Applications of Latent Variable Models in Modeling Influence and Decision Making

Sean M. Gerrish

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Adviser: David M. Blei

May 2012

Abstract

The abstract goes here.

Acknowledgements

This work was supported by grants from the Office of Naval Research, ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P. Sloan foundation, and a grant from Google.

Foremost, I owe my advisor, David Blei, many thanks for his mentorship and support for the past four years. The preponderance of this mentorship has been on research, but Dave's support during the "middle years" also helped me to press through the program.

I would also like to thank Leon Bottou, from whom I learned more than I should admit while TAing for him: Leon's deep understanding of mathematics and pleasant manner are a ...

I owe many thanks to a number of current and former graduate students, who have helped me in this program both academically and personally. Jordan Boyd-Graber – tips on organization Jonathan Chang – stupendous music collection Chong Wang Indraneel – inspiring chats about our futures Gungor Sam and John

Friends

My parents deserve many thanks, for their support in the past three decades. Those early years, giving whatever career goals I have.

My brother Jason has provided me with steadfast encouragement and reinforcement through this period, and I may have never become interested in computer science if my brother Josh hadn't shown me things like BASIC, Robot Odyssey, logic puzzles, division, fractals,

I would like to thank my committee, who ...

- (reader)
- (reader)
- (reader)
- (non-reader)
- (non-reader)

And I thank my dear sister Kim, who would have liked to see me graduate. Kim shared many conversations with my my first year of graduate school and inspired me the rest of the years, through my generals, several tough exams, and through to the end.

This thesis is dedicated to my sister, Kimberly Mesa.

Contents

Chapter 1

Previous Publications

The work here represents expanded versions of the following publications:

- a
- b

The current presentation attempts to provide deeper and broader understanding to the reader.

Chapter 2

Introduction

Information is all around us. Society interacts with this information in a complicated dance: information affects what we do, and we create further information – books, scientific papers, legislation, and tweets – as a result. Information influences everything we say, think, and do.

The point of this thesis is not to prove this point; that information influences us in everything we do is self-evident. This thesis will describe a number of quantitative methods for modeling the influence of text on society.

We begin by ...

We also ...

2.1 Influence and decision making

2.2 The availability of observational social science data on a massive scale

2.2.1 Text

2.2.2 Dyadic data: networks and interaction between entities

2.2.3 Time-series data

2.2.4 The insufficiency of traditional methods

2.2.5 Causal inference in modern observational setting

2.3 The role of statistical machine learning

2.3.1 Probabilistic latent variable models

2.3.2 Tools and abstractions for probabilistic inference

2.3.3 Ability to perform inference on large-scale datasets

Chapter 3

Preliminary material: quantitative methods

3.1 Standards and naming conventions

We begin by outlining naming and variable conventions in this work. Random variables and their instantiations are given by roman or greek characters; the role of a variable will typically evident from its context. Multivariate random variables such as vectors are given by boldface, and collections of random variables are given by uppercase Roman characters.

The reader may find Table **??** a helpful resource in the subsequent chapters. This table summarizes many of the variables described in this work.

Table 3.1: table:notation

Variable	**Description**
d	Document (subscript)
θ_d	Topic mixture for document d
z_n	$K-$variate topic indicator for term n
w	A collection of words, as in a document
α	Dirichlet parameter for LDA
u	Person, e.g., a lawmaker (subscript)
x_u	An ideal point indicating an individual's sentiment
a_d	A document's polarization
b_d	A document's popularity

Figure 3.1: Data analysis pipeline

3.2 Latent-variable models for exploratory analysis

We develop the ideas outlined in the last chapter using the process of data analysis outlined in Figure **??**. This pipeline, which is driven by a specific question, proceeds with the development of a latent-variable model to answer that question or to answer many questions like it. Once a model is selected, we then derive and implement an algorithm to estimate the values of the latent random variables in it. We will variously refer to this stage of the process as *fitting a model*, *performing inference*, and *fitting the posterior*.

This is then followed by a process of data analysis and exploration, along with the drawing of conclusions. This may include visualization tools such as **?**(chaney and blei).

In some cases, the model may be revised. This revision is ideally driven by findings in the data analysis step, although our decisions in the model development step are of course informed by limitations in the tools available for performing inference and the ease of subsequent data analysis.

3.2.1 Latent-variable models

As alluded to above, we will focus on latent-variable models in stage 2 of the pipeline. While some alternatives exist, latent variable models have several benefits that appeal to us:

1. Flexibility. These models can describe, summarize, and explain a wide variety of phenomena in the physical and social sciences.

2. Embeddability and Interpretability. Any quantifiable metric in the dataset can be encoded as a random variable in a probabilistic model. The relationship between metrics can be likewise encoded explicity.

3. Modularity. Parts of these models can be re-used across different models. This leads to efficient transfer of resources and common paradigms.

4. Existing toolbox of statistical tools. There is a large and growing body of literature around how to fit these models. Practitioners no longer need to be experts in statistics to correctly apply many of these tools.

The risk with applying latent-variable models is that the credibility and careful deliberation we often associate with statistics suggests that an estimated posterior must be credible. This may not

be true, particularly when the model is poorly embedded into a statistical model, or when it is incorrectly interpreted. Both of these happen when Item **??** above is carried out carelessly.

3.2.2 Text as a medium for social science analysis

Text data is the low-hanging fruit of most social science research questions. Its ubiquity is due to the ease with which it can be both created and digitized. At the same time, it provides a rich source of data: documents, one of the basic units of information in text analysis, are an observation in an extremely high-dimensional and interpretable space **?**.

? provides an excellent overview of text analysis; we will summarize several methods here.

Text is inherently extremly high-dimensional. A large collection of documents represented by a sequence $\boldsymbol{w}_n$ of words would be unweildly for even a human to describe. A number of tools have been developed over the past several decades to simply find the *gist* of documents, making it possible to describe

Before describing these tools in more detail, we note that a simplifying assumption that we can make about a text document d is that it can be represented as a vector $\boldsymbol{w}_d \in \mathcal{R}^V$ of word counts. This assumption, known as the *bag of words* assumption, removes most of the information in a document, but it allows us to capture the "gist" of a document very well.

Latent Dirichlet Allocation

Latent Dirichlet Allocation

Unigram models

Text regression

Text regrssion is a

3.2.3 Ideal point models and matrix factorization

- relevant sources - mathematical foundation - examples - hinge loss - inference (stochastic)

3.2.4 Hidden Markov Models and Kalman Filters

3.3 Posterior inference and posterior evaluation

3.3.1 MAP estimation

3.3.2 Variational inference

- objective

3.3.3 Stochastic optimization

- examples

3.3.4 Posterior Predictive Checks

- examples

Chapter 4

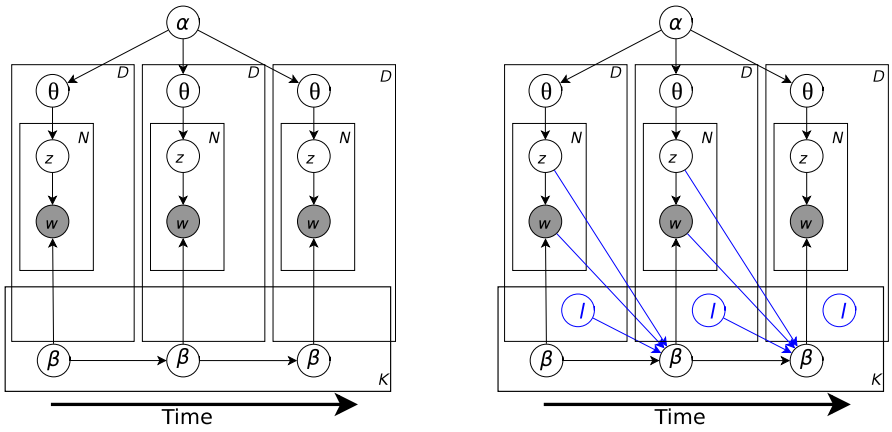
A model of influence in text documents

4.1 Introduction

Measuring the influence of a scientific article is an important and challenging problem. Influence measurements are used to assess the quality of academic instruments, such as journals, scientists, and universities, and can play a role in decisions surrounding publishing and funding. They are also important for academic research: finding and reading the influential articles of a field is central to good research practice.

The traditional method of assessing an article's influence is to count the citations to it. The impact factor of a journal, for example, is based on aggregate citation counts **?**. This is intuitive: if more people have cited an article, then more people have read it, and it is likely to have had more impact on its field. Citation counts are used with other types of documents as well: the Pagerank algorithm, which uses hyperlinks of web-pages, has been essential to Google's early success in Web search **?**.

Though citation counts can be powerful, they can be hard to use in practice. Some collections, such as news stories, blog posts, or legal documents, contain articles that were influential on others but lack explicit citations between them. Other collections, like OCR scans of historical scientific literature, do contain citations, but they are difficult to read in reliable electronic form. Finally, citation counts only capture one kind of influence. All citations from an article are counted equally in an impact factor, when some articles of a bibliography might have influenced the authors more



(a) (b)

Figure 4.1: The Dynamic Topic Model (a) and the the Document Influence Model (b).

than others.

We take a different approach to identifying influential articles in a collection. Our idea is that an influential article will affect how future articles are written and that this effect can be detected by examining the way corpus statistics change over time. We encode this intuition in a time-series model of sequential document collections.

We base our model on dynamic topic models, allowing for multiple threads of influence within a corpus **?**. Though our algorithm aims to capture something different from citation, we validate the inferred influence measurements by comparing them to citation counts. We analyzed one hundred years of the Proceedings of the National Academy, one hundred years of *Nature*, and a forty year corpus of articles on computational linguistics. With only the language of the articles as input, our algorithm produces a meaningful measure of each document's influence in the corpus.

4.2 The Document Influence Model

We develop a probabilistic model that captures how past articles exhibit varying influence on future articles. Our hypothesis is that an article's influence on the future is corroborated by how the language of its field changes subsequent to its publication. In the model, the influence of each article is encoded as a hidden variable and posterior inference reveals the influential articles of the collection.

Our model is based on the dynamic topic model (DTM) **?**, a model of sequential corpora that allows language statistics to drift over time. Previous probabilistic models of text assumed that the underlying distributions over words were fixed. The DTM introduced a Markov chain of term distributions to capture probabilities that drift over the course of the collection.

Let V be the number of words in a vocabulary and consider the natural parameters β_t of a term distribution at time t, where the probability of a word w is given by the softmax transformation of the unconstrained vector,

$$p(w \mid \beta_t) \propto \exp(\beta_{t,w}). \tag{4.1}$$

The corresponding distribution over terms, i.e., the "topic," is a point on the vocabulary simplex. In the logistic normal Markov chain, this distribution drifts with the stationary autoregressive process

$$\beta_{t+1} \mid \beta_t \sim \mathcal{N}(\beta_t, \sigma^2 I), \tag{4.2}$$

where σ^2 is the transition variance.

Now consider a corpus with D articles at each time t and let the rows $\mathbf{w}_{t,d}$ of $\mathbf{W}_{t,1:D}$ denote the articles as vectors of word counts. In the simplest DTM, a single distribution over words drifts according to Equation **??**. For each time point, the words of its articles are drawn independently from Equation **??**. One can then compute the posterior distribution of the sequence of topics $\beta_{1:T}$ conditioned on the observed documents. This summarizes the corpus as a smooth trajectory of word frequencies.

We now turn to our central idea: some articles influence the topic more than others. In our model, each article is assigned a normally distributed *influence score* ℓ_d, which is a scalar value that describes the influence that article d has on the topic. The higher the influence, the more the words of the article affect how the topic drifts.

This is encoded in the time series model. The more influential a document is, the more its words "nudge" the topic's natural parameters at the next time step,

$$\begin{aligned} \beta_{t+1} \mid \beta_t, (w, \ell)_{t,1:D} \sim \\ \mathcal{N}\left(\beta_t + \exp(-\beta_t) \textstyle\sum_d w_{d,t} \ell_{t,d}, \sigma^2 I\right). \end{aligned} \tag{4.3}$$

The words of an article with a high influence will have a higher expected probability in the next epoch; the words of an article with zero influence will not affect the next epoch.

We call this model the *document influence model* (DIM). Conditioned on a corpus, the posterior distribution of the topic and influence scores gives a trajectory of term frequencies and a retrospective estimate of the influence of each article. An article whose words can help explain the way the word frequencies change will have a high posterior influence score. We show in Section **??** that this estimate of influence is meaningful.

Multiple topics. Corpora typically contain multiple persistent themes. Accordingly, the full dynamic topic model contains multiple topics, each associated with a time series of distributions. Conditioned on the topics, articles at each time are modeled with latent Dirichlet allocation (LDA). Each article exhibits the topics with different random proportions θ_d; each word of each article is drawn by choosing a topic assignment from those proportions $z_{d,n}$, and choosing a word from the corresponding topic **?**.

Modeling multiple topics is important to the influence model because an article might have different impact in the different fields that it discusses. For example, an article about computational genomics may be very important to biology but less important to computer science. We want to discern its influence on each of these topics separately.

As with the DTM, we posit K topic trajectories, and each document of each time point is modeled with LDA. For each document, we now associate an influence score $\ell_{d,k}$ for each topic k. Each of the K topics drifts according to an adapted version of Equation **??**, where we restrict attention to the influence score for that topic and to the words of each document that were assigned to it,

$$\begin{aligned} &\beta_{k,t+1} \,|\, \beta_{k,t}, (w, \ell, z)_{t,1:D} \sim \\ &\qquad \mathcal{N}\left(\beta_{k,t} + \exp(-\beta_{k,t}) \textstyle\sum_d \ell_{d,k} \sum_n w_{d,n} z_{d,n,k}, \sigma^2 I\right). \end{aligned} \tag{4.4}$$

Here, $z_{d,n,k}$ is the indicator that the nth word in document d is assigned to topic k and we have dropped the index t on z and w. The graphical model is illustrated Figure **??**.

Although we presented our model in this section with influence spanning one year, we also adapted it to accomodate an "influence envelope", where an article's influence spans W years. This provides a more realistic model of influence **?**, but it complicates the inference algorithm and may not be necessary, as we note in section **??**.

To use this model, we analyze a corpus through posterior inference. This reveals a set of K changing topics and influence scores for each article and each topic. The posterior provides a thematic window into the corpus and can help identify which articles most contributed to the development of its themes.

Related work. There is a large literature on citation analysis and bibliometrics. See **?** for a review. Much work in this area uses the link structure of citation networks to extract higher level structure. **?** for example, have used author and citation networks to understand the evolution of ideas in the history of science. There has also been work that proposes features and models for predicting future citation counts. Successful features often include the publishing journal's impact factor, previous citations to last author, key terms, and number of authors **??**.

A number of algorithms include the text of the articles in their analysis. This work often models the information in citations by predicting them or modeling them with topics **????** or other semantic tools **??**. Other work in this area uses the text of documents along with citations to summarize documents **?** or to propose new bibliometrics: **?** use topic models and citations to map topics over time and define several new bibliometric measurements such as topic Impact Factor, topical diffusion, and topic longevity.

Our model has a different flavor from this research. We are interested in identifying the influential articles in a collection, but we do not assume that there are any notions of reference within them: there is no training data that contains citations. While we validate our model by looking at the relationship between our measure of influence and citation counts, our model is applicable to collections for which this kind of information does not exist.

Two important pieces of recent research have similar goals. **?** describe a framework for tracking the spread of memes, or ideas, in document collections, and investigate the direction in which ideas tend to percolate. **?** describe a measure of influence by modeling documents as unigram mixtures of earlier documents and use a likelihood ratio test to predict citations between documents. In contrast to this work, the DIM uses dynamic topics to explicitly model the change in *topic* language. Further, we do not attempt to model links between documents, as in **?**.

4.3 Inference and parameter estimation

Our computational challenge is to compute the posterior distribution of the latent variables—the sequences of topics and the per-document influence values—conditioned on an observed corpus. As for simpler topic models, this posterior is intractable to compute exactly. We employ variational methods to approximate it. Variational methods posit a simpler distribution over the latent variables with free parameters (called variational parameters). These parameters are fit to minimize the KL divergence between the variational distribution and the true posterior, which is equivalent to maximizing a lower bound on the marginal probability of the observations. See **?** for a review of this class of approximate inference methods.

We begin by specifying a variational distribution for the DIM posterior. First, the word assignments z_n and topic proportions θ_d are governed by multinomial parameters ϕ_d and Dirichlet parameters γ_d, as in LDA **?**.

The variational distribution for topic trajectories $\{\beta_{k,1}, \ldots, \beta_{k,T}\}$ is described by a linear Gaussian chain. It is governed by parameters $\{\tilde{\beta}_{k,1}, \ldots, \tilde{\beta}_{k,T}\}$, which are interpreted as the "variational

observations" of the chain. These induce a sequence of means $\tilde{m}_t$ and variances $\tilde{V}_t$. **?** call this a "variational Kalman filter."

Finally, the variational distribution of the document influence value $\ell_{d,k}$ is a Gaussian with mean $\tilde{\ell}_{d,k}$ and fixed variance σ_ℓ^2.

The variational distribution is

$$
\begin{aligned}
&q(\beta, \ell, z, \theta | \tilde{\beta}, \tilde{\ell}, \phi, \gamma) = \\
&\quad \textstyle\prod_{k=1}^{K} q(\beta_{k,1:T} | \tilde{\beta}_{k,1:T}) \\
&\quad \textstyle\prod_{t=1}^{T} \prod_{d=1}^{D_t} q(\theta_{t,d} | \gamma_{t,d}) q(\ell_d | \tilde{\ell}_d) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n} | \phi_{t,d,n}).
\end{aligned}
$$

Using this variational family, our goal is to maximize the following lower bound on the model evidence of the observed words $\mathbf{W}$:

$$
\begin{aligned}
\ln p(\mathbf{W}) \geq & \textstyle\sum_T \mathbb{E}_q \left[\ln p(\beta_t | \beta_{t-1}) \right] \\
& + \textstyle\sum_T \sum_{D_t} \mathbb{E}_q \left[\ln p(\ell_d) \right] + \mathbb{E}_q \left[\ln p(\theta_d | \alpha) \right] \\
& + \textstyle\sum_T \sum_{D_t} \sum_{N_d} \mathbb{E}_q \left[\ln p(z_n | \theta_d) \right] + \mathbb{E}_q \left[\ln p(w_n | z_n, \beta_t) \right] \\
& + H(q).
\end{aligned}
$$

This bound is optimized by coordinate ascent, with the variational parameters optimized sequentially in blocks. These updates are repeated until the relative increase in the lower bound is below a threshold.

Topic trajectories. The variational update for $\tilde{\beta}$ is similar to that in **?**. For each topic, we update the variational Kalman "observations" by applying gradient ascent:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \tilde{\beta}_{sw}} = & -\tfrac{1}{\sigma^2}\textstyle\sum_{t=1}^{T}(\tilde{m}_{tw} - \tilde{m}_{t-1,w} - G_{t-1,w}) \\ & \qquad \times \left(\tfrac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}} - \tfrac{\partial \tilde{m}_{t-1,w}}{\partial \tilde{\beta}_{sw}} + G_{t-1,w}\tfrac{\partial \tilde{m}_{t-1,w}}{\partial \tilde{\beta}_{sw}}\right) \\ & + \textstyle\sum_T \left(N_{w,t} - N_t \zeta_t^{-1}\exp(\hat{m}_{\beta_{tw}} + \tfrac{\tilde{V}_{tw}}{2})\right)\tfrac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}} \\ & + \tfrac{1}{\sigma^2}\textstyle\sum_{t=1}^{T}\tfrac{\partial \tilde{m}_{t-1,w}}{\partial \tilde{\beta}_{sw}}\left(H_{t-1,w} - G_{t-1,w}^2\right) \\ & + \tfrac{1}{\sigma^2}\textstyle\sum_{t=0}^{T-1}\tfrac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}} G_{tw}\tilde{V}_{tw},\end{aligned}$$

where

$$\begin{aligned} G_{sn} &= \mathbb{E}_q\left[\exp(-\beta_{s,k,n})(\mathbf{W}_{s,k,n} \circ z_{s,k,n})\ell_{s,k}\right] \\ H_{sn} &= \mathbb{E}_q\left[\exp(-2\beta_{s,k,n})\left((\mathbf{W}_{s,k,n} \circ z_{s,k,n})\ell_{s,k}\right)^2\right]. \end{aligned}$$

We expand H_t in the supplementary materials. Note also the variational parameter ζ_t and the term $\frac{\partial \tilde{m}_{tn}}{\partial \tilde{\beta}_{sn}}$, both described in **?**. The former can be updated once per iteration with $\zeta_t \leftarrow \sum_w \exp(\tilde{m}_{t,n} + \tilde{V}_{t,n}/2)$. The latter can be derived from the variational Kalman filter updates (see the supplementary materials).

Influence values. In the DIM, changes in a topic's mean parameters are governed by a normal distribution. As a consequence of this choice, updates for the influence parameters $\tilde{\ell}_{t,k}$ solve a linear regression. In this regression, documents' words at time t explain the expected topic drift $\boldsymbol{\Delta_{\beta,t,k}} = (\beta_{t+1,k} - \beta_{t,k})$, where the contributions of each document's words are given by the design matrix $X = \text{Diag}\left(\exp(-\beta_{t,k})\right)(\boldsymbol{W}_{t,k} \circ \phi_{t,k})$. ($\text{Diag}\left(\vec{x}\right)$ refers to the matrix having the elements of $\vec{x}$ on its diagonal.)

The parameter updates for document influence $\tilde{\ell}_{t,k}$ are defined, for each time t and each topic k, by the variational normal equation

$$\tilde{\ell}_{t,k} \leftarrow \left(\tfrac{\sigma^2}{\sigma_d^2}\boldsymbol{I} + \mathbb{E}_q\left[X^T X\right]\right)^{-1}\mathbb{E}_q\left[X^T \boldsymbol{\Delta_{\beta,t,k}}\right].$$

The expectation $\mathbb{E}_q\left[X^T X\right]$ is a matrix with dimension $D_t \times D_t$. Its elements are

$$\begin{aligned}\mathbb{E}_q\left[X^T X\right]_{d,d'} = \quad & \textstyle\sum_n \exp(-2\tilde{m}_{t,k,n} + 2\tilde{V}_{t,k,n}) \\ & \times (w_{t,d,n} w_{t,d',n} \phi_{t,k,d,n} \phi_{t,k,d',n})\end{aligned}$$

when $d \neq d'$ and

$$\begin{aligned}\mathbb{E}_q\left[X^T X\right]_{d,d} = & \textstyle\sum_n \exp(-2\tilde{m}_{t,k,n} + 2\tilde{V}_{t,k,n}) \\ & \times (w_{t,d,n}^2 \phi_{t,k,d,n})\end{aligned}$$

otherwise. The expectation $\mathbb{E}_q\left[X^T \boldsymbol{\Delta}_{\boldsymbol{\beta},\boldsymbol{t},\boldsymbol{k}}\right]$ is a D_t-dimensional matrix with elements

$$\begin{aligned}\mathbb{E}_q\left[X^T \boldsymbol{\Delta}_{\boldsymbol{\beta},\boldsymbol{t},\boldsymbol{k}}\right]_d = & \textstyle\sum_n w_{t,d,n} \phi_{t,k,d,n} \\ & \times (\tilde{m}_{t+1,k,n} - \tilde{m}_{t,k,n} + \tilde{V}_{t,k,n}/2) \\ & \times \exp(-\tilde{m}_{t,k,n} + \tilde{V}_{t,k,n}/2).\end{aligned}$$

Topic proportions and topic assignments. Updates for the variational Dirichlet on the topic proportions $\theta_{d,k}$ have a closed-form solution, exactly as in LDA **?**; we omit details here.

The variational parameter for each word w_n's hidden topic z_n is the multinomial ϕ_n. We solve for $\phi_{n,k}$ by the closed-form updates

$$\begin{aligned}\log(\phi_{n,k}) \quad \leftarrow \quad & \Psi(\gamma_k) + \tilde{m}_{t,k,n} \\ & + \frac{1}{\sigma^2} w_t \tilde{\ell}_{d_n,k} \exp(-\tilde{m}_{t,k} + \tilde{V}_{t,k}/2)(\tilde{m}_{t+1,k} - \tilde{m}_{t,k} + \tilde{V}_{t,k}) \\ & - \frac{1}{\sigma^2} w_{t,n} \Big[\tilde{\ell}_{d_n,k} \exp(-2\tilde{m}_{t,k} + 2\tilde{V}_{t,k}) \\ & \qquad \times (\boldsymbol{W}_{t,n,\setminus d_n} \circ \phi_{t,n,k,\setminus d_n}) \tilde{\ell}_{t,k,\setminus d_n}\Big] \\ & - \frac{1}{\sigma^2} w_{t,n}^2 \exp(-2\tilde{m}_{t,k} + 2\tilde{V}_{t,k})(\tilde{\ell}_{d,n,k}^2 + \sigma_l^2)\end{aligned}$$

where Ψ is the digamma function. Solving the constrained optimization problem, this update is followed by normalization, $\phi_{w,k} \leftarrow \frac{\phi_{w,k}}{\sum_K \phi_{n,k}}$.

4.4 Empirical study

We studied three sequential corpora of scientific articles. For each corpus, we estimated and examined the posterior distributions of its articles' influence.

In this section, we demonstrate that the estimate of an article's influence is robustly correlated to the number of citations it received. While the DIM model is designed for corpora without citations—and, indeed, only the documents' text and dates are used in fitting the model—citations remain an established measure of influence. This study provides validation of the DIM as an exploratory tool of influential articles.

4.4.1 Data

The three corpora we analyzed were the *ACL Anthology*, *The Proceedings of the National Academy of Science*, and the journal *Nature*. For each corpus, we removed short documents, terms that occurred in too few documents, and terms that occurred in too many documents. We also removed terms whose statistics did not vary over the course of the collection, as such terms would not be useful for assessing change in language (a sample of such non-varying terms from *Nature* is "ordinarily", "shake", "centimetre", "traffic", and "themselves").

ACL Anthology. The *Association for Computational Linguistics Anthology* is a digital collection of publications about computational linguistics and natural language processing **?**. We analyzed a 50% sample from this anthology, spanning 1964 to 2002. Our sample contains 7,561 articles and 11,763 unique terms after preprocessing. For this corpus we used article citation counts from the *ACL Anthology Network* **?**.

PNAS. The *Proceedings of the National Academy of Sciences* is a leading, highly-cited, multidisciplinary scientific journal covering biological, physical, and social sciences. We sampled one seventh of the collection, spanning 1914 (when it was founded) to 2004. Our sample contains 12,145 articles and 14,504 distinct terms after preprocessing. We found citations using Google Scholar for 78% of this collection.

Nature. The journal *Nature* is the world's most highly cited interdisciplinary science journal **?** with content on a range of scientific fields. We analyzed a 10% sample from this corpus, spanning 1869 (when it was founded) to 2008. Our sample contains 34,418 articles and 6,125 distinct terms after preprocessing. We found citations using Google Scholar for 31% of these documents.

Inference for 10 topics on each corpus above took about 11 hours to converge on a desktop Intel 2.4GHz Core 2 Quad cpu. Our convergence criterion was met when the evidence lower bound

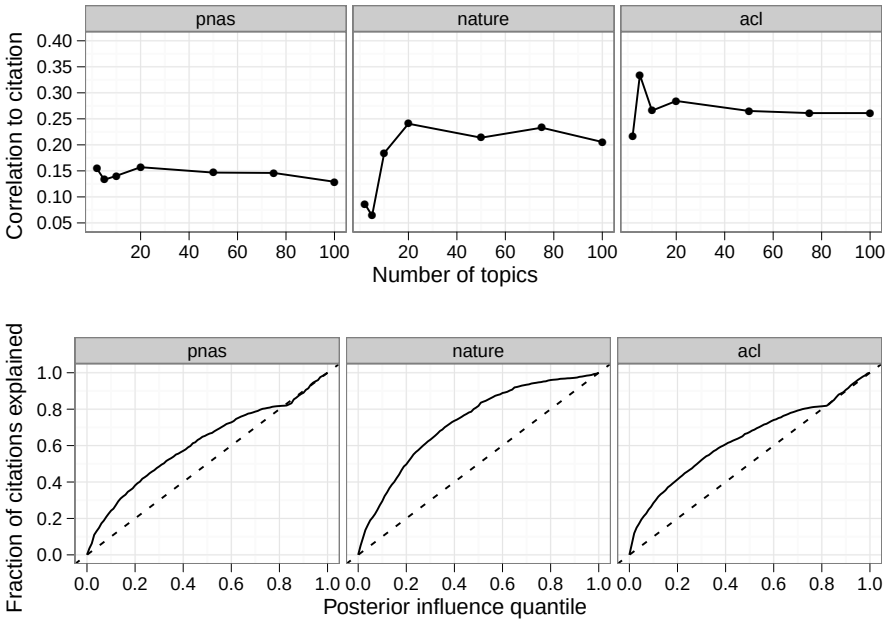


Figure 4.2: Spearman rank correlation between citation counts and posterior influence score, controlling for date (top) and fraction of citations explained by posterior influence (bottom).

increased by no more than 0.01%. For the experiments described below, we set topics' Markov chain variance $\sigma^2 = 0.005$ and $\sigma_d = \sigma_l = 0.0001$.

4.4.2 Relating posterior influence and citation

We studied the DIM with varying numbers of topics. We measured the relationship between the posterior influence values of each article $\tilde{\ell}_d$ and its citation count c_d.

We first aggregate the influence values across topics. Recall that each document has an influence value for each topic. For each word, we compute its expected posterior influence score, with the expectation taken with respect to its (random) topic assignment. We then sum these values over all words in the document,

$$f(\tilde{\ell}_d) = \sum_{n=1}^{N_d} \mathrm{E}[z_{d,n} \cdot \tilde{\ell}_d]. \tag{4.5}$$

This weights each word by the influence associated with its assigned topic. (Using the maximum value of influence across topics yielded similar results.)

Figure **??** displays the Spearman rank correlation between the aggregated posterior influence score of Equation **??** and citation counts. The DIM posterior—which is estimated only from the texts of the articles—has a positive correlation to the number of citations. All of these numbers were found significant up to $p < 1e-4$, using permutation tests on the influence scores.

Correlation goes up when we model multiple topics within a corpus. Moving from 2 to 5 topics in the *ACL* corpus increases correlation from 0.25 to 0.37. *Nature* is likewise better with more topics,

with a correlation of 0.28 at 20 topics; while *PNAS* performs best near 5 topics, with a correlation of 0.20.

Figure **??** also shows the fraction of citations explained by DIM scores: *Nature* documents with the highest 20% of posterior influence, for example, received 56% of citations. The flat regions in *ACL* and *PNAS* are due to aggregate influence scores very close to zero.

Heuristic model. The DIM is a complicated model. To justify its complexity, we describe a simple baseline (the *heuristic*) which captures our intuition with a single topic, is easy to implement, and runs quickly. For this heuristic, we define a word's weight at time t as:

$$w_t := \frac{\text{Frequency of } w \text{ in } [t,t+f]}{\text{Frequency of } w \text{ in } [t-p,t]},$$

for fixed distances f into the future and p into the past. A document's score is the weighted average of its words' weights. This heuristic captures the intuition that influential documents use language adopted by other documents.

The heuristic performed best with large values of its parameters ($f = p = 200$). With these settings, it achieves a correlation of 0.20 for the *ACL*, 0.20 for *PNAS*, and 0.26 for *Nature*. For *Nature*, the model is more correlated with citations than the heuristic for 20, 50, and 75 topics. Correlation is matched for *PNAS*, the model slightly beating the heuristic at 5 topics. *ACL* outperforms the heuristic for all numbers of topics.

Shuffled corpus Though we have eliminated date as a confounder by controlling for it in correlations, there may be other confounders such as document length or topic distribution. We therefore measured the DIM's relationship to citations when dates were randomly shuffled, keeping all documents which share a date together. If non-date confounders exist, then we might see correlation in the shuffled data, marking observed correlation as dubious.

We shuffled dates in the corpora and refit the DIM. We found a *maximum* date-controlled correlation of 0.018 for 29 shuffles of *ACL*; 0.001 for 5 shuffles of *Nature*; and 0.012 for 28 shuffles of *PNAS*. While this shuffled experiment and controlling for date do not entirely preclude confounding, they eliminate many potential confounders.

4.4.3 A closer look

Experiments showing correlation with citations demonstrate consistency with existing bibliometrics. However, the DIM also finds qualitatively different articles than citation counts. In this section we

describe several documents to give the reader an intuition behind the kind of analysis that the DIM provides.

IBM Model 3 The second-most cited article in the *ACL Anthology Network* is *The Mathematics of Statistical Machine Translation: Parameter Estimation* **?**. It has 450 intra-*ACL* citations and 2,130 total citations listed on Google Scholar. This seminal work describes parameter estimation for five word-based statistical models of machine translation; it provided widely accepted statistical models for word alignment and introduced the well-known "IBM models" for machine translation. The posterior influence score for **?** ranked 6 out of 7,561 articles in a 10-topic model.

This article was most influential in a topic about translation, which had a trend toward "alignment for machine translation." The largest-moving words are shown in Figure **??** (left). Upward trends for "alignment", "brown", and "equation" are evident (although it is not clear whether "brown" refers to the author or the corpus).

The Penn Treebank The most-cited article in our subset of the *ACL Anthology Network* is *Building a large annotated corpus of English: the Penn Treebank* **?**, with 1,622 *ACL* citations and 2,810 citations on Google Scholar. This article describes the large-scale part-of-speech and syntax tagging of a 4.5-million word corpus. It falls in a topic about part-of-speech tagging and syntax trees; "treebank" had become one of the top words in the topic by 2004.

The DIM assigned a relatively low influence score to this article, ranking it 2,569 out of 7,561 articles. While **?** introduces a powerful *resource*, most of the article uses conventional language and ideas to detail the annotation of the Penn Treebank. As such, the paper does not discuss paradigm-changing ideas and the model scores it low. We emphasize that this does not undermine the tremendous influence that the Penn Treebank has had on the field of natural language processing. The DIM is not designed to discover this kind of influence.

Success in 1972 In 1967, The College Science Improvement Program was established to assist predominantly undergraduate institutions. Two years later *Nature* published a short column, which has the highest of our posterior influence in a 20-topic model, out of 34,418 *Nature* articles. No citation information was available about this article in Google Scholar. The column, *How to be Overtaken by Success*, discusses a debate about the "Miller bill", which considers funding for post-graduate education **?**. *Overtaken by Success* provides few research resources to researchers, which may explain lack of citation information. Instead, it presciently discusses a paradigm shift in a topic about science, industry, research, and education: "The record of the hearings [on the bill] is not

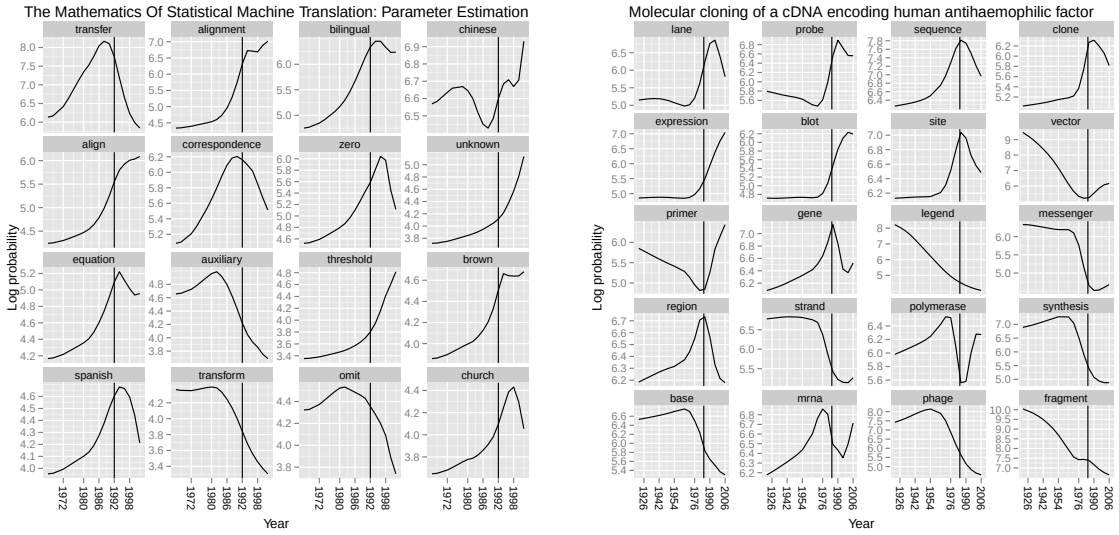


Figure 4.3: Most active words appearing in **?** (left) which have changed the most in a topic about translation. On right are words appearing in **?** in a topic about DNA and genetics. Terms are sorted by increase over 10 years.

merely an indication of the way the wind is blowing but an important guide to some of the strains which are now accumulating within the system of higher education..."

In 1972, three years after this article's publication, The NSF Authorization Act of 1973 made the NSF explicitly responsible for science education programs *at all levels* **?**. Where this may have been missed by those using citation counts to study the history of science education, the DIM has provided a metric with which to gauge interest in the article.

Genetics in *Nature* The sixth most influential document by the DIM in a 20-topic model of *Nature* is *Molecular cloning of a cDNA encoding human antihaemophilic factor*, an article describing successful cloning of a human mRNA sequence important in blood clotting **?**. With 584 citations, this article is among the top 0.2% of these 34,418 documents. The most active words appearing in this article are shown in Figure **??** (right). The plot shows some of the document's key words – "expression", "primer", "blot" – become prominent words in the topic.

4.5 A parallel implementation of the model

The algorithm described in Section **??** takes approximately 11 hours on a modern desktop computer [1], for about 30,000 documents. For a larger dataset – such as all scientific articles in *Nature*, *Science*, and *PNAS* combined, it takes considerably longer to complete. In this section, we describe a parallel

[1] This was a 2.2GHz, 1MB cache, Dual core AMD Opteron 275 processor

algorithm for this model. As with the standard algorithm, this one optimizes the evidence lower bound by local coordinate ascent. Here, however, many of these steps are made in parallel.

4.5.1 Algorithm overview

With both the parallel implementation and the standard implementation, we initialize the model with LDA topics. The parallel implementation distributes the work of the E step among the many workers during each iteration. The LDA M step for each iteration (which simply aggregates sufficient statistics) is then run on a single processor.

Following this initialization, the actual model is fit. This is driven by a single master program which alternates between two steps: a *topic* M-step and a *document* E-step.

4.5.2 The topics M-step

In the topic step of the original algorithm, topics are re-fit by adjusting the "variational observations" using documents' parameters. The parallel algorithm performs the same operations in parallel, simply splitting the work among N workers. After this, the likelihood of each topic chain is printed to a sentinel file to alert the master that the task is complete.

4.5.3 The documents E-step

In the documents step of the algorithm, our goal is to re-fit each document using topics. As with the original algorithm, we re-fit γ_d using alternating updates of ϕ_{dn} and γ_d. We also update the influence I_d of all documents at each time $t = 1 \dots T$, as before.

We perform inference of I_d by distributing the D documents among W workers W sequentially contiguous blocks

$$d_{n_1} \dots d_{n_2} \qquad d_{n_2+1} \dots d_{n_3} \qquad \dots \qquad d_{n_W+1} \dots d_{n_{W+1}},$$

for $0 = n_1 < \dots n_i < \dots < n_{W+1} = D$, and performing the E-step on each worker $1 \dots W$ in parallel.

This update is problematic, however, because the influence of documents at *different* times interact. A document d_j handled by worker W_i and a document d_k handled by worker $W_i + 1$ might both influence the same topic, for example: we cannot therefore simply update I_{d_j} and I_{d_k} independently because doing so would over-explain a topic's influence. This could lead to instability, as documents attempt to explain ever-diverging topics and topics must adjust to ever-divergent

influence. We address this with **update dampening**: on each iteration s, the influence of any document d is the geometric average of the exact update $I_{d,s,D_{s-1}\beta_{s-1}}$ and its previous value $I'_{d,s-1}$:

$$I'_{d,s} \leftarrow \frac{W-1}{W} I'_{d,s-1} + \frac{1}{W} I_{d,s,D_{s-1},\beta_{s-1}},$$

or, equivalently,

$$I'_{d,s} \leftarrow I'_{d,s-1} + \frac{1}{W}(I_{d,s,D_{s-1},\beta_{s-1}} - I'_{d,s-1})$$

This update has two nice properties:

1. The influence of each document is no more than 1/W closer to the total influence than it was before; this prevents global fluctuations.

2. Posterior modes are stationary points.

4.6 Influence in all of science

Chapter 5

Models of Spatial Voting and text

Introduction: Legislative voting in Western Democracies

5.1 The role of text in lawmaking bodies

5.2 The Ideal Point Model

5.3 A model for predicting votes with the text of new bills

5.3.1 An empirical analysis

5.4 A study of lawmakers' voting on specific issues

5.4.1 A model of exceptional voting patterns

5.4.2 An empirical analysis of voting patterns

Conclusions

Chapter 6

The influence of Judges' opinions in higher courts

6.1 Introduction: Decisions and influence in the higher courts

6.1.1 Influence and the progress of ideas in the higher courts

6.1.2 The Cardozo Topic

6.2 Influence of Judges' dissenting opinions

6.2.1 A model of influence among opinions

6.2.2 A posterior-predictive checks among dissenting opinions

6.2.3 Do lower court opinions influence higher courts?

Conclusions

Chapter 7

A time-series model of foreign affairs: sentiment between nation-states

Introduction

7.1 Foreign relations and their influence on news reporting

7.2 A latent-space model of foreign relations

7.2.1 Posterior inference

7.2.2 Supervised sentiment analysis

- Introduce Amazon Mechanical Turk - Describe supervised relationships and the model of sentiment from

7.2.3 Unsupervised sentiment analysis

- Describe relational topic model

7.3 Empirical studies: comparisons with ground truth

7.3.1 Comparisons with Ground Truth

7.3.2 Relation to ideal points inferred from the UN General Assembly

Conclusions

Chapter 8

Appendix

8.1 Processing text data

8.1.1 Stemming and parsing

8.1.2 Current libraries and software

8.2 Model estimation basics

8.2.1 Multimodal distributions and identification

8.2.2 Setup of a typical program

8.2.3 The role of traditional dimensionality reduction in evaluating underconstrained distributions

8.3 Stochastic optimization

8.3.1 Round-robin updates

8.3.2 Bounded step size