Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (B)

Extending Information Retrieval Methods to Personalized Genomic-Based Studies of Disease

Shuyun Ye[1], John A. Dawson[1] and Christina Kendziorski[2]

[1]Department of Statistics. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA.

ABSTRACT: Genomic-based studies of disease now involve diverse types of data collected on large groups of patients. A major challenge facing statistical scientists is how best to combine the data, extract important features, and comprehensively characterize the ways in which they affect an individual's disease course and likelihood of response to treatment. We have developed a survival-supervised latent Dirichlet allocation (survLDA) modeling framework to address these challenges. Latent Dirichlet allocation (LDA) models have proven extremely effective at identifying themes common across large collections of text, but applications to genomics have been limited. Our framework extends LDA to the genome by considering each patient as a "document" with "text" detailing his/her clinical events and genomic state. We then further extend the framework to allow for supervision by a time-to-event response. The model enables the efficient identification of collections of clinical and genomic features that co-occur within patient subgroups, and then characterizes each patient by those features. An application of survLDA to The Cancer Genome Atlas ovarian project identifies informative patient subgroups showing differential response to treatment, and validation in an independent cohort demonstrates the potential for patient-specific inference.

KEYWORDS: latent Dirichlet allocation, time-to-event, survival, cancer, genomics

SUPPLEMENT: Array Platform Modeling and Analysis (B)

CITATION: Ye et al. Extending Information Retrieval Methods to Personalized Genomic-Based Studies of Disease. *Cancer Informatics* 2014:13(S7) 85–95 doi: 10.4137/CIN.S16354.

RECEIVED: August 7, 2014. **RESUBMITTED:** October 22, 2014. **ACCEPTED FOR PUBLICATION:** October 23, 2014.

ACADEMIC EDITOR: J T Efird

TYPE: Original Research

FUNDING: This work was supported by the National Institutes of Health (GM 102756). JAD receives funding through a pre-doctoral training grant from NIGMS, T32GM74904. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: kendzior@biostat.wisc.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

Technological advances continue to increase both the ease and accuracy with which measurements of the genome and phenome can be obtained and, consequently, genomic-based studies of diseases such as cancer often involve highly diverse types of data collected on large groups of patients. The primary goals of such studies involve identifying genomic features useful for characterizing patient subgroups as well as predicting patient-specific disease course and/or likelihood of response to treatment. Doing so requires computational methods that handle complex interactions, accommodate genetic heterogeneity, and allow for data integration across multiple sources.

A number of statistical methods are available for feature identification and prediction of a time-to-event phenotype such as overall survival or time to recurrence (for a review, see Chen et al.[1], Li and Li,[2] and Wei and Li[3]). Most often, classical models for a survival response are coupled with some dimension reduction methods for individual[4–6] or grouped predictors,[1,2,7,8] providing a concise representation of the genomic features affecting patient outcome. Although useful, the majority of these methods identify a set of covariates common to all patients and as a result may "distort what is observed" in the presence of heterogeneity.[9] Survival-supervised clustering approaches naturally accommodate heterogeneity, providing for efficient and effective identification of patient subgroups.[10,11] However, these approaches do not identify salient features associated with subgroups and, as with the aforementioned methods, may

sacrifice power and accuracy by focusing on one (or a few) data set(s) in isolation.

Latent Dirichlet allocation (LDA)[12] models are particularly well tailored for accommodating heterogeneity, selecting features, and characterizing complex interactions in a high-dimensional textual setting, but their application in genomics has been limited. By far, the most common application concerns identifying groups of words that co-occur frequently (topics) across large collections of text (eg, a collection of articles or abstracts). The derived topics provide insights into the collections' content overall as well as into the specific content within a document, and estimated document-specific distributions over topics are useful in classifying new documents.[12–14]

An extension allows for topic estimation to be supervised by a response that is suitably described by a generalized linear model.[15] So-called supervised LDA (sLDA) debuted with a study of movie reviews (each considered a document) and estimated topics (collections of co-occurring words in a review) that determined the number of stars (supervising response) a movie received. Derived topics included ones having highest weight on words such as "power", "perfect", "fascinating", and "complex"; another with highest weight on "routine", "awful", "featuring", "dry"; a third on "unfortunately", "least", "flat", "dull"; and so on. The movie-review–specific distribution over topics proved useful in classifying movies. Those with highest weight on the "power" topic generally had a high number of stars while those with highest weight on the "unfortunately" topic had a low number; those with weight on the "routine" topic most often ended up in the middle. Differences between the distributions also provided insights into differences between movies that received a similar number of stars.

Our interest here is not in evaluating movies. However, it is important to note that the questions addressed in Blei and McAuliffe[15] are identical in structure to the most important questions we face in cancer genomics. In the former, questions include: "Given reviews and ratings for a group of movies, can we identify collections of words (topics) that discriminate movie reviews? Can each movie be described by a distribution over those topics? Can distributions over topics provide insights into differences between similarly rated movies? And can a movie-specific distribution over topics be used to predict what the rating of a new movie will be?" In cancer genomics, the questions include: "Given genomic, clinical, and survival information on a group of patients, can we identify collections of genomic and clinical features (topics) that define and discriminate among patient subgroups? Can a patient be well described by a distribution over those topics? Can distributions over topics provide insights into the genomic differences between two patients with similar survival? And can a patient-specific distribution over topics be used to predict survival of a new patient?"

To address these types of questions, we extend LDA for use in a clinical and genomic setting. Specifically, survival-supervised LDA (survLDA) is developed in the second section to facilitate topic supervision by a time-to-event response with censoring. Unlike in the textual domains of Blei et al.[12], Porteous et al.[13], and Biro et al.[14], the definition of a document is not obvious in this setting. The Methods section details the construction of documents, one for each patient, where words describe clinical events, treatment protocols, and genomic information from multiple sources. As we show in the Application of survLDA to the TCGA data section, application of survLDA to this collection of documents provides for the identification of topics useful in characterizing patient subpopulations as well as individual patients in a study of ovarian cancer conducted as part of The Cancer Genome Atlas (TCGA) project.[16] Classification of new patients is considered in the third section, and we conclude with the Discussion section.

Methods

The LDA model. We briefly review the LDA model as detailed in Blei et al.[12] Assume there are D documents indexed by $i = 1,..., D$, each of which consists of N_i words. The vocabulary is the unique set of length V indexed by $v = 1,..., V$, from which the documents' words arise and is usually taken to be the union of all words over documents. Further, assume that there are K latent "topics" indexed by $k = 1,..., K$, that govern the assignment of words to documents. Each topic corresponds to a discrete distribution over the V words in the vocabulary, with parameters given by the V-vector τ_k. Likewise, each document is assumed to be a mixture over the K topics with mixing coefficients θ_i (a K-vector parameter), indicating the proportion of words sampled from each topic.

For a given document i, N_i words arise from the following generative process, given the system-wide hyperparameters α (a K-vector Dirichlet parameter) and the $\tau_{1:K}$ (the topic V-vectors):

1. Draw topic proportions $\theta_i \sim \text{Dirichlet}(\alpha)$.
2. For each of the N_i words, indexed by j:
   a. Draw a topic assignment $Z_{ij}|\theta_i \sim \text{Multinomial}(1,\theta_i)$, where $Z_{ij} \in \{1, \ldots, K\}$.
   b. Draw a word $W_{ij}|Z_{ij}, \tau_{1:K} \sim \text{Multinomial}(1,\tau_{Z_{ij}})$, where $W_{ij} \in \{1, \ldots, V\}$.

With this model in place, a variational expectation-maximization (EM) algorithm may be used to estimate the joint posterior distribution of θ_i and $Z_{i,1:N_i}$, given $w_{i,1:N_i}$, α, and $\tau_{1:K}$ for each document i (expectation step [E-step]) and then to estimate the system-wide hyperparameters α and $\tau_{1:K}$ (maximization step [M-step]). Upon convergence, the variational EM yields optimal values for the key quantities of interest, namely posterior estimates of the topics ($\tau_{1:K}$) and document-specific distributions over topics ($\theta_{1:D}$). An extension of LDA by Blei and McAuliffe in 2008 allows for topic estimation to be supervised by a response that is suitably described by a generalized linear model. When time-to-event

responses such as survival times are of interest, sLDA is not directly applicable since it does not accommodate censoring.

The survLDA model. The survLDA model assumes the same setup as in Section 2, but allows for topics to be supervised by a time-to-event outcome. For document i, the survival outcome is denoted by T_i; an indicator variable for death/censoring is also observed for each document, denoted by δ_i. The survival response $T_i \mid \bar{Z}_i, \beta, h_0$ is described by a Cox proportional hazards model[17] with hazard function $h(t \mid \bar{Z}_i) = h_0(t) \exp\{\beta' \bar{Z}_i\}$, where $\bar{Z}_i$ is a K-vector with components $\bar{Z}_{ik} = \#\{Z_{ij} = k\} / N_i$. In this Cox proportional hazards model, each regression coefficient β_k exhibits the beneficent (negative) or deleterious (positive) effect of topic k on survival. We use a Weibull model for h_0, noting that alternative specifications (eg, nonparametric)[18] may be used. The system-wide model parameters for the survLDA model include a K-vector Dirichlet parameter α and the topic V-vectors $\tau_{1:K}$, just as in the LDA model described above. Specific to survLDA are survival response parameters β (a K-vector of regression coefficients) and $h_0(\cdot)$ (the baseline hazard). As in LDA, a variational EM algorithm is used to estimate the joint posterior distribution of θ_i and $Z_{i,1:N_i}$ given $w_{i,1:N_i}$, T_i, δ_i, α, $\tau_{1:K}$, β, and h_0 for each document i (E-step) and then to estimate the system-wide hyperparameters α, $\tau_{1:K}$, β, and h_0 (M-step). The derivation is given in the Appendix.

Document construction in the TCGA cohort. Unlike in the textual domains of Blei et al.[12], Porteous et al.[13], and Biro et al.[14] or in the movie review example described above, the definition of document is not obvious in this setting. To push the review analogy a bit further, whereas a movie review describes what is going on in a movie and provides an opinion on how the events were conveyed overall, we imagine patient reviews that describe what is going on in a patient with respect to genomic and clinical features. The analogy breaks down there, as the patient review does not contain an opinion on whether the features are positive or negative overall. Rather, the survLDA model is used to identify important features and estimate how these features relate to patient outcome as summarized by a time-to-event phenotype such as survival.

We use data from the TCGA ovarian project to construct patient-specific reviews or documents that summarize clinical and genomic features. For each of 511 patients in the TCGA ovarian cohort, clinical information such as age at diagnosis, date of surgery, surgical outcome, adjuvant therapies, time to recurrence, treatment at recurrence, overall survival, and dozens of other variables are available. Also available are high-throughput measurements of gene expression, methylation, single-nucleotide polymorphism (SNP)/copy number variation (CNV)s, and microRNAs.

For document construction, we use words associated with drugs, gene expression, and methylation, noting that other data sources could be integrated in a similar way. Specifically, the vocabulary (the union of words across all documents) includes words associated with commonly administered drugs (platinum, taxol, doxorubicin, topotecan, and gemcitabine) as well as words derived from potentially relevant genes. For gene words, we consider the 991 genes from the 12 cancer-related pathways defined in Jones et al.[19], since studies suggest that the vast majority of cancer-causing mutations lie in genes within these pathways. We also include the 5000 genes having mRNA expression that is most correlated with overall survival in the TCGA cohort as well as the 5000 having methylation that is most correlated. Given the considerable overlap between these lists, the two combined give 7452 unique genes for a total of 7897 genes from which words are derived.

Ideally, a patient's document will provide a comprehensive description of his/her clinical and genomic state. Toward this end, a patient's document received a drug word for each drug the patient received and a gene word for gene 'X' if the patient showed aberrant expression for that gene. To determine the direction of aberrant expression, we considered the association between gene expression and survival time. If increased expression was associated with decreased survival time for gene X, then any patient with expression in the uppermost 10th percentile for that gene received a gene word. Similarly, if decreased expression was associated with decreased survival time, then any patient with expression in the lowest 10th percentile received a gene word. The same procedure was applied to methylation data. Once all documents were constructed, the term-frequency inverse-document frequency ($tf - idf$) statistic was applied to identify words with discriminating power, as is commonly done in LDA applications.[20] Term-frequency, $tf(t,d)$, is the normalized frequency of a word t in a document d: specifically, $tf(t,d) = \frac{f(t,d)}{\max(f(w,d) : w \in d)}$ where $f(t,d)$ is the frequency of the term t in document d, and $\max f(w,d)$: $w \in d$ is the maximum frequency over all terms in the document. The inverse-document frequency is given by $idf(t,D) = \log \frac{|D|}{|d \in D : t \in d|}$ where $|D|$ is the number of documents in the corpus and $|d \in D: t \in d|$ is the number of documents in the corpus for which t appears. The idf of a rare term is high, whereas the idf of a common term is low. The $tf - idf$ statistic $tf - idf(t,d,D) = tf(t,d) * idf(t,D)$ combines these two measures[21] and is used here to identify terms that are relatively rare across documents (ie, discriminating), but relatively common within some sub-collection of documents.

To provide further detail, if a word shows up exactly the same number of times across all the documents (eg, the word "the" shows up 10 times in each of all the documents we have), then the $tf - idf$ value of this word in every document will be 0 since idf will be zero (since the document frequency is the proportion of documents that contain this word [in this example, 1] and idf is log of the inverse-document frequency [here, 0]). On the other hand, a word's $tf - idf$ value in a document will be higher if it shows up in some documents but not others. For example, if we have a word that appears in 10% of the documents, then $tf - idf$ is 2.3 (assuming documents of equal length since then tf is 1 and idf is log(10)). In a TCGA cohort,

words with $tf-idf \geq 0.25$ were retained in the final collection of documents following the study by Horacek et al (2010).[22]

Prediction. Given a new patient with clinical and genomic data, it may be of interest to construct a document $w_{1:N}$ and use it to predict survival. With a fitted model $\{\alpha, \tau_{1:K}\}$, the posterior mean $\bar{Z}_{new} = \bar{Z} \mid w_{1:N}, \alpha, \tau_{1:K}$ can be obtained in order to estimate from which topics this new patient draws words and in what proportions. As was the case during model fitting, this posterior must be approximated via variational inference. We do so by following the same procedure as outlined in the first subsection of the Appendix, except that all survival-related terms in the evidence lower bound are dropped; see the Prediction section in the Appendix for details.

Given $\bar{Z}_{new}$ measures related to topic membership can be predicted for the new patient. This may be done qualitatively (eg, "This patient is predicted to belong strongly to the first topic and survival for that topic is poor, hence her prognosis is bad.") or quantitatively (eg, predicting median survival time using the parametric survival model).

Application of survLDA to the TCGA Data

Given documents constructed as described above for each of the 511 women are considered, we applied survLDA. The supervising outcome of interest is all-cause mortality; and in all analyses, we used $K = 7$ topics, the last being the background topic. Application of survLDA provides two quantities of primary interest. The topics $\tau_{1:K}$ or estimated distributions over words identify clinical and genomic features that co-occur frequently in some groups of patients, but less frequently in others; and the document-specific distributions over topics $\theta_{1:D}$ characterize individual patients by specifying the proportions of their features coming from each topic. Of interest is determining the salient features in patient-specific documents that are represented by these topics and ultimately how the topics relate to overall survival.

Results

The left panel of Figure 1 shows a heat map with patients (columns) clustered according to topic membership for the six nonbackground topics (rows). The proportion of a patient's document words coming from a topic ranges from near 0 (almost no words, deep blue) to near 1 (virtually all words, red). As shown, most patient documents have the majority of words coming from a single topic, while some are best described by mixtures over topics. To see how differences among topics translates to differences in overall survival, the right panel of Figure 1 shows Kaplan–Meier curves for TCGA patients grouped by topic membership. Specifically, each patient is assigned to the topic having the highest weight in her document, as estimated by $\theta_{1:D}$. Patients with documents having highest weight on topic 1, for example, show dramatically reduced survival (44% at 1.5 years), whereas patient documents best described by topic 2 show average survival (76% at 1.5 years). A closer look at the words underlying these topics provides some insight into the differences and identifies features that may be worthy of further investigation.

The left panel of Figure 2 presents the topic-specific distributions over words for each topic. Red (blue) indicates an overabundance (dearth) of a word's weight in the corpus belonging to a particular topic. The right panel of Figure 2 shows a close-up view, highlighting 40 high-weight words that in part differentiate topics 1 and 2. A number of the results observed are consistent with prior studies. For example, CD163 expression levels have recently been shown to be prognostic of outcome in ovarian cancer patients, with higher expression associated with poor outcome.[23,24] This is consistent with what we observe, with an abundance of CD163 words in the poor outcome (topic 1) group. Similarly, increased expression of IGF2 has also been associated with poor survival in ovarian cancer patients.[25] Here we observe high methylation of IGF2AS (which is correlated with IGF2)[26] in the poor outcome group,

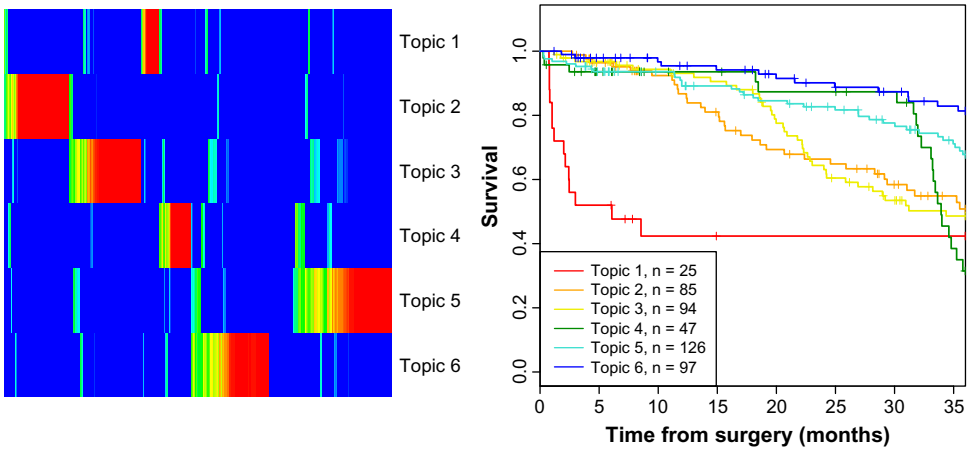

Figure 1. The left panel shows a heat map of the estimated patient-specific distributions over topics (θ) for each of 511 patients (the background topic is not shown). Topics are given in the rows; patients are clustered along the columns. Colors range from deep blue (topic underrepresented in the patient's document) to red (topic overrepresented). The right panel shows Kaplan–Meier survival curves for patients classified into one of the six nonbackground topics. Each patient was assigned to the topic having highest weight in his/her document, as estimated by $\theta_{1:D}$.

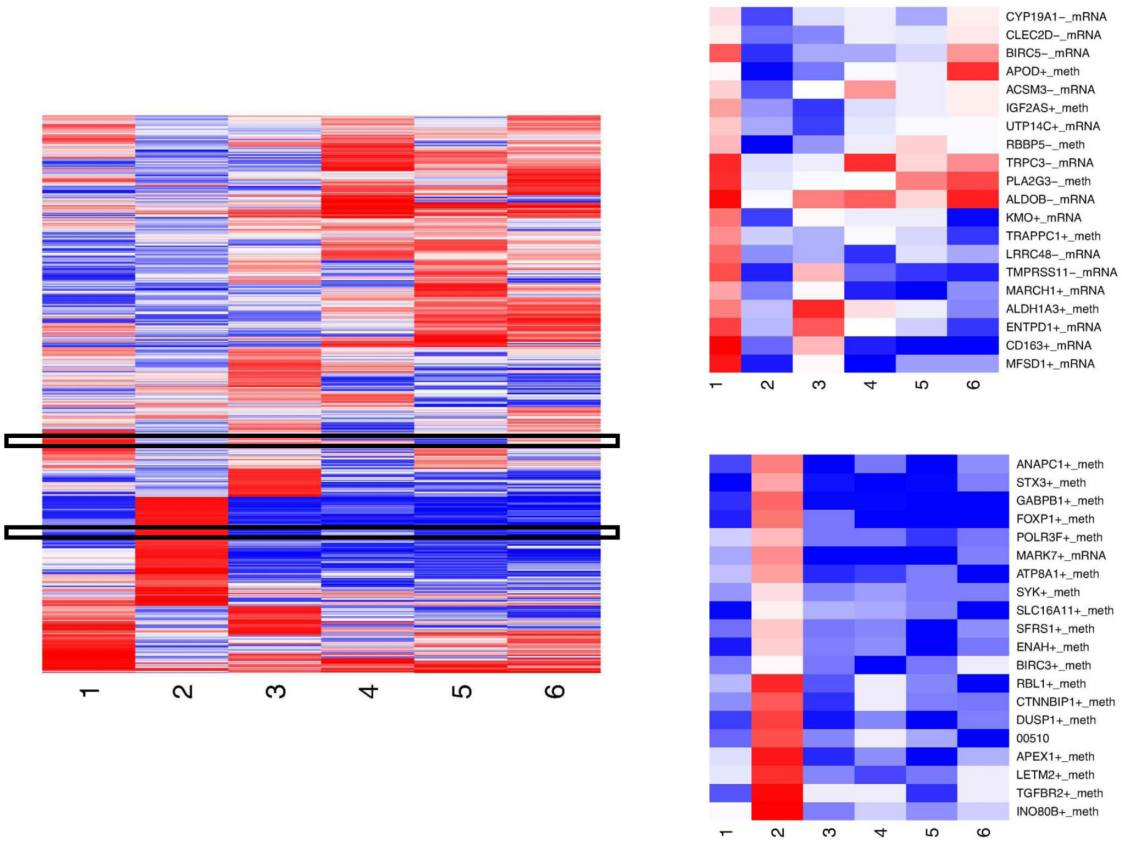

Figure 2. The left panel shows a heat map of the topics derived from survLDA. Topics are shown in the columns; words are clustered along the rows. The colors range from blue (word underrepresented in the topic) to red (word overrepresented), with white in the middle (average representation). To aid in interpretation, we add the risk direction and data source from which each word was derived. For example, *CYP19A1 – mRNA* indicates that underexpression of CYP19A1 is associated with increased risk and that CYP19A1 words were entered into a document for patients with underexpression of CYP19A1. As the heat map shows, there are many words that distinguish topics 1 and 2, having high weight in one topic but not the other. The insets highlight 40 such words; those having high weight in topic 1 (topic 2) are shown in the upper (lower) right.

which at first may seem to be a contradiction given that increased methylation often results in decreased expression. However, that is not the case for IGF2, where increased methylation correlates with increased expression.[25]

Other genes such as TRPC3, ALDH1A3, and FOXP1 have been studied in other cancers; and our results suggest that these genes may play important roles in ovarian cancer as well. Underexpression of TRPC3 has been correlated with poor prognosis in lung cancer,[27,28] as has hypermethylation of ALDH1A3 in bladder cancer.[29] FOXP1 is a relatively well-known tumor suppressor gene with increased expression associated with improved outcomes among breast cancer patients.[30] As in these studies, we observe TRPC3 underexpression and ALDH1A3 hypermethylation in our poor prognosis group and increased FOXP1 expression in patients with longer survival.

It is interesting to note that with the exception of CD163, these genes would not likely have been identified in this cohort using other approaches, as the marginal *P*-values from a Cox proportional hazards test are far from overwhelming (CD163 $P = 0.013$, IGF2AS $P = 0.631$, ALDH1A3 $P = 0.951$; FOXP1 $P = 0.188$; TRPC3 $P = 0.282$), indicating that although there are differences in the expression and/or methylation of these genes between patients primarily described by topics 1 and 2, those differences are obscured by heterogeneity in the full cohort. Although further investigation of these and other genes that display markedly different abundance patterns between patient subtypes might improve our understanding of the mechanisms that underlie differences between the groups, we note that a main advantage of LDA models in general and survLDA in particular is that topics describe co-occurrence of groups of words, not just occurrence of high-frequency words. The left panel of Figure 3 is a co-occurrence heat map showing the percentage of topic 1 patients having a given pair of words in their document. It is clear that the majority of topic 1 patients show high co-occurrence of topic 1 words and low co-occurrence of topic 2 words. The same holds true of patients best described by topic 2 words (right panel). Consequently, characterization of patient subtypes by these collections of genes taken together may prove to be more informative than characterization by individual genes. To further investigate whether these gene groups are meaningful, in the following section we evaluate their prognostic utility in independent patient populations.

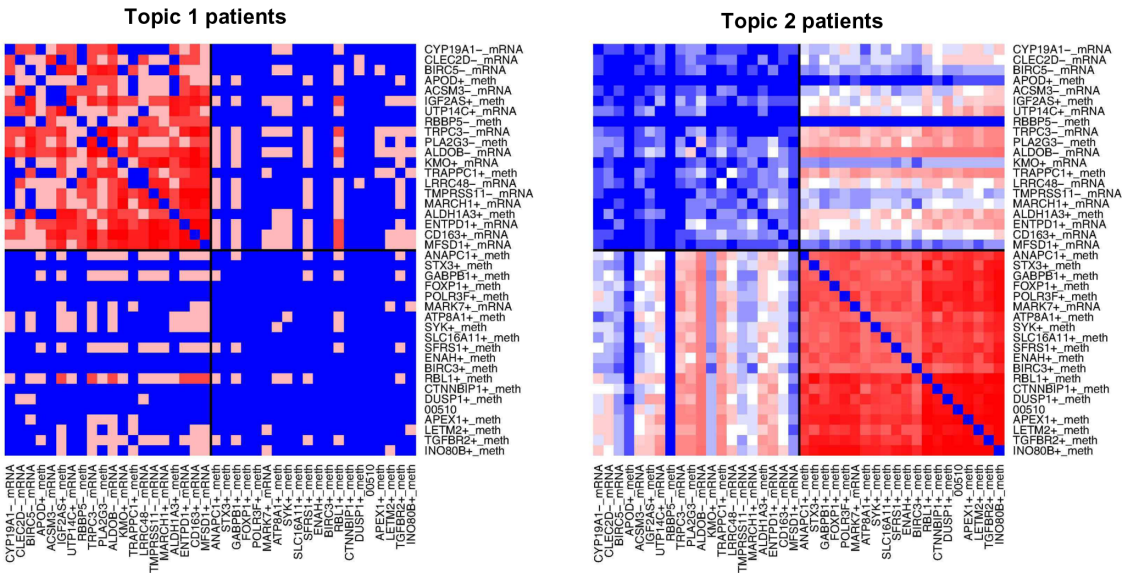

Figure 3. Heat maps showing co-occurrence of the 40 high-weight topic 1 and topic 2 words shown in Figure 2. The left heat map considers the 25 patients having documents with highest weight on topic 1. Shown are the percentages of those patients having both words in their document, ranging from 0 (blue) to 100% (red). The black line separates topic 1 and topic 2 words. The right panel is similar, showing percentages of co-occurrence in documents of the 85 patients best described by topic 2 words.

Prediction on independent data sets. To evaluate survLDA for patient-specific prediction, we consider two independent data sets. Specifically, we consider 240 patients from the study by Tothill et al.[31], conducted in Australia consisting of patients with ovarian, tubal, and peritonial cancers; we also consider 260 patients from the study by Yoshihara et al.[32] conducted in Japan. These independent populations are referred to hereinafter as the validation patients. Although the TCGA data we have used is restricted to patients with stage III or IV serous ovarian adenocarcinomas, these independent studies are more heterogeneous and thus present a more challenging (and realistic) validation data set. Documents for the validation patients were derived as described in Section 2 with the quantile thresholds taken from the training set (TCGA) data. Identifying thresholds in the training data allows us to construct documents for validation patients one at a time, as would be required in any setting where patient-specific prediction was of interest. Drug and methylation words were not included as the validation data sets did not contain this information.

With documents in hand, the survLDA output was used to predict topic membership for patients in the validation set, using the prediction approach given in Section 2. The left panel of Figure 4 shows survival for patients best described by words from topics 1 and 2, the two topics discussed earlier. There is a significant difference between survival in the two groups ($P = 0.037$). As in the training set, those patients predicted to belong to topic 2 have better survival, on average, than those patients predicted to belong to topic 1. Although statistically significant, the difference in the survival curves is attenuated relative to that observed in the training set (73% vs. 93% at 1.5 years in the validation set; 44% vs. 76% in the training set). Of course, some decrease in performance is expected as one moves to an independent validation set. Here we also lose some predictive ability as the validation set does not contain information about methylation or treatment, and so our predictor was built using a single data source (expression). Nevertheless, the ability to recover at least some information regarding outcome suggests that the topics are biologically relevant.

Discussion

A problem pervasive in genomic-based studies of disease concerns taking large, diverse data sets collected on a cohort

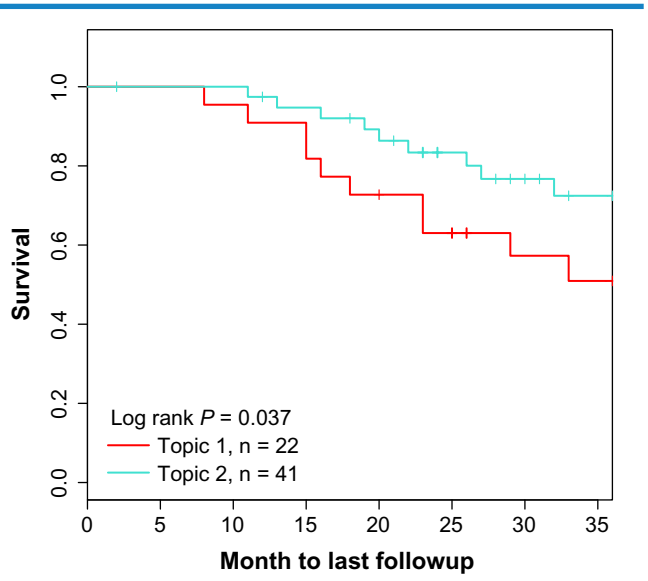

Figure 4. Topic-based prediction of overall survival in an independent patient cohort.

of patients and using the information contained therein to characterize patient subtypes as well as individuals. Computational scientists often address this problem by performing analysis within a single data type and comparing results subsequently in an effort to identify a signal supported by the disparate analyses (eg, a gene's SNPs, expression, and methylation all associate with a phenotype). Comparing results manually has its obvious disadvantages. At the same time, meta-analysis approaches such as Fisher's combined probability test can be limited by low power[33]; and efforts to combine data directly are challenged by measurements on different scales with differential dependencies. The survLDA-based framework proposed here addresses these challenges by transforming the information contained in high-throughput genomic screens into text. Doing so has both advantages and disadvantages.

One advantage is that data integration is seamless. In the implementation presented, a word for a gene is assigned to a patient's initial document if the gene shows extreme expression; the same word is assigned if the gene shows extreme methylation. In this way, a document may contain copies of words associated with extreme genomic features, measured from expression and/or methylation. Other types of data are easily incorporated into the framework. For example, just as with extreme expression or methylation, a gene word could be included in a document if that gene harbored a CNV.

A second advantage is that the threshold required for a gene to be included in the analysis is much lower than would be required with other methods. As detailed in the Methods section, some preselection of genes is done, but the selection does not require even nominally significant association with a survival endpoint, as is often required in survival studies with high-dimensional covariates.[4,7,34] This allows for the identification of many important genes, some previously known to be involved in cancers other than ovarian, which would not otherwise have been considered.

Although the identification of individual genes may prove useful, a main advantage of LDA in general and survLDA in particular is that it reveals *groups of genomic aberrations* that co-occur together (topics) and then characterizes individual patients by those groups. The topics themselves are useful in that they define collections of genes, methylations, or other covariates among which undiscovered interactions might occur, while the patient-specific distributions over topics give insights into the similarities and differences among patients that go beyond the information that can be gained from grouping by like outcome.

Our findings of the predictive ability of the approach are mixed. In our earlier work,[35] we conducted simulation studies to assess predictive ability under a variety of settings. As that work suggested for the sample size considered here, there is some ability for prediction, but improvements are expected with increases in the number of patients as well as improvements in document creation strategies. As detailed in,[35] sample sizes larger than those considered here are required to improve prediction significantly. In general, more work is needed to better understand the specific effects of sample size, document size, word frequencies, and replication, which are determined in part by the method used for document construction. Our approach to assign a word for any gene showing extreme expression or methylation was motivated by the study of Zilliox and Irizarry,[35] where the authors identify bimodal genes and, for each individual and each gene, assign a binary variable indicative of mode membership. The resulting gene expression 'barcode' for each patient proved useful in classifying patients into biologically meaningful groups[36] and the extrapolation of their approach proved to be an effective strategy here. Another possibility is to assign an increasing number of words in direct proportion with signal. For example, consider breaking a gene's expression into deciles, say, and assign 1–10 words for each document (eg, a value between the sixth and seventh deciles gets seven words). We did not favor this approach for two main reasons. First, the approach assumes linearity of expression and methylation, which is often not the case. Second, the approach results in documents having few unique words, which reduces specificity of topics as well as document-specific distributions over topics. Document construction continues to be explored, and improvements are expected to prove useful in a number of settings.

In addition to the means by which covariates are translated into words, there are many aspects of the proposed methods that require further development. In particular, survLDA assumes the simplest of Dirichlet priors on the distributions of topics over patients and therefore the documents are considered conditionally independent given α. While this is a reasonable assumption for the TCGA data set we considered, there are other realms where correlation among the documents could arise. For example, one could have multiple documents arising from the same subject, one for each time point or tissue; or, when integrating multiple cancer types, subjects with the same type of cancer would be expected to be more alike than subjects with differing cancer types. Adding such hierarchy has already been explored to some extent for traditional LDA,[37] presenting a starting point for future methodological work.

Similarly, the composition of the topics themselves is essentially free. Were it not for our imposition of a background topic, the topics would be completely unstructured a priori. As it is, $K - 1$ topics are still governed solely by the data. This need not be the case, as methods similar to those proposed for construction of a background topic (see the Appendix) could be extended. In particular, the Dirichlet prior could be modified directly or a set of restrictions could be imposed for each topic and groups of words so that certain words cannot appear together or may only appear together in certain topics.

In summary, it is becoming increasingly clear that studies aimed at solving the most challenging problems in cancer genomics involve highly diverse types of data collected

on large groups of patients. Many methods will prove useful. We expect that advantage will be gained from methods that are able to integrate data and account for cohort heterogeneity, allow supervision by outcomes of interest such as survival, provide for patient-specific inference, and facilitate prediction of unobserved outcomes. The proposed approach provides tools for these purposes in an effort to help ensure that maximal information is obtained from genomic-based studies of disease.

Acknowledgments

The authors wish to thank Michael Jordan for conversations that helped motivate this work and Michael Newton and Ning Leng for conversations that helped to improve the manuscript.

Author Contributions

CK conceived the model and application, and wrote much of the paper. JD implemented an initial version of the model, figured out the extension of sLDA to survival data, analyzed data, wrote some of the paper, and wrote the appendix. SY improved upon the initial implementation and conducted further analysis. All authors reviewed and approved the final manuscript.

REFERENCES

1. Chen X, Wang L, Ishwaran H. An integrative pathway-based clinical-genomic model for cancer survival prediction. *Stat Probab Lett*. 2010;80:1313–9.
2. Li L, Li H. Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*. 2004;20:3406–12.
3. Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*. 2007;8:265–84.
4. Li H, Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data in pacific symposium on biocomputing; 2003: 65–76.
5. Ghosh D, Yuan Z. Combining multiple models with survival data: the PHASE algorithm, technical report. Penn State University Department of Statistics: University Park, PA 2010.
6. Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes. *Bioinformatics*. 2010;26:250–8.
7. Chen X, Wang L. Integrating biological knowledge with gene expression profiles for survival prediction of cancer. *J Comput Biol*. 2009;16:265–78.
8. Ma S, Song X, Huang J. Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*. 2007;8:60.
9. Aalen OO. Heterogeneity in survival analysis. *Stat Med*. 1988;7:1121–37.
10. Dettling M, Bühlmann P. Supervised clustering of genes. *Genome biology*. 2002; 3(12):1–0069.
11. Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*. 2004;20(suppl 1):i208–15.
12. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993–1022.
13. Porteous I, Newman D, Ihler A, Asuncion A, Smyth P, Welling M. Fast collapsed Gibbs sampling for latent Dirichlet allocation in KDD '08. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA, 2008:569–77.
14. Biro I, Szabo J, Benczur A. Latent dirichlet allocation in web spam filtering. In: Castillo C, Chellapilla K, Fetterly D, eds. *AIR-Web*. Beijing: ACM International Conference Proceeding Series; 2008:29–32.
15. Blei DM, McAuliffe JD. Supervised topic models in *advances*. In: Platt JC, Koller D, Singer Y, Roweis S, eds. *Neural Information Processing Systems* 20. Cambridge, MA: MIT Press; 2008:121–8.
16. National Cancer Institute and National Human Genome Research Institute. The cancer genome atlas. 2011. Available from: http://cancergenome.nih.gov/2011.
17. Cox DR. Regression models and Life-tables. *J R Stat Soc Series B Stat Methodol*. 1972;34:187–220.
18. Cook TD, DeMets DL. Introduction to statistical methods for clinical trials. *Statistical Science*. US: Chapman and Hall/CRC; 2008:366–82.
19. Jones S, Zhang X, Parsons DW, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*. 2008;321:1801–6.
20. Hong L, Davison BD. Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics. Washington DC, USA, 2010, 80–8.
21. Salton G, Fox E, Wu H. Extended Boolean information retrieval. *Commun ACM*. 1983;26:1022–36.
22. Horacek et al. Natural Language Processing and Information Systems: 14th International Conference on Applications of Natural Language to Information Systems. Helmut Horacek, Elisabeth Metais, Rafael Munoz, Magdalena Wolska, editors. June 24–26, 2009. NLDB 2009, Saarbrücken, Germany.
23. No JH, Moon JM, Kim K, Kim YB. Prognostic significance of serum soluble CD163 level in patients with epithelial ovarian cancer. *Gynecol Obstet Invest*. 2013;75:263–7.
24. Lim R, Lappas M, Riley C, et al. Investigation of human cationic antimicrobial protein-18 (hCAP-18), lactoferrin and CD163 as potential biomarkers for ovarian cancer. *J Ovarian Res*. 2013;75:5.
25. Huang Z, Murphy SK. Increased intragenic IGF2 methylation is associated with repression of insulator activity and elevated expression in serous ovarian carcinoma. *Front Oncol*. 2013;3:131.
26. Vu TH, Chuyen NV, Li T, Hoffman AR. Loss of imprinting of IGF2 sense and antisense transcripts in Wilms' tumor. *Cancer Res*. 2003;63:1900–5.
27. Saito H, Minamiya Y, Watanabe H, et al. Expression of the transient receptor potential channel c3 correlates with a favorable prognosis in patients with adenocarcinoma of the lung. *Ann Surg Oncol*. 2011;18:3377–83.
28. Yang SL, Cao Q, Zhou KC, Feng YJ, Wang YZ. Transient receptor potential channel C3 contributes to the progression of human ovarian cancer. *Oncogene*. 2009;28:1320–8.
29. Kim YJ, Yoon HY, Kim JS, et al. HOXA9, ISL1, and ALDH1A3 methylation patterns as prognostic markers for non-muscle invasive bladder cancer: array-based DNA methylation and expression profiling. *Int J Cancer*. 2013;133(5):1135–42.
30. Fox SB, Brown P, Han C, et al. Expression of the forkhead transcription factor FOXP1 is associated with estrogen receptor alpha and improved survival in primary human breast carcinomas. *Clin Cancer Res*. 2004;10:3521–7.
31. Tothill RW, Tinker AV, George J, et al; Australian Ovarian Cancer Study Group. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res*. 2008;14:5198–206.
32. Yoshihara K, Tajima A, Yahata T. Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. *PLoS One*. 2010;5:9615.
33. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. *Genet Epidemiol*. 2002;22:170–85.
34. Liu Z, Gartenhaus RB, Chen XW, Howell CD, Tan M. Survival prediction and gene identification with penalized global AUC maximization. *J Comput Biol*. 2009;16:1661–70.
35. Korthauer K, Dawson JA, Kendziorski C. Survival-supervized latent Dirichlet allocation models. Kim-Anh Do, Zhaohui Steve Qin, Marina Vannucci editors. *Advances in Statistical Bioinformatics*: *Models and Integrative Inference for High-Throughput Data*. Cambridge: Cambridge University Press; 2013:366–82.
36. Zilliox MJ, Irizarry RA. A gene expression bar code for microarray data. *Nat Methods*. 2007;4:911–3.
37. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. *J Am Stat Assoc*. 2006;101:1566–81.
38. Wainwright M, Jordan M. Graphical models, exponential families, and variational inference. *Found Trends Mach Learn*. 2008;1:1–305.
39. Jordan M, Ghahramani Z, Jaakkola T, Saul L. Introduction to variational methods for graphical models. *Mach Learn*. 1999;37:182–233.
40. Breslow N. Covariance analysis of censored survival data. *Biometrics*. 1974;30:89–99.

Appendix

The survLDA variational EM. *Posterior inference.* For a given document i with survival response dyad (T_i, δ_i), the key quantity of interest is

$$p(\theta_i, Z_{i,1:N_i} \mid w_{i,1:N_i}, T_i, \delta_i, \alpha, \tau_{1:K}, \beta, h_0) = \frac{p(\theta_i \mid \alpha)\left(\prod_{j=1}^{N_i} p(Z_{ij} \mid \theta_i)\, p(W_{ij} \mid Z_{ij}, \tau_{1:K})\right) p(T_i, \delta_i \mid Z_{i,1:N_i}, \beta, h_0)}{\int p(\theta_i \mid \alpha) \sum_{Z_{i,1:N_i}} \left(\prod_{j=1}^{N_i} p(Z_{ij} \mid \theta_i)\, p(W_{ij} \mid Z_{ij}, \tau_{1:K})\right) p(T_i, \delta_i \mid Z_{i,1:N_i}, \beta, h_0) \mathrm{d}\theta} \tag{1}$$

where the normalizing value is known as the *evidence*. As in LDA[12] and sLDA,[15] the evidence cannot be exactly computed efficiently, so we will use mean-field variational inference using Jensen's inequality to approximate it. For reviews of this and other variational methods, see Wainwright and Jordan[38] and Jordan et al.[39]

Let $\pi = \{\alpha, \tau_{1:K}, \beta, h_0\}$ and $q_i(\theta_i, Z_{i,1:N_i})$ denote a *variational distribution* of the latent variables. For computational tractability, we choose a fully factorized variational distribution:

$$q_i(\theta_i, Z_{i,1:N_i} \mid \gamma_i, \phi_{1,1:N_i}) = q_i(\theta_i / \gamma_i) \prod_{j=1}^{N_i} q_i(Z_{ij} \mid \phi_{ij}) \tag{2}$$

where

$$\theta_i | \gamma_i \sim Dir(\gamma_i) \text{ and } Z_{ij} | \phi_{ij} \sim Discrete(\phi_{ij}).$$

With this quantity defined, the lower bound for the evidence given by Jensen's inequality is

$$\begin{aligned}
&\log p(W_{i,1:N_i}, T_i, \delta_i \mid \pi) \\
&= \log \int_{\theta_i} \sum_{Z_{i,1:N_i}} p(\theta_i, Z_{i,1:N_i}, W_{i,1:N_i}, T_i, \delta_i \mid \pi) \mathrm{d}\theta \\
&= \log \int_{\theta_i} \sum_{Z_{i,1:N_i}} p(\theta_i, Z_{i,1:N_i}, W_{i,1:N_i}, T_i, \delta_i \mid \pi) \frac{q(\theta_i, Z_{i,1:N_i})}{q(\theta_i, Z_{i,1:N_i})} \mathrm{d}\theta \\
&= \log E_{qi}\left[p(\theta_i, Z_{i,1:N_i}, W_{i,1:N_i}, T_i, \delta_i \mid \pi) \frac{1}{q(\theta_i, Z_{i,1:N_i})} \right] \\
&\geq E_{qi}\left[\log p(\theta_i, Z_{i,1:N_i}, W_{i,1:N_i}, T_i, \delta_i \mid \pi) \frac{1}{q(\theta_i, Z_{i,1:N_i})} \right] \\
&= E_{qi}[\log p(\theta_i, Z_{i,1:N_i}, W_{i,1:N_i}, T_i, \delta_i \mid \pi)] + -E_{qi}[\log q(\theta_i, Z_{i,1:N_i})]
\end{aligned} \tag{3}$$

where the second term in the lower bound is the entropy $H(q_i)$ of the variational distribution. We will use $L(\cdot)$ to refer to the so-called *evidence lower bound* (ELBO) given in (3). We can expand the ELBO:

$$\begin{aligned}
L(W_{i,1:N_i}, T_i, \delta_i \mid \pi) &= E_{qi}[\log p(\theta_i \mid \alpha)] \\
&+ \sum_{j=1}^{N_i} E_{qi}[\log p(Z_{ij} \mid \theta_i)] + \sum_{j=1}^{N_i} E_{qi}[\log p(W_{ij} / Z_{ij}, \tau_{1:K})] \\
&+ E_{qi}[\log p(T_i, \delta_i \mid Z_{i,1:N_i}, \beta, h_0)] + H(q_i)
\end{aligned} \tag{4}$$

Thus, an approximation of the posterior given in (1) is obtained by maximizing L with respect to γ_i and $\phi_{i,1:N_i}$. The first, second, and third terms in (4), as well as the entropy $H(q_i)$, are identical to the corresponding terms in the ELBO for LDA[12] and sLDA[15]:

$$\begin{aligned}
E_{qi}[\log p(\theta_i \mid \alpha)] &= \log \Gamma\left(\sum_{k=1}^{K} \alpha_k \right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k) \\
&+ \sum_{k=1}^{K} (\alpha_k - 1)\left(\Psi(\alpha_k) - \Psi\left(\sum_{g=1}^{K} \alpha_g \right) \right)
\end{aligned} \tag{5}$$

$$\sum_{j=1}^{N_i} E_{qi}[\log p(Z_{ij} \mid \theta_i)] = \sum_{j=1}^{N_i} \sum_{k=1}^{K} \phi_{ijk} \left(\Psi(\alpha_k) - \Psi\left(\sum_{g=1}^{K} \alpha_g \right) \right) \tag{6}$$

$$\sum_{j=1}^{N_i} E_{qi}[\log p(W_{ij} \mid Z_{ij}, \tau_{1:K})] = \sum_{j=1}^{N_i} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{ijk} W_{ijv} \log \tau_{kv} \tag{7}$$

$$\begin{aligned}
H(q_i) &= -[\log \Gamma\left(\sum_{k=1}^{K} \gamma_{ik} \right) - \sum_{k=1}^{K} \log \Gamma(\gamma_{ik}) \\
&+ \sum_{k=1}^{K} | (\gamma_{ik} - 1)\left\{ \Psi(\gamma_{ik}) - \Psi\left(\sum_{g=1}^{K} \gamma_{ig} \right) \right\} \\
&+ \sum_{j=1}^{N_i} \sum_{k=1}^{K} \phi_{ijk} \log \phi_{ijk}]
\end{aligned} \tag{8}$$

where Ψ denotes the digamma function. All that remains is to derive the fourth term of (4):

$$
\begin{aligned}
&E_q[\log p(T_i,\delta_i \mid Z_{i,1:N_i},\beta,h_0)]\\
&= E_q[\log[\{h_0(T_i)\exp(\beta' \bar{Z}_i)\}_i^{\delta} \times \exp\{-H_0(T_i)\exp(\beta' \bar{Z}_i)\}]]\\
&= E_q[\delta_i \log h_0(T_i) + \delta_i \beta' \bar{Z}_i - H_0(T_i)\exp(\beta' \bar{Z}_i)]\\
&= \delta_i \log h_0(T_i) + \delta_i E_q[\beta' \bar{Z}_i] - H_0(T_i) E_q[\exp(\beta' \bar{Z}_i)] \qquad (9)\\
&= \delta_i \log h_0(T_i) + \delta_i \beta' \bar{\phi}_i - H_0(T_i)\left[\prod_{j=1}^{N_i}\left(\exp(\frac{\beta}{N_i})' \phi_{ij}\right)\right]
\end{aligned}
$$

where the K-vector

$$\bar{\phi}_i = (1/N_i)\sum_{j=1}^{N_i} \phi_{ij}.$$

We use block coordinate-ascent variational inference, maximizing (4) with respect to γ_i and then each $\varnothing_{ij}$ in turn. As in sLDA,[15] the terms of (4) involving γ_i are unchanged from LDA, and hence, the update for γ_i is

$$\gamma_i^{new} = \alpha + \sum_{j=1}^{N_i} \phi_{ij} \qquad (10)$$

The update for a given $\varnothing_{ij}$, however, must be derived anew. We first define the following quantities:

$$
\begin{aligned}
\psi_i &= \left(\Psi(\gamma_{i1}) - \Psi\left(\sum_{g=1}^{K}\gamma_{ig}\right),\ldots,\Psi(\gamma_{iK}) - \Psi\left(\sum_{g=1}^{K}\gamma_{ig}\right)\right)\\
\xi_{ij} &= \left(\sum_{v=1}^{V} \mathrm{I}(W_{ij}=v)\log\tau_{1v},\ldots,\sum_{v=1}^{V}\mathrm{I}(W_{ij}=v)\log\tau_{Kv}\right)
\end{aligned}
$$

and then take the partial derivative of (4) with respect to $\varnothing_{ijk}$:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \phi_{ijk}} &= 0 + \psi_{ik} + \xi_{ijk} + [-\log\phi_{ijk} - 1]\\
&\quad + \partial_i \frac{\beta_k}{N_i} - H_0(T_i)\left(\prod_{m\neq j}\exp(\frac{\beta}{N_i})'\phi_{im}\right)\exp(\frac{\beta_k}{N_i})
\end{aligned} \qquad (11)
$$

Setting this equal to zero and plugging in ϕ_{ijk}^{new} yields:

$$\phi_{ijk}^{\text{new}} \propto \exp\left[\psi_{ik} + \xi_{ijk} + \delta_i \frac{\beta_k}{N_i} - H_0(T_i)\left(\prod_{m\neq j}\exp(\frac{\beta}{N_i})'\phi_{\text{im}}\right)\exp(\frac{\beta_k}{N_i})\right] \qquad (12)$$

where proportionality means that the components of ϕ_{ij}^{new} are evaluated according to (12) and then normalized so that their sum is one. Variational inference proceeds by iteratively updating the variational parameters $\{\gamma_i, \phi_{i,1:N_i}\}$ according to (10) and (12) in order to find a local optimum for the ELBO, which in turn best approximates the evidence given in (1).

Parameter estimation. We use maximum likelihood estimation based on variational EM. Our data are $D = \{W_{i,1:N_i}, T_i, \delta_i\}$.

$$
\begin{aligned}
&L(\alpha,\tau_{1:K},\beta,h_0;D)\\
&= \sum_{i=1}^{D}\{E_{qi}[\log p(\theta_i, Z_{i,1:N_i}, W_{i,1:N_i}, T_i, \delta_i)] + H(q_i)\}
\end{aligned} \qquad (13)
$$

In the E-step, we use the variational inference algorithm outlined here in the first subsection to estimate the approximate posterior distribution for each document–response pair. In the M-step, we maximize the corpus-level ELBO with respect to π, subject to some constraints. First, we take α to be $(\alpha_o/K, \ldots, \alpha_0/K)$, where α_0 is specified a priori. This is not necessary; further structure could be placed on a, ranging from a simple Dirichlet prior as in (12) to more complicated structures allowing dependence among the documents more complex than simple conditional independence as in the study by The et al.[37] However we, like Blei and McAuliffe,[15] prefer letting α be user-defined, which is simple and straightforward, yet allows some flexibility in the model specification.

The $\tau_{1:K}$ updates are unchanged from unsupervised LDA (12; 15) and are thus calculated in this manner:

$$\hat{\tau}_{kv}^{\text{new}} \propto \sum_{i=1}^{D}\sum_{j=1}^{N_i} \mathrm{I}(W_{ij}=v)\phi_{ijk} \qquad (14)$$

where proportionality means that each $\hat{\tau}_{kv}^{\text{new}}$ is normalized to sum to one.

The regression coefficients that comprise β and the baseline hazard h_0 must be numerically optimized with respect to maximizing the portion of the joint ELBO that depends on them. Numerical optimization is required as no closed form can be derived in general for the maximizing choice of β. The specific computations this process entails depend on the choice for h_0. For example, when an exponential survival model is chosen, so that $h_0 = \lambda$, β, and λ are numerically optimized by finding the solutions:

$$
\begin{aligned}
&(\hat{\beta}^{\text{new}}, \lambda^{\text{new}}) = \operatorname{argmax} L(\beta,\lambda)\\
&= \operatorname{argmax}\sum_{i=1}^{D}\left[\delta_i \log\lambda + \delta_i\beta'\bar{\phi}_i - \lambda T_i \times \prod_{j=1}^{N_i}\exp(\frac{\beta}{N_i})'\phi_{ij}\right]
\end{aligned} \qquad (15)
$$

Numerical optimization for a Weibull survival model is similar. In contrast, if we use a nonparametric Breslow estimate[40] for h_0, we first update β given the current value for h_0:

$$\hat{\beta}^{\text{new}} = \text{argmax}\, L(\beta)$$
$$= \text{argmax} \sum_{i=1}^{D} \left[\delta_i \log h_0 + \delta_i \beta' \bar{\phi}_i - H_0(T_i) \prod_{j=1}^{N_i} \exp(\frac{\beta}{N_i} \gamma \phi_{ij} \right] \quad (16)$$

Then, given the updated $\beta = \hat{\beta}^{\text{new}}$, the maximum likelihood estimate of the baseline hazard h_0 at the rth ordered survival time t_r is given by[40]:

$$\hat{h}_0^{\text{new}}(t_r \mid \beta) = \frac{m_r}{(t_r - t_{r-1}) \sum_{j \in R_r} \exp(\beta' \bar{\phi}_j)} \quad (17)$$

where m_r is the number of failures at time t_r and R_r is the set of patients who have not failed or been censored by time t_r. Regardless, once h_0 has been updated an estimate of the cumulative baseline hazard H_0 follows immediately.

Prediction. Given a new patient with document $w_{1:N}$ and a fitted model $\{\alpha, \tau_{1:K}\}$, the posterior mean $\bar{Z}_{\text{new}} = \bar{Z} \mid w_{1:N}, \alpha, \tau_{1:K}$ can be obtained in order to estimate from what topics this new patient draws words and in what proportion. This is similar to the procedure outlined in the *Posterior Inference* subsection, except that all survival-related terms in the ELBO are dropped. Thus, under the same variational distribution as given in (2), the coordinate ascent updates are

$$\gamma^{\text{new}} = \alpha + \sum_{j=1}^{N} \phi_j \quad (18)$$

$$\phi_{jk}^{\text{new}} \propto \exp[\upsilon_i k + \xi_{jk}] \quad (19)$$

where again j indexes words, k indexes topics, and proportionality means that the components of ϕ_j^{new} are evaluated according to the above update and then normalized so that their sum is one. Note that this variational sequence is identical to that in Blei and McAuliffe,[15] as they point out that it does not depend on the particular response type.

Given $\bar{Z}_{\text{new}}$, measures related to topic membership can be predicted for a new document. This may be done qualitatively or quantitatively using the chosen survival model. For example, the predicted median lifetime can be obtained by solving the following equation for $\widehat{t_{\text{med}}}$:

$$\exp\left[-H_0(\widehat{t_{\text{med}}}) \exp(\beta' \bar{Z}_{\text{new}})\right] = \frac{1}{2} \quad (20)$$

where H_0 and β are taken from the fitted survLDA model.

Including a background topic. Finally, we introduce into the variational EM an 'uninteresting' background topic to act as a benchmark with respect to the supervising outcome. The background topic is used because in most documents (in our application as well as others), there is an imbalance in word frequency that is not totally overcome by considering term-frequency and/or inverse-document frequency measures. In our setting, for example, there are far fewer drug-related words as compared with gene-related words due to the nature of the data. To address this, the background topic puts proportional weight on the words according to their final frequency sum in all the documents. Assuming the vocabulary contains V words in total and that each word in the vocabulary appears $n_1, n_2, \ldots, n_V$ times through all the documents, the weight for each word in the background topic should be $\frac{n_1}{\sum_{i=1}^{V} n_i}, \frac{n_2}{\sum_{i=1}^{V} n_i}, \ldots, \frac{n_V}{\sum_{i=1}^{V} n_i}$. Non-background topics are then different deviations from this benchmark that express themselves through differential survival. In our application, the background topic would describe 'featureless' documents that contain only the ubiquitous adjuvant therapy information, nothing more. Upon convergence, the variational EM yields posterior estimates for the key quantities of interest: posterior estimates for the $\theta_{1:D}$ as well as for the composition $(\tau_{1:K})$ and outcome effect (β) of the K topics.