthy
12	maple	**37**	bit	**62**	blueberry	**87**	spicy
13	ipa	**38**	pale	**63**	red	**88**	body
14	caramel	**39**	grapefruit	**64**	lemon	**89**	nose
15	cherry	**40**	raspberry	**65**	cranberry	**90**	fruity
16	wheat	**41**	smoke	**66**	pine	**91**	grassy
17	pepper	**42**	belgian	**67**	pineapple	**92**	barleywine
18	citrus	**43**	rye	**68**	finish	**93**	coriander
19	apricot	**44**	tea	**69**	head	**94**	bready
20	floral	**45**	flavor	**70**	cinnamon	**95**	toffee
21	oatmeal	**46**	style	**71**	aroma	**96**	hibiscus
22	ale	**47**	ginger	**72**	sour	**97**	herbal
23	bourbon	**48**	banana	**73**	roasted	**98**	glass
24	unibroue	**49**	medium	**74**	sweetness	**99**	tropical
25	malt	**50**	bitter	**75**	licorice	**100**	decent

Table 5.1: Anchor words found by QR decomposition with column pivots.

rounding out the theme initiated by the anchor word. The diversity of topics is particularly striking, as is their lack of crossover. Because anchor words are confined to a particular topic and in some sense dictate that topic's content, we avoid the sort of vague and repetitious topics that are sometimes present using other approaches. Note anchor words always appear as the most prevalent word within a topic; this is not guaranteed theoretically but is almost always observed in practice.

5.2 Permutation Tests

The relevant inferential question is whether beer styles and/or preferences vary geographically. In addition to the reviews, we obtained geographic information about where beers are brewed, and were thus able to partition the documents as formalized in chapter 4. Note that regions are defined by where a beer is brewed, and not where reviews are written. Partitioning according to province would produce wildly disparate sample sizes as some provinces have very few breweries. Instead, we combined all Atlantic provinces into a region `Atlantic` and the Prairie provinces into `Prairies`. The remaining regions are `BC`, `Quebec`,

anchor	1	2	3	4	5	6	7	8	9	10
beer	beer	beers	tastes	time	lot	drink	drinking	mouth	hint	hard
bit	bit	lot	overly	worth	feels	stick	expected	reminds	left	tad
bitterness	bitterness	lacing	mouthfeel	bodied	pick	retention	balance	time	pronounced	lot
black	black	opaque	tan	tastes	sticky	worth	tobacco	day	background	pours
blueberry	blueberry	blueberries	artificial	fresh	berry	summer	beers	refreshing	grainy	faint
bottle	bottle	sticky	time	lots	huge	complexity	bottles	palate	slowly	pour
carbonation	carbonation	lot	fine	aromas	enjoyable	type	feel	coming	alcohol	ale
creamy	creamy	balanced	feel	mouthfeel	cream	mouth	lingering	time	lingers	flavors
dark	dark	fruits	tan	complex	raisins	deep	plums	plum	molasses	rich
flavor	flavor	flavors	color	mouthfeel	drinkability	drink	slight	bite	tongue	feel
flavour	flavour	flavours	colour	retention	ontario	coloured	lace	cap	nutty	final
hibiscus	hibiscus	pink	flowers	rose	tart	flowery	refreshing	unique	flower	smells
honey	honey	golden	pear	pears	esters	tastes	gold	duvel	cloying	syrupy
lemon	lemon	refreshing	yellow	golden	straw	hazy	zest	lemony	gold	hot
lime	lime	soda	salt	bud	seltzer	fizzy	beers	natural	bubbles	summer
maple	maple	syrup	scotch	sugar	boozy	heavy	syrupy	booze	wee	molasses
pineapple	pineapple	golden	mango	abv	stuff	gold	palate	touch	bright	fresh
porter	porter	porters	watery	roast	blackberry	flavours	baltic	sleeman	cocoa	cola
smooth	smooth	brew	drinkable	excellent	perfect	slight	hint	tongue	delicious	highly
spicy	spicy	spice	spices	spiciness	strong	clove	complex	drink	color	brew
stout	stout	stouts	mocha	roast	roasty	barley	pitch	silky	hints	char
style	style	finger	lots	clean	heavy	initially	true	prefer	tongue	drinkability
tripel	tripel	golden	clove	tripels	yellow	abv	complex	balanced	perfect	triple
tropical	tropical	mango	juicy	tangerine	citra	melon	papaya	rind	dank	peach
wheat	wheat	summer	refreshing	golden	wheaty	wit	witbier	tart	mouthfeel	finishes

Table 5.2: 10 most prevalent words within 25 randomly chosen topics.

and `Ontario`. Territories were omitted due to the small number of breweries contained therein. Figure (5.2) depicts all 10 region-region tests conducted according to algorithm (4). The histogram of permutation test statistics is displayed as well as the observed test statistic and associated p-value. We find that the topic makeup of all regions are pairwise significantly different, save for `Atlantic` vs. `Quebec` and `Atlantic` vs. `Ontario`. This is corroborated by examining the TwR weights as displayed in table (5.3), which lists the topics given the highest and lowest weight within each region. After balancing the region-topic matrix with Sinkhorn-Knopp, the most popular topics within each region are quite distinct, suggesting a kind of "flavour profile" that each exhibits. The failure to reject in the two aforementioned tests is likely down to `Atlantic`'s sample size, the smallest by far. We need not worry about inflated type I error due to multiple testing, as all significant p-values would survive Bonferroni corrections or similar. Those cases with $p = 0$ result from the limited number of random permutations; such cases should be interpreted as having tiny (but non-zero) p-values. Needless to say, the above region-region results are accompanied by a rejection of the null in the overall test of regional differences. In this case the gap between the observed test statistic and the permutation statistics is quite large and we omit the resulting plot.

5.3 RwT and TwR Weights

To explore the differences between regions with respect to specific topics, we employ the RwT weights of chapter 4 which are displayed as a series of bootstrap sampling distributions over the relevant entries of the balanced region-topic matrix. Recall that these are the weights assigned to each region under a particular topic, balanced by Sinkhorn-Knopp. This

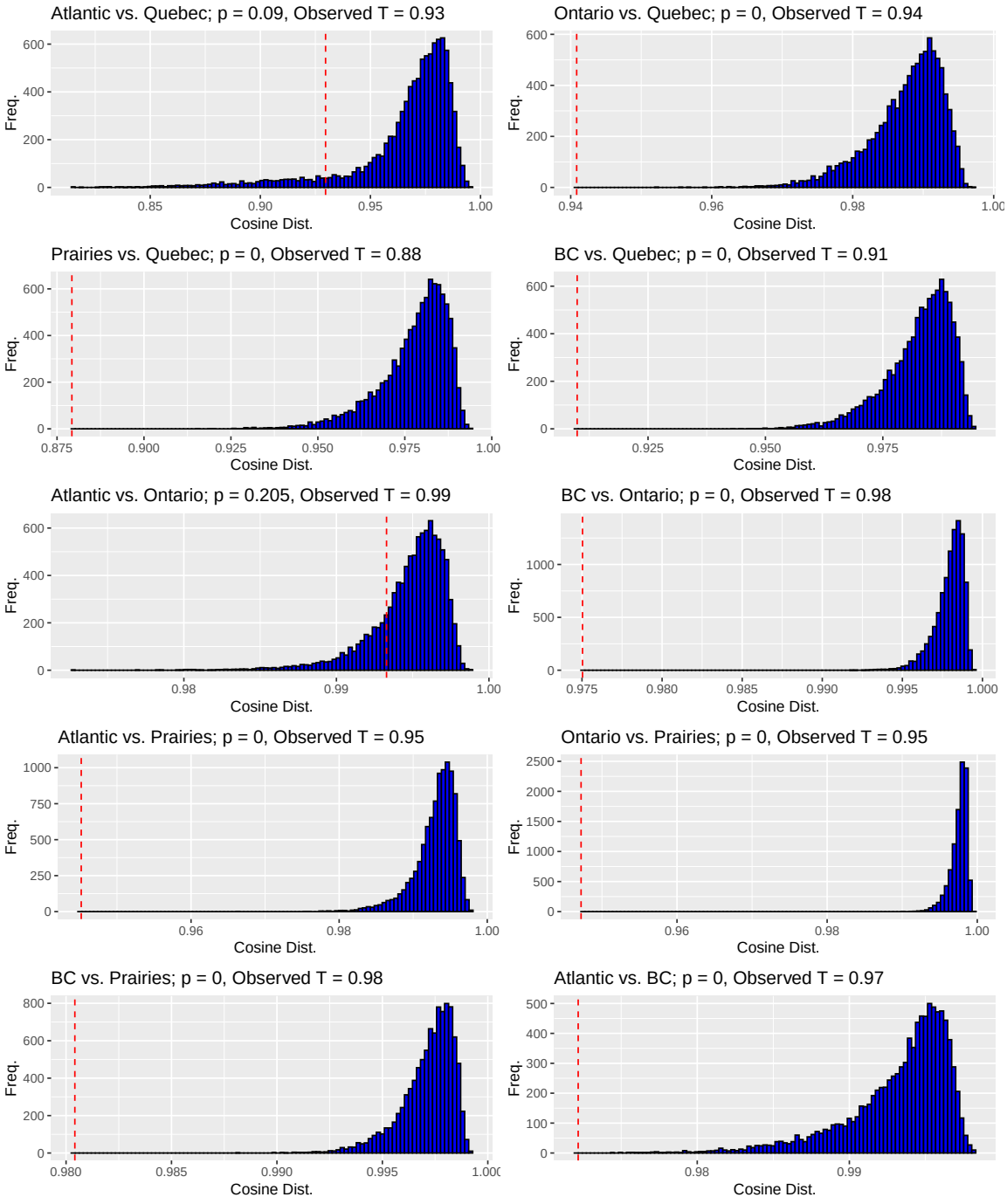


Figure 5.1: Tests of region equality using $s = 10000$ random permutations. Tests are based on the cosine distance between the topic weight vectors from two regions. The dashed red line gives the value of the observed test statistic.

	Atlantic	BC	Ontario	Prairies	Quebec
1	blueberry	unibroue	cranberry	earthy	pumpkin
2	lime	tripel	rye	bready	ginger
3	lager	quad	tea	floral	pine
4	corn	hibiscus	pineapple	herbal	pineapple
5	honey	belgian	corn	glass	tropical
6	ipa	coriander	porter	malt	grapefruit
7	aroma	alcohol	oak	barleywine	ipa
8	maple	apple	malts	dry	bourbon
9	bitter	yeast	nose	pale	citrus
10	light	pepper	lager	amber	licorice
91	pepper	grassy	hibiscus	pumpkin	belgian
92	oatmeal	earthy	oatmeal	oak	cranberry
93	apricot	grapefruit	pepper	cranberry	apple
94	saison	pale	yeast	flavor	lager
95	belgian	ipa	belgian	blueberry	corn
96	bourbon	pine	quad	belgian	tripel
97	tripel	porter	tripel	hibiscus	quad
98	unibroue	blueberry	blueberry	tripel	hibiscus
99	hibiscus	lager	barleywine	quad	blueberry
100	quad	corn	unibroue	unibroue	unibroue

Table 5.3: Topics ranked within each region according to their TwR weights. The top and bottom 10 topics are listed for each region.

balancing allows us to meaningfully compare across both regions and topics; figures (5.2) and (5.3) give examples. Recall that the RwT weights are constrained so that their squares sum to $R/T = .05$ within each topic. We overly a line at $1/10$ – the value which would be attained for each region under a completely neutral topic. Polarizing topics like `yeast` tend to have relatively little uncertainty associated with them, whereas the universality of `stout`, say, gives rise to diffuse RwT weights which preclude association to a particular set of regions. The large variance exhibited by `Atlantic`'s RwT weights is due to the region's relatively small sample size (i.e., number of breweries). A similar treatment of TwR weights is not given here, as the resulting plots would contain 100 overlaid densities. Instead, the aforementioned table (5.3) ranks the TwR weights for each region and shows the top and bottom 10 topics.

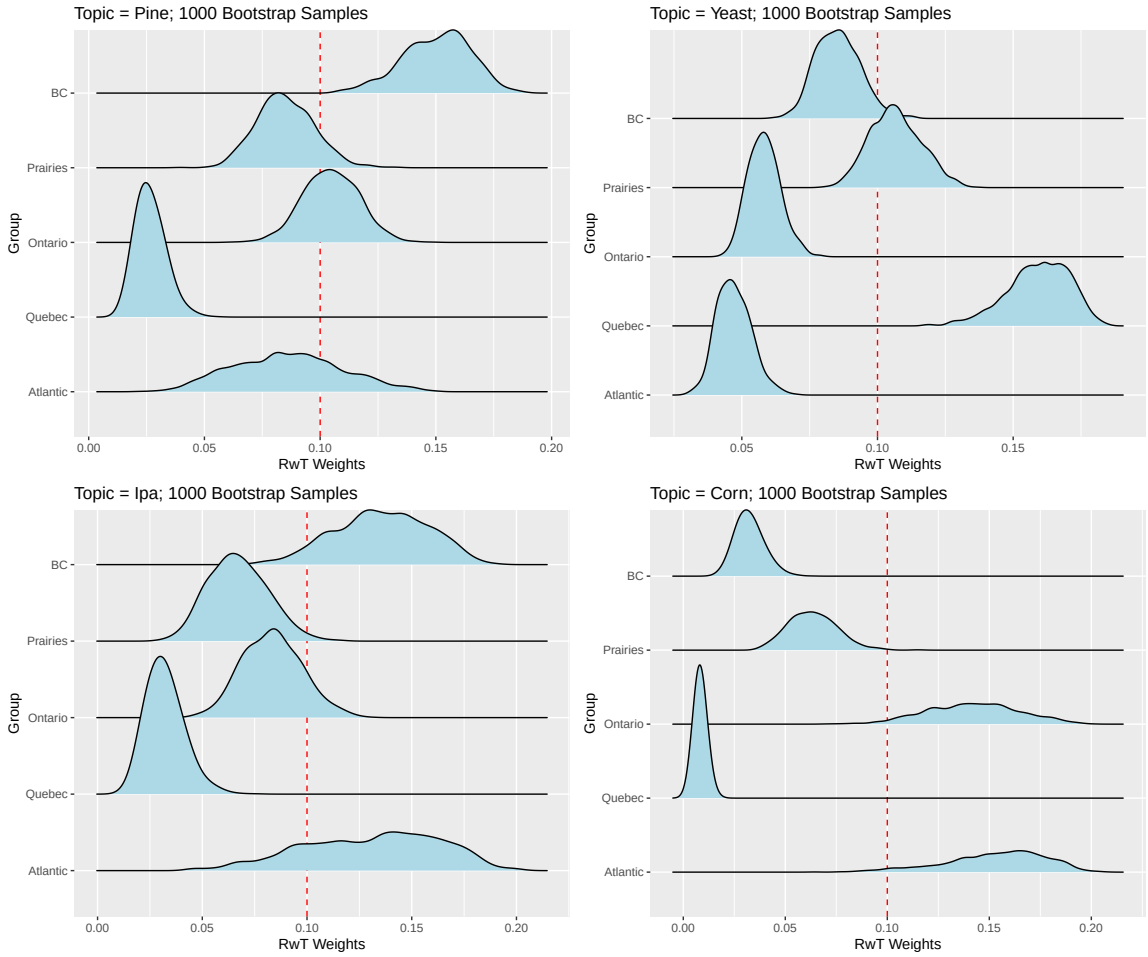


Figure 5.2: RwT weights for the four indicated topics (as identified by their anchor words) using $B = 1000$ bootstrap samples to estimate the sampling distribution.

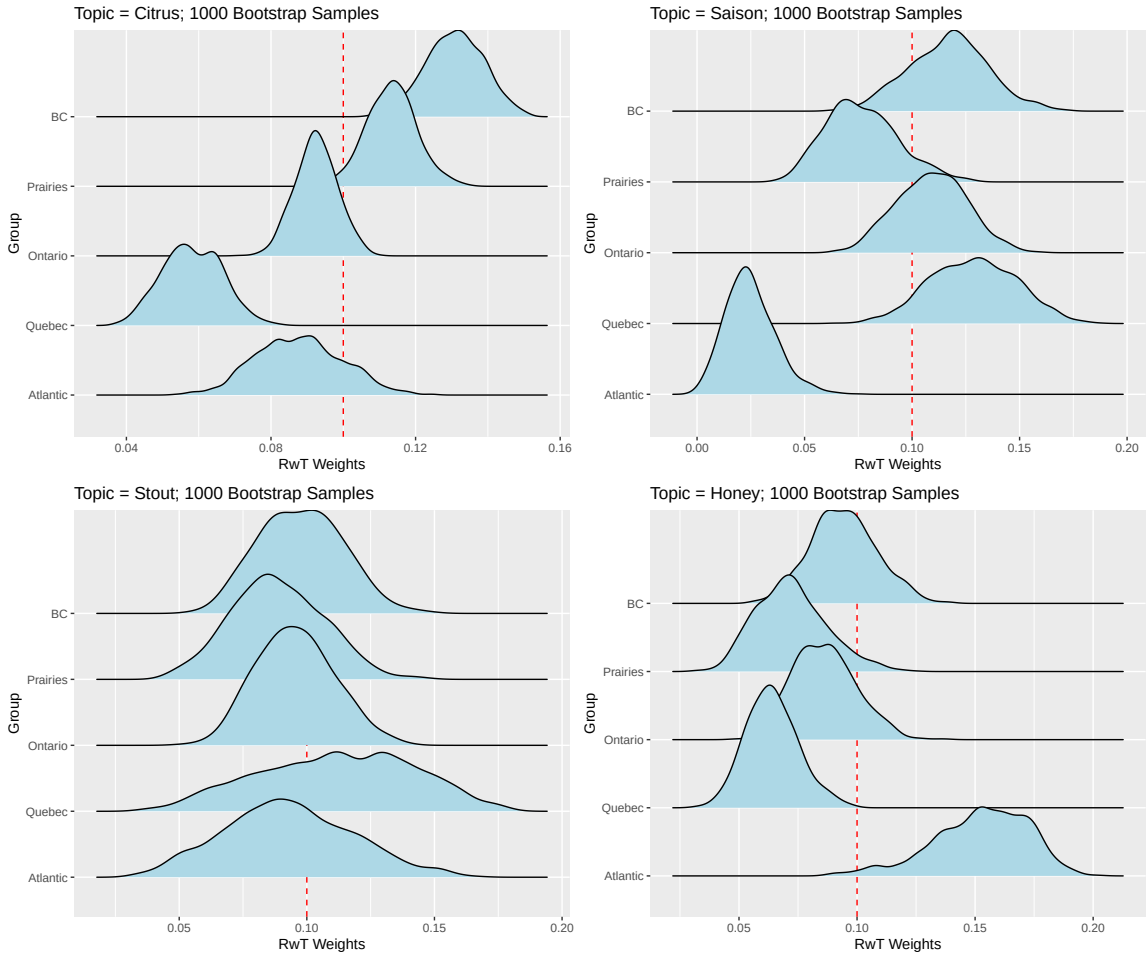


Figure 5.3: RwT weights for an additional four topics with $B = 1000$ as in figure (5.2).

Chapter 6

Conclusions and Future Work

This thesis explored a variety of ideas related to non-negative matrix factorization and resampling-based inference, culminating in their combination for inference in grouped topic models. Despite this concrete end-goal, much of the intermediary material is worth highlighting in its own right. First, we drew attention to recent developments in non-negative matrix factorization, without which the subsequent material would have been impossible. In particular, anchor words allow for NMF to be solved globally and efficiently, a fact that – despite years of development – few practitioners seem to know about. Anchor words can be found using tools from numerical linear algebra, and the approximate non-negative factorization found by convex programming leads to rich and diverse topics when applied in the context of text analysis. Our approach to this problem was novel in that we deliberately eschewed a generative stochastic model, embracing an optimization-based perspective that merged particularly well with permutation tests and the bootstrap when considering the NMF solution a statistic under a weakly stochastic sampling model. Our inference techniques are underscored by their simplicity, especially when compared with hard-to-compute and approximation-laden probabilistic topic models. The efficacy of these methodologies were demonstrated on a data set of online beer reviews. A general philosophy permeated the entire work, which we believe deserves further attention in the statistics community: the application of classical nonparametric resampling methods to general, distribution-free statistics (like the solution to an NMF problem). This approach allows for the amalgamation of statistical learning techniques with genuine frequentist inference, computational efficiency being a noteworthy driver of the process. Making these ideas more robust and user-friendly is a worthy future goal of the statistics community.

Many future research questions, which we hope to address, were motivated while completing this work. For example:

1. We cited work improving upon the QR decomposition with column pivoting. Do these subset selection methods offer any improvement over vanilla QR? Do their theoretical guarantees manifest in practice?

2. Are there better procedures for finding anchor words in general, perhaps unrelated to QR?

3. We recently became aware of a method known as *Correspondence Analysis* [Ben92], which elegantly addresses the fact that text data is not really Euclidean in nature (though NMF treats it as such). Can this be made to work with the methods provided herein?

4. Are there other suitable test statistics for conducting permutation tests in grouped topic models (where we strictly enforce computational tractability)?

5. How can the permutation methodology be used to efficiently construct confidence intervals or similar in this setting?

6. How can other resampling methods be reposed on top of NMF in an efficient way?

7. Are there different ways in which the Sinkhorn-Knopp algorithm can be utilized (by applying it directly to the TDM, for example)?

8. Can regional variation be modelled continuously while retaining the attractive features of our approach? Does this provide inferential advantages not found treating regions as discrete entities?

We suspect and hope that most of these questions can be answered in the positive.

Bibliography

[ABB+99] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[AGH+18] Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. Learning Topic Models – Provably and Efficiently. *Commun. ACM*, 61(4):85–93, March 2018.

[AGKM12] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a Nonnegative Matrix Factorization – Provably. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '12, page 145–162, New York, NY, USA, 2012. Association for Computing Machinery.

[AGM12] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning Topic Models – Going Beyond SVD. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, page 1–10, USA, 2012. IEEE Computer Society.

[BBP+10] Mary Broadbent, Martin Brown, Kevin Penner, Ilse Ipsen, and Rizwana Rehman. Subset Selection Algorithms: Randomized vs. Deterministic. *SIAM Undergraduate Research Online*, 3, 01 2010.

[BD08] Christos Boutsidis and Petros Drineas. Random Projections for the Nonnegative Least-Squares Problem. *Linear Algebra and its Applications*, 431:760–771, 12 2008.

[Ben92] Jean-Paul Benzecri. *Correspondence Analysis Handbook*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1992.

[BG65] Peter Businger and Gene H. Golub. Linear Least Squares Solutions by Householder Transformations. *Numer. Math.*, 7(3):269–276, June 1965.

[BGY+01] Saldanha Bezerra, Roberto Galvão, Takashi Yoneyama, Henrique Chame, and Valeria Visani. The Successive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis. *Chemometrics and Intelligent Laboratory Systems*, 57:65–73, 07 2001.

[BHPJ10] O. Berné, A. Helens, P. Pilleri, and C. Joblin. Non-negative Matrix Factorization Pansharpening of Hyperspectral Data: An Application to Mid-infrared

Astronomy. In *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pages 1–4, 2010.

[BL05] David M. Blei and John D. Lafferty. Correlated Topic Models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, page 147–154, Cambridge, MA, USA, 2005. MIT Press.

[BM07] David M. Blei and Jon D. McAuliffe. Supervised Topic Models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 121–128, Red Hook, NY, USA, 2007. Curran Associates Inc.

[BMW20] Kenneth Benoit, David Muhr, and Kohei Watanabe. *stopwords: Multilingual Stopword Lists*, 2020. R package version 2.0.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Cha87] Tony F. Chan. Rank Revealing QR Factorizations. *Linear Algebra and its Applications*, 88-89:67 – 82, 1987.

[CP] Donghui Chen and Robert J. Plemmons. *Nonnegativity Constraints in Numerical Analysis*, pages 109–139.

[DDF+90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[DG03] Sanjoy Dasgupta and Anupam Gupta. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[DLP+16] V. Duong, Y. Lee, B. Pham, P. T. Bao, and J. Wang. Nmf-based Image Segmentation. In *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2, 2016.

[DM16] Petros Drineas and Michael W. Mahoney. RandNLA: Randomized Numerical Linear Algebra. *Commun. ACM*, 59(6):80–90, May 2016.

[DS04] David Donoho and Victoria Stodden. When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts? In S. Thrun, L. K. Saul, and B. Scholkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1141–1148. MIT Press, 2004.

[EAX11] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse Additive Generative Models of Text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1041–1048, Madison, WI, USA, 2011. Omnipress.

[Eng97] H. Engler. The Behavior of the QR-Factorization Algorithm with Column Pivoting. *Applied Mathematics Letters*, 10(6):7 – 11, 1997.

[EOSX10] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA, October 2010. Association for Computational Linguistics.

[GE96] Ming Gu and Stanley C. Eisenstat. Efficient Algorithms for Computing a Strong Rank-Revealing QR Factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[Gil14] Nicolas Gillis. The Why and How of Nonnegative Matrix Factorization, 2014.

[Goo94] P.I. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer series in statistics. Springer-Verlag, 1994.

[GS04] Thomas L. Griffiths and Mark Steyvers. Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[GVL83] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in Atlantic History & Culture. Johns Hopkins University Press, 1983.

[Hof99] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA, 1999. Association for Computing Machinery.

[HW78] A. Hedayat and W. Wallis. Hadamard Matrices and Their Applications. *The Annals of Statistics*, 6, 11 1978.

[IM98] Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, page 604–613, New York, NY, USA, 1998. Association for Computing Machinery.

[JL84] William Johnson and Joram Lindenstrauss. Extensions of Lipschitz Maps into a Hilbert Space. *Contemporary Mathematics*, 26:189–206, 01 1984.

[JM09] D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artifi. Pearson Prentice Hall, 2009.

[Kni07] Philip Knight. The Sinkhorn–Knopp Algorithm: Convergence and Applications. volume 30, 01 2007.

[LS96] D. D. Lee and H. S. Seung. Unsupervised Learning by Convex and Conic Coding. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 515–521. MIT Press, Cambridge, MA, USA, 1996.

[LS99] Daniel Lee and H. Seung. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 401:788–91, 11 1999.

[LS01] Daniel D. Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

[MB01] P.W. Mielke and K.J. Berry. *Permutation Methods: A Distance Function Approach*. Springer series in statistics. Springer, 2001.

[MM08] David Mimno and Andrew McCallum. Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 411–418, Arlington, Virginia, USA, 2008. AUAI Press.

[Mv12] Katharine M. Mullen and Ivo H. M. van Stokkum. *nnls: The Lawson-Hanson Algorithm for Non-negative Least Squares (NNLS)*, 2012. R package version 1.4.

[RGB14] Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[Rou19] Tim Roughgarden. Beyond Worst-Case Analysis. *Commun. ACM*, 62(3):88–96, February 2019.

[RSA16] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003, 2016.

[RST19] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2):1–40, 2019.

[SK67] Richard Sinkhorn and Paul Knopp. Concerning Nonnegative Matrices and Doubly Stochastic Matrices. *Pacific J. Math.*, 21(2):343–348, 1967.

[SOAC+18] Genevieve Stein-O'Brien, Raman Arora, Aedin Culhane, Alexander Favorov, Lana Garmire, Casey Greene, Loyal Goff, Yifeng Li, Alioune Ngom, Michael Ochs, Yanxun Xu, and Elana Fertig. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends in Genetics*, 34, 08 2018.

[TB97] L.N. Trefethen and D. Bau. *Numerical Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1997.

[TGCD16] Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

[Vav10] Stephen A. Vavasis. On the Complexity of Nonnegative Matrix Factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

[Was] Larry Wasserman. Modern Two-Sample Tests. Blog Post.

[YCH+11] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. Geographical Topic Discovery and Comparison. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 247–256, New York, NY, USA, 2011. Association for Computing Machinery.