AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2023 American Psychological Association
ISSN: 1082-989X

2023, Vol. 28, No. 5, 1178–1206
https://doi.org/10.1037/met0000541

Supervised Latent Dirichlet Allocation With Covariates: A Bayesian Structural and Measurement Model of Text and Covariates

Kenneth Tyler Wilcox, Ross Jacobucci, Zhiyong Zhang, and Brooke A. Ammerman
Department of Psychology, University of Notre Dame

Abstract

Text is a burgeoning data source for psychological researchers, but little methodological research has focused on adapting popular modeling approaches for text to the context of psychological research. One popular measurement model for text, topic modeling, uses a latent mixture model to represent topics underlying a body of documents. Recently, psychologists have studied relationships between these topics and other psychological measures by using estimates of the topics as regression predictors along with other manifest variables. While similar two-stage approaches involving estimated latent variables are known to yield biased estimates and incorrect standard errors, two-stage topic modeling approaches have received limited statistical study and, as we show, are subject to the same problems. To address these problems, we proposed a novel statistical model—supervised latent Dirichlet allocation with covariates (SLDAX)—that jointly incorporates a latent variable measurement model of text and a structural regression model to allow the latent topics and other manifest variables to serve as predictors of an outcome. Using a simulation study with data characteristics consistent with psychological text data, we found that SLDAX estimates were generally more accurate and more efficient. To illustrate the application of SLDAX and a two-stage approach, we provide an empirical clinical application to compare the application of both the two-stage and SLDAX approaches. Finally, we implemented the SLDAX model in an open-source R package to facilitate its use and further study.

Translational Abstract

Text data is an increasingly popular data source in psychological research that can be analyzed with a variety of models and algorithms. Topic models are a popular measurement model that use latent variables to represent constructs underlying a set of documents (e.g., clinical interviews, survey open responses, written or spoken educational assessments). Recent applications have used estimates of these "topics" as predictors of other variables in a regression model, but the statistical behavior of this approach has not been well studied. Similar approaches with other latent variable models are known to yield incorrect regression coefficient estimates and incorrect inferences. We showed that the use of topic estimates as regression predictors is also prone to these problems. As a solution, we proposed a model that jointly estimates the topic model and regression model—supervised latent Dirichlet allocation with covariates (SLDAX). Using a simulation study under typical psychological text data conditions, we found that SLDAX estimates were generally more accurate and more precise than the two-stage approach. We illustrate the SLDAX and two-stage approaches in a clinical study of nonsuicidal self-injury and emotional dysregulation with participant interpersonal narratives. To allow researchers to apply the SLDAX model, we developed an open-source R software package.

Keywords: text mining, supervised topic modeling, mixture modeling, Bayesian estimation, regression

Text data, a facet of the big data whose availability has rapidly increased for psychology (e.g., Adjerid & Kelley, 2018), is a popular, rich, and challenging data source for psychological research. Psychological researchers have used text from a wide variety of sources such as social media (e.g., Schwartz et al., 2013), open-ended questions (Ammerman et al., 2021; Popping, 2015), and

This article was published Online First January 5, 2023.

Kenneth Tyler Wilcox https://orcid.org/0000-0003-3969-0427
Ross Jacobucci https://orcid.org/0000-0001-7818-7424
Brooke A. Ammerman https://orcid.org/0000-0002-0367-6589

An early version of this work was presented at the 2019 Society for Multivariate Experimental Psychology (SMEP) Conference and the 2020 Psychometric Society Conference. The conference abstract from the 2019 SMEP conference was published in *Multivariate Behavioral Research*. A draft of this article was posted as a preprint on PsyArXiv: https://doi.org/10.31234/osf.io/62tc3.

Data presented in this article was supported by National Institute of Mental Health 1F31MH107156-01A1 awarded to Brooke A. Ammerman.

Correspondence concerning this article should be addressed to Kenneth Tyler Wilcox, who is now at Department of Psychology, Cornell University, G126 Martha Van Rensselaer Hall, NY 14853, United States. Email: ktwilcox@cornell.edu

medical health record notes (Obeid et al., 2019). However, researchers face a dizzying array of algorithms to choose from to analyze and interpret text data. While overviews on the use of text mining algorithms for psychological research exist (Finch et al., 2018; Iliev et al., 2015; Jacobucci et al., 2021; Kjell et al., 2019; Rohrer et al., 2017), many of these algorithms were developed for much larger data sets than may be common in psychology and little research has focused on adapting these algorithms to questions unique to psychological research. Ad hoc multistage approaches have been dominant where a text algorithm (e.g., a topic model) is utilized and the resulting estimates of interest (e.g., topic proportion estimates) are used in a subsequent analysis (e.g., Finch et al., 2018; Kim, Kwak, Cardozo-Gaibisso, et al., 2017; Kim, Kwak, & Cohen, 2017; Packard & Berger, 2020; Rohrer et al., 2017; Schwartz et al., 2013). These approaches, however, remain underevaluated and their statistical performance may not be well understood. Therefore, psychology still needs rigorous statistical models that link text and nontext data with theory.

Quantitative analysis of text in psychology can be traced back to the development of the general inquirer system (Stone et al., 1966), which defined and measured psychological processes such as affection or distress based on a dictionary of words. This was motivated by the recognition that text contains both latent (complex interpretations of manifest features constructed from manifest features) and manifest (e.g., word frequencies, usage of parts of speech, etc.) content (Stone et al., 1966). This idea was more recently popularized through the development of the linguistic inquiry and word count software (LIWC; Pennebaker et al., 2015; Tausczik & Pennebaker, 2010), which uses a set of dictionaries to measure psychological constructs (e.g., affective and social processes). These resultant scores can be the variables of interest or can serve as inputs in a subsequent analysis (e.g., Kovacs & Kleinbaum, 2020; Packard & Berger, 2020). However, the relevance and scope of these categories may not be valid for new data in some applications. Pennebaker et al. (2003) cautioned that dictionary-based methods such as LIWC cannot handle the full scope of constructs that may occur in natural language, pointing to a limitation in the generalizability of predefined dictionaries. As a (partial) remedy to this limitation, new domain-specific dictionaries can be constructed to measure constructs of interest. This can be appealing if these constructs are not well-represented by predefined dictionaries such as those used by LIWC. In practice, manually constructing a dictionary can be time-consuming and expensive as it is challenging to exhaustively identify all terms that are both relevant and context invariant (Garten et al., 2018).

LIWC and other dictionary methods share another critical limitation. Words are often polysemic (i.e., a word can possess multiple semantically related meanings in different contexts or uses) or homonymic (i.e., a word can possess multiple unrelated meanings in different contexts or uses). “Happiness,” for example, can exhibit polysemy by referring to (a) current positive emotion or (b) positive evaluation of one’s life overall. “Lie” can exhibit homonymy by referring to, nonexhaustively, (a) resting or (b) deception. By assigning words to categories using dictionaries, this semantic complexity can be easily lost. Therefore, Pennebaker et al. (2003) and Kjell et al. (2019) have argued that data-driven methods such as latent semantic analysis may be better approaches to content or thematic analysis than dictionary methods. Latent semantic analysis (LSA, or indexing), an older but popular data-driven approach, is similar to principal component analysis where a matrix of word frequencies in each document is decomposed by singular value decomposition (Deerwester et al., 1990). This decomposition is designed to maximize the amount of variability in word frequencies using a set of q weighted linear combinations of the word frequencies. Typically, q is much smaller than V, the total number of unique words in the corpus, so each document can be represented by a set of these q eigenvector scores rather than the original word frequencies, achieving what is often a large degree of dimensionality reduction. Each of the q eigenvectors can be interpreted based on the coefficients corresponding to each of words. The resulting decomposition is often described as a “semantic space” (Deerwester et al., 1990) where proximity of words is indicative of semantic similarity and proximity of documents is indicative of content similarity. These eigenvectors can also be used as starting points for constructing dictionaries, allowing for faster development of new dictionaries for use with dictionary methods like LIWC. One limitation, however, is that LIWC and LSA are not statistical models and can be prone to overfitting (e.g., Blei et al., 2003).

Consequently, a more general and extensible approach is needed. Because LSA does not provide a generative model of text, we cannot perform inference nor can we falsify the “semantic” representations it produces. In light of more recent approaches that do provide a generative model of text along with similar dimensionality reduction and construction of semantic representations without the degree of overfitting of LSA, it is difficult to justify the choice of LSA on statistical grounds (Blei et al., 2003). One popular approach that addresses these limitations is topic modeling (Blei et al., 2003; Blei & Lafferty, 2009). We discuss the seminal latent Dirichlet allocation (LDA) topic model in detail later in this article. Briefly, topic models provide a fully generative statistical model of text that uses latent variables to represent text as a mixture of K latent categories (i.e., topics) that can be interpreted based on how probable each word in a vocabulary is given each topic. Documents can be represented based on the relative document-specific probability of each topic. The model is data-driven, rather than dictionary-based, and is capable of accounting for polysemy and homonymy while providing dimensionality reduction. In short, topic models not only share the desirable aspects of LSA, but also provide a full statistical model of text. This article does not intend to provide a comparison of LSA and topic modeling, but we note that the distinction between the two is analogous to the difference between principal component analysis and factor analysis. Although, superficially, the two appear to accomplish similar outcomes, principal component analysis is better suited to maximizing the explained variability in a set of variables with a mathematical model, whereas factor analysis is preferable when the aim is to accurately model correlations among variables with a generalizable statistical model (see, e.g., Widaman, 1993). LSA may provide a convenient mathematical decomposition of the variance in word frequencies, while topic modeling provides a generalizable statistical model of word co-occurrences.

Since its inception, topic modeling has attracted interest from psychological researchers (see, e.g., Griffiths & Steyvers, 2004). Recently, topic modeling has seen increasing use in psychological research in a variety of applications in order to study a range of research questions with domains including, for example, clinical, moral, and educational research questions (e.g., Finch et al., 2018; He, 2013; Kim, Kwak, Cardozo-Gaibisso, et al., 2017; Kim, Kwak,

& Cohen, 2017; Packard & Berger, 2020; Rohrer et al., 2017). Nearly every application of topic modeling to psychological research has used a two-stage approach in which (a) a topic model is first fit to text; and (b) estimated topic proportions are then used as predictors or outcomes in a subsequent (e.g., regression) model. However, the two-stage approach can be problematic because it treats the estimated topic proportions as though they were the true latent topic proportions. Problems associated with two-stage approaches with latent variable models are well known, particularly in the mixture modeling literature (e.g., Bolck et al., 2004; Vermunt, 2010) and the factor analysis literature (e.g., Devlieger et al., 2019; Hayes & Usami, 2020). As we will show in our simulation study, two-stage estimates of the relationships (i.e., regression coefficients) between the latent topic proportions and another manifest variable can be biased and may have incorrectly estimated standard errors. The same biased estimates and incorrect standard errors also occur in the case of two-stage regression approaches with continuous latent variables (Croon, 2002; Devlieger et al., 2016; Levy, 2017; Lu & Thomas, 2008; Skrondal & Laake, 2001). Given the growing popularity of topic modeling, an alternative to two-stage estimation procedures is needed. Three recent developments in this direction are the structural topic model (STM; M. E. Roberts et al., 2014), the supervised latent Dirichlet allocation model (SLDA; Blei & McAuliffe, 2008), and a hybrid of the STM and SLDA models proposed by Ansari et al. (2018). Both STM and SLDA extended topic modeling to incorporate other manifest variables. STM is similar to a multiple indicators–multiple causes (MIMIC; Joreskog & Goldberger, 1975) model as it models the effects of manifest predictors on the latent topics underlying the text. SLDA, on the other hand, treats a manifest outcome variable as an effect indicator of the topics by allowing the latent topics to predict the outcome. Finally, the Ansari et al. (2018) model allows covariates to predict the latent topics (like STM) and the topics to predict an outcome (like SLDA). Notably, the covariates predicting the latent topics do not predict the outcome. Choosing among these and other kinds of extended topic models should be performed with care as the different model specifications make different assumptions about covariate, topic, and outcome relationships and aim to answer different kinds of research questions.

SLDA, in particular, is more closely related to the two-stage procedure that is currently popular in psychological research where the latent topic estimates are used as predictors. However, SLDA does not allow for other manifest predictors of the outcome to be included. Recently, psychological applications of topic modeling have been interested in assessing the contribution of topics from text in conjunction with or in addition to other predictors. For example, Packard and Berger (2020) controlled for topical content from song lyrics in a regression analysis as a possible confounder of second-person pronoun usage (their predictor of interest) when predicting song popularity. Rohrer et al. (2017) used topics from survey responses regarding sources of worry as predictors of five personality traits from the Big Five Inventory while controlling for participant gender and age. In the former, the topic proportion estimates were used as control variables while in the latter, the topic proportion estimates were the primary predictors of interest while controlling for age and gender differences. Packard and Berger (2020) could not use SLDA to control for topical effects because SLDA does not allow for the inclusion of other manifest predictors. Rohrer et al. (2017) could have used SLDA to assess the *unconditional* relationships between topics and personality, but SLDA would not have allowed them to control for possible age and gender differences. In both cases, a two-stage approach was used to accomplish these aims.

Our goal in this article is to describe a new statistical model that jointly incorporates a latent variable measurement model of text and a structural regression model that links these latent variables with an additional set of manifest variables. This new method—termed supervised latent Dirichlet allocation with covariates (SLDAX)—generalizes related work on Bayesian topic modeling (Blei et al., 2003; Blei & McAuliffe, 2008) for text to accommodate both latent variables underlying the text and manifest variables as predictors without resorting to a two-stage estimation procedure (e.g., Finch et al., 2018; Packard & Berger, 2020; Rohrer et al., 2017). The rest of the article is organized as follows. First, we review the latent Dirichlet allocation (LDA) topic model as it serves as the foundation of our proposed model. Second, we discuss two-stage and one-stage approaches to using topic proportions to predict an outcome. Third, we describe the proposed SLDAX model and two MCMC algorithms for estimation and inference with continuous and dichotomous outcomes; we developed a freely available R package (*psychtm*; Wilcox, 2021) for estimating SLDAX and related models. Fourth, we discuss interpretation and statistical inference for SLDAX and related models. Fifth, we compare the performance of SLDAX and several variants of the commonly used two-stage approach in a simulation study. Sixth, we illustrate the application of SLDAX in an empirical example where clinical measures and interpersonal narratives are used to model emotional dysregulation. We end the article by discussing implications and limitations of the study and potential future research directions.

Topic Modeling: Latent Dirichlet Allocation

To help understand our proposed SLDAX model, this section summarizes latent Dirichlet allocation (LDA; Blei et al., 2003). LDA is a measurement model for a corpus (i.e., a set of documents or audio transcripts) that is designed to model relationships among word co-occurrences via a set of probability distributions. This is accomplished by introducing a set of K categorical latent random variables commonly known as *topics*. Typically, one must specify K before fitting the LDA model, although it is possible to treat K as a random variable and estimate it (Teh et al., 2006). We first describe the generative model for LDA and then describe the meaning and interpretation of the key model parameters.

Generative Model

Consider a corpus of D documents each of length N_d words, $d = 1, \ldots, D$. We refer to the set of all V unique words[1] in the corpus as a vocabulary—a given document may only contain a subset of these V words. Let $\vec{w}_d$ denote a $N_d \times 1$ vector of the observed words in document d and let $w_{dn} \in \{1, 2, \ldots, V\}$ denote an integer representation of the word in position n of document d, $n = 1, \ldots, N_d$. Let $\vec{z}_d$ be a $N_d \times 1$ vector of latent topic assignments corresponding to the observed words in document d and let $z_{dn} \in$

[1] We use words here for simplicity, but one can alternatively model phrases or other units of text.

$\{1, 2, \ldots, K\}$ be the latent topic assigned to w_{dn}. Let $\vec{\theta}_d$ denote a $K \times 1$ vector of topic probabilities or proportions under the constraint $\sum_{k=1}^{K} \theta_{dk} = 1$. Let $\vec{\beta}_k$ denote a $V \times 1$ vector of word probabilities for topic k under the constraint $\sum_{v=1}^{V} \beta_{kv} = 1$. The generative model for LDA (see Figure 1) is,

1. For each topic $k = 1, \ldots, K$:

   (a) Draw the word probability vector $(\vec{\beta}_k) \sim \text{Dir}(\vec{\gamma})$ with $\vec{\gamma}$ denoting the hyperparameters that reflect the relative concentrations of the word probabilities,

2. For each document $d = 1, \ldots, D$:

   (a) Draw topic probabilities $(\vec{\theta}_d) \sim \text{Dir}(\vec{\alpha})$ with $\vec{\alpha}$ denoting the hyperparameters that reflect the concentration of the topic probabilities,

   (b) For each word $n = 1, \ldots, N_d$:

      i. Draw topic assignment $(z_{dn} \mid \vec{\theta}_d) \sim \text{Cat}(\vec{\theta}_d)$,

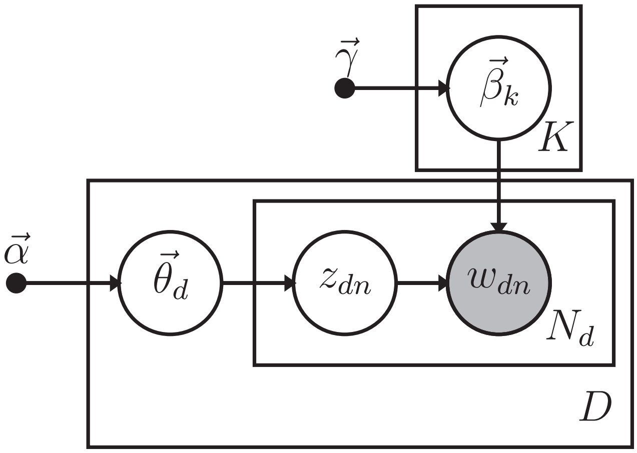
      ii. Draw word $(w_{dn} \mid z_{dn} = k, \vec{\beta}_k) \sim \text{Cat}(\vec{\beta}_k)$,

where Dir$(\cdot)$ denotes a Dirichlet distribution and Cat$(\cdot)$ denotes a categorical distribution.

Although the components of the V-dimensional hyperparameters $\vec{\gamma}$ and the K-dimensional hyperparameters $\vec{\alpha}$ can be freely specified, exchangeable Dirichlet distributions (equal hyperparameters for all components) are commonly used (e.g., Blei et al., 2003). The Dirichlet priors are convenient due to conjugacy, but alternative distributions can be used. Instead of using a Dirichlet prior for $\vec{\theta}_d$, for example, a logistic-normal prior more flexibly models the correlation structure of the topics (i.e., correlated topic model; Blei & Lafferty, 2006). The LDA likelihood function depends on the $D \times K$ matrix of topic proportions $\vec{\Theta} = [\vec{\theta}'_1, \ldots, \vec{\theta}'_d, \ldots, \vec{\theta}'_D]$ and the $K \times V$ matrix of topic-word probabilities $\vec{B} = [\vec{\beta}'_1, \ldots, \vec{\beta}'_k, \ldots, \vec{\beta}'_K]$. Under the assumptions that the documents are conditionally independent given $\vec{\Theta}$ and the words are conditionally independent given each document's topic assignments $\vec{z}_d$, the likelihood function is

Figure 1

Directed Acyclic Graphical Representation of the LDA Model



Note. The nth observed word w_{dn} in document d is represented by a shaded circle. Latent variables are represented by unshaded circles: z_{dn} denotes topic assignments for each word in each document; $\vec{\theta}_d$ denotes the K topic proportions for each document; $\vec{\beta}_k$ denotes the V topic-word probabilities for topic k. Fixed parameters are represented by dots: $\vec{\alpha}$ denotes the hyperparameters of the topic probabilities; $\vec{\gamma}$ denotes the hyperparameters of the topic-word probabilities. A set of (conditionally) independent replicates (i.e., words given topics; documents; word probabilities given a topic) is represented by a rectangle. LDA = latent Dirichlet allocation.

$$\begin{aligned} L(\vec{\Theta},\vec{B}) &= \prod_{d=1}^{D}\prod_{n=1}^{N_d}\Pr[z_{dn}=k \mid \vec{\theta}_d]\Pr[w_{dn}=v \mid z_{dn}=k,\vec{\beta}_k] \\ &= \prod_{d=1}^{D}\prod_{n=1}^{N_d}\theta_{dz_{dn}}\beta_{z_{dn}w_{dn}}, \end{aligned} \tag{1}$$

where $\theta_{dz_{dn}}$ is the probability of assigning topic z_{dn} to word w_{dn} in document d and $\beta_{z_{dn}w_{dn}}$ is the probability of word w_{dn} given the assigned topic z_{dn} in document d. Combining the prior distributions $f(\vec{\theta}_d), d = 1, \ldots, D$, and $f(\vec{\beta}_k), k = 1, \ldots, K$, with the likelihood, the posterior distribution is

$$\begin{aligned} &f(\vec{\Theta},\vec{B},\vec{z}_1,\ldots,\vec{z}_D \mid \vec{w}_1,\ldots,\vec{w}_D) \\ &= \frac{L(\vec{\Theta},\vec{B})\prod_{d=1}^{D} f(\vec{\theta}_d)\prod_{k=1}^{K} f(\vec{\beta}_k)}{f(\vec{w}_1,\ldots,\vec{w}_D)}. \end{aligned} \tag{2}$$

Parameter Interpretation and Estimation

The parameters of interest in LDA can be classified into document-level (person-level) and word-level (item-level) sets. For document d, $\vec{\theta}_d$ summarizes the content of the document as a mixture of the K topics with mixture proportions $\theta_{dk} = Pr[z_{dn} = k]$ for each topic. This allows for a lower-dimensional representation of each document using K topic proportions instead of V word frequencies. In many applications, some of the topics in a document may have probabilities near zero and can be considered ignorable so that document can be represented using a subset of $K^* < K$ topics. Furthermore, we can examine between-person variability in the documents by comparing the topic proportions across any pair of documents: that is, we can compare documents d and d' by comparing $\vec{\theta}_d$ and $\vec{\theta}_{d'}$.[2] For topic k, $\vec{\beta}_k$ summarizes topic k as a distribution over V words with probabilities $\beta_{kv} = Pr[w_{dn} = v \mid z_{dn} = k]$ for word v. Typically, many words will have ignorably small probabilities for a particular topic. In this case, that topic can be interpreted using this subset of $V^* < V$ words. As we illustrate later in our empirical example, the topic-specific word probabilities β_{kv} can be used like factor

[2] Between-person comparisons are limited to rank comparisons because the topic proportions are ipsative (Clemens, 1966). This is often desired because comparisons can be made based on the relative prevalence of each topic across individuals. If, instead, comparisons were made between individuals using the frequency of words assigned to each topic, more weight will be given to longer documents rather than the relative prevalence of each topic (this is also relevant for dictionary-based methods; Kovacs & Kleinbaum, 2020).

loadings in factor analysis to interpret each topic.[3] Overall, we can conceptualize $\vec{\Theta}$ as document or person-level parameters and $\vec{B}$ as word- or item-level parameters.

To obtain the posterior distribution of the LDA model in Equation 2, the marginal distribution $f(\vec{w}_1, \ldots, \vec{w}_D)$ needs to be computed, but this marginal distribution is intractable (Blei et al., 2003; Dickey, 1983). Therefore, exact inference using the posterior distribution is intractable. Instead, approximation algorithms such as Markov chain Monte Carlo (MCMC) algorithms (e.g., Gibbs sampling; Griffiths & Steyvers, 2004) or variational expectation-maximization (Blei et al., 2003) are used to estimate the model.

Use of Topic Proportions as Predictors

While the topic proportions $\vec{\Theta}$ can be used to summarize the corpus, it is common for researchers to use $\vec{\Theta}$ as variables (often predictors) in a subsequent model to study relationships between the content of the corpus and other measures (e.g., Do topics from free responses explain variability in self-reported emotional dysregulation?). This can be of interest for several reasons. Text data can provide auxiliary or complementary information to augment other scales and measures (Ercikan et al., 1998). By including both text and other measures as predictors of an outcome, researchers can study the effects of the topics on an outcome above and beyond or controlling for the other predictors. Alternatively, as in Packard and Berger (2020), information accounted for by the topics can be controlled for while evaluating the effects of the other predictors of interest on an outcome. To answer these types of research questions, two general approaches exist, namely, a two-stage procedure and a one-stage procedure.

Two-Stage Estimation

In a two-stage approach, the topic proportion estimates $\hat{\vec{\Theta}}$ obtained from a topic model (e.g., LDA) are used in a subsequent model. Perhaps because of its convenience, this two-stage procedure is popular within psychological research (Kim, Kwak, Cardozo-Gaibisso, et al., 2017; Kim, Kwak, & Cohen, 2017; Packard & Berger, 2020; Rohrer et al., 2017; Schwartz et al., 2013). A typical two-stage approach is as follows: (a) fit a LDA model to a corpus and determine the number of topics to use; (b) obtain estimates of the topic proportions $\hat{\vec{\Theta}}$; (c) estimate a (generalized) linear regression model[4] with $\hat{\vec{\Theta}}$ as predictors. Other predictors could also be included.

However, the two-stage approach may be problematic because it treats the estimated topic proportions as though they were the true latent topic proportions. This ignores estimation error and uncertainty. Two possible consequences are that the relationships (i.e., regression coefficients) between the latent topics and the dependent variable can be biased and that the estimated standard errors of the regression coefficients can be incorrect.[5] This has been studied in the mixture modeling literature in the context of latent class models where a two-stage approach is used to model relationships from covariates to latent class assignments (Bolck et al., 2004). As we will show in a simulation study, a two-stage approach using $\hat{\Theta}$ can lead to severely biased topic proportion regression coefficients and incorrect standard errors. Consequently, two-stage estimates of relationships between topic proportions and other variables can be inaccurate and lead to incorrect inference. To our knowledge, these problems have not been studied in the topic modeling literature.

One-Stage Estimation

The supervised LDA model (SLDA) proposed by Blei and McAuliffe (2008) is a one-stage model that extends LDA to include an outcome variable that is observed along with each document. Let $\vec{\bar{z}}_d = [\bar{z}_{d1}, \ldots, \bar{z}_{dk}, \ldots, \bar{z}_{dK}]'$ be the vector of proportions of topic assignments in a document, which we refer to as *empirical topic proportions* throughout the paper; $\bar{z}_{dk} = N_d^{-1} \sum_{n=1}^{N_d} I(z_{dn} = k)$ and $I(\cdot)$ is an indicator function equal to 1 if topic assignment n in document d is equal to k and 0 otherwise. Note that $\sum_{k=1}^{K} \bar{z}_{dk} = 1$. In SLDA, $\vec{\bar{z}}_d$ predicts the observed outcome y_d using a (generalized) linear regression model with a $K \times 1$ vector of regression coefficients $\vec{\eta}$,

$$y_d = \sum_{k=1}^{K} \bar{z}_{dk} \eta_k + \epsilon_d, \tag{3}$$

where ϵ_d is a residual with the usual assumption that $\epsilon_d \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$. The regression component of SLDA in Equation 3 corresponds to a first-degree canonical polynomial model for mixture proportions (Scheffé, 1958) and introduces a set of population-level parameters $\vec{\eta}$. No intercept is included in the regression model because of the constraint that $\sum_k \bar{z}_{dk} = 1$ in every document—inclusion of an intercept term would result in perfect collinearity. Equivalently, the model can be reparameterized to include an intercept and $K-1$ regression coefficients, but interpretation and inference for the regression coefficients becomes more challenging due to its dependence on the choice of the reference topic (Cornell, 2002). Using the generalized linear modeling framework (McCullagh & Nelder, 1989), SLDA can accommodate nonnormal outcomes with an appropriate canonical link function (Blei & McAuliffe, 2008). Like the posterior for LDA, the exact form of the posterior is intractable and approximation methods are needed for estimation. Blei and McAuliffe (2008) derived a variational inference algorithm to estimate the SLDA model.

Like LDA, SLDA models the corpus using topic proportions and word-topic probabilities (a measurement model of the text), but it jointly models the relationship between the topic assignments and the outcome (a structural model). As a result, the estimated topic assignments and topic proportions in SLDA will be related to the outcome provided that the population regression coefficients of the topics vary. In contrast, topics estimated by LDA will not necessarily have any relation to the outcome of interest because the outcome is not linked to the topics during the estimation procedure (Magnusson et al., 2020).[6] This distinction can impact performance: Blei and McAuliffe (2008), for example,

[3] In factor analysis, factors can be interpreted based on the loading of each manifest variable on each factor where interpretation is typically based on which items have nonzero loadings on a factor.

[4] Correlation analyses between estimated topic proportions and other manifest variables are a special case (i.e., univariate linear regression).

[5] The same bias and incorrect standard errors can also occur in the case of continuous latent variables (Croon, 2002; Lu & Thomas, 2008; Skrondal & Laake, 2001).

[6] Magnusson et al. (2020) compared the distinction between such a two-stage procedure (i.e., LDA followed by regression) and a supervised topic model to the difference between principal components regression (a two-stage procedure) and partial least squares regression (a one-stage procedure).

showed that SLDA had better predictive performance on new data in two empirical examples than a two-stage approach. Although SLDA is a one-stage approach, it does not incorporate other predictors besides the empirical topic proportions. In the next section, we introduce a generalization of SLDA that incorporates other manifest predictors in a one-stage approach.

Supervised Latent Dirichlet Allocation With Covariates (SLDAX)

Model Overview

As we discussed previously, text data can provide auxiliary or complementary information to complement other measures (Ercikan et al., 1998). By including topic proportions and other measures as predictors of an outcome, researchers can study the conditional effects of topics on an outcome and the conditional effects of other predictors on an outcome. For example, we can control for auxiliary information captured by topics from open-ended survey responses while evaluating the effects of other manifest predictors. Alternatively, we can assess the incremental contribution of complementary or auxiliary information accounted for by topics (e.g., from clinical interviews) after accounting for standard measures to model clinical outcomes.

Because SLDA does not incorporate other predictors besides the topics, an alternative model is needed. For example, in our empirical example, we model relationships between emotional dysregulation, nonsuicidal self-injury, and subjective interpersonal distress along with participant narratives of an interpersonal conflict. SLDA can only model the relationship between the text responses and a single outcome (e.g., emotional dysregulation). On the other hand, we could estimate a LDA model for the text and then use the topic proportion estimates with the other measures to model emotional dysregulation, but this is a two-stage approach with potentially poor performance (see our simulation results). To address this limitation, we developed the supervised latent Dirichlet allocation with covariates (SLDAX) model, a generalization of SLDA, that incorporates manifest variables and latent topics as predictors of an outcome.

Suppose that a set of p predictor variables denoted by the $p \times 1$ vector $\vec{x}_d$ is observed with document d and dependent variable y_d for D participants. The generative model for SLDAX with a normally distributed outcome (see Figure 2) is shown below,

1. Draw residual variance $\sigma^2 \sim \mathrm{IG}(\frac{a_0}{2}, \frac{b_0}{2})$,
2. Draw regression coefficients $\vec{\eta} \sim \mathrm{N}(\vec{\mu}_0, \vec{\Sigma}_0)$
3. For each topic $k = 1, \ldots, K$:
   (a) Draw $\vec{\beta}_k \sim \mathrm{Dir}(\vec{\gamma})$,
4. For each document $d = 1, \ldots, D$:
   (a) Draw $\vec{\theta}_d \sim \mathrm{Dir}(\vec{\alpha})$,
   (b) For each word $n = 1, \ldots, N_d$:
      i. Draw topic $z_{dn} \mid \vec{\theta}_d \sim \mathrm{Cat}(\vec{\theta}_d)$,
      ii. Draw word $w_{dn} \mid z_{dn} = k, \vec{\beta}_k \sim \mathrm{Cat}(\vec{\beta}_k)$,
   (c) Draw $y_d \sim \mathrm{N}(\vec{r}'_d \vec{\eta}, \sigma^2)$,

where σ^2 is the residual variance of y_d, IG($\cdot$) denotes an inverse-gamma distribution, $\vec{\eta}$ is a $(p+K) \times 1$ vector of regression coefficients, and $\vec{r}_d = (\vec{x}_d, \bar{\vec{z}}_d)$ is a $(p+K) \times 1$ vector that augments the empirical topic proportions with the vector of observed predictors. As in the case of SLDA, alternative link functions and distributions can be used by modifying step (4c) of the SLDAX generative model. In this article, we consider two cases: (a) y_d is normally distributed and (b) y_d is Bernoulli-distributed (via a logit link function). If the outcome is Bernoulli-distributed, we can replace Step 4c in Algorithm 4 with

$$(4c') \text{ Draw } y_d \sim \mathrm{Ber}(\pi_d),$$

where $\mathrm{logit}(\pi_d) = \vec{r}'_d \vec{\eta}$ and $\mathrm{logit}(\pi_d) = log(\frac{\pi_d}{1-\pi_d}), 0 < \pi_d < 1$. No intercept is included in the regression model because $\sum_k \bar{z}_{dk} = 1$ in every document. Regardless of the choice of link function and distribution of y_d, the exact form of the posterior distribution is intractable, so we next describe two MCMC algorithms to approximate the posterior distribution. We developed a freely available R package (*psychtm*; Wilcox, 2021) that implements these MCMC algorithms for estimating and summarizing SLDAX and related models.

We note that SLDAX is distinct from the STM. Like SLDA, the STM incorporates a LDA-style topic model of text with other manifest variables. The key distinction is that STM assumes that the manifest variables predict the topic proportions and/or topic-word probabilities. In comparison, SLDA assumes that the empirical topic proportions predict a manifest outcome. Our proposed SLDAX model is aligned with SLDA: The empirical topic proportions and an additional set of manifest variables jointly predict a manifest outcome. Consequently, the choice between STM and SLDAX depends on the research question(s) of interest.

Estimation

This section addresses the case where y is assumed to be normally distributed. For a sampling algorithm when the outcome is dichotomous, see Appendix A. Assuming that the outcomes y_d, documents, and the words are conditionally independent, the likelihood function is

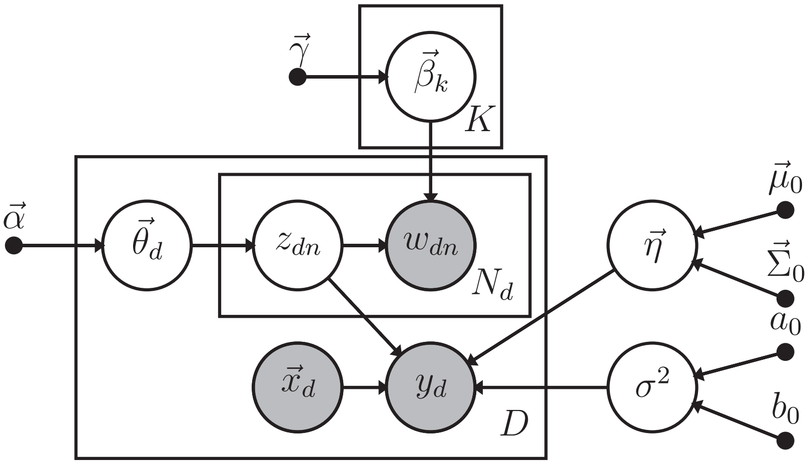
$$L(\vec{\Theta}, \vec{B}, \vec{\eta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{D}{2}} exp \left\{ -(2\sigma^2)^{-1} \sum_{d=1}^{D} (y_d - \vec{r}'_d \vec{\eta})^2 \right\} \prod_{d=1}^{D} \prod_{n=1}^{N_d} \theta_{d z_{dn}} \beta_{z_{dn} w_{dn}}. \quad (4)$$

Combining the priors and the likelihood, the posterior distribution is

$$f(\vec{\eta}, \sigma^2, \vec{\Theta}, \vec{B}, \vec{z}_1, \ldots, \vec{z}_D \mid \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D) = \frac{L(\vec{\Theta}, \vec{B}, \vec{\eta}, \sigma^2) f(\vec{\eta}) f(\sigma^2) \prod_{d=1}^{D} f(\vec{\theta}_d) \prod_{k=1}^{K} f(\vec{\beta}_k)}{f(\vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D)}. \quad (5)$$

Figure 2

Directed Acyclic Graphical Representation of the SLDAX Model With a Normal Outcome



Note. Observed variables are represented by shaded circles: w_{dn} denotes the nth word in document d; for subject d, $\vec{x}_d$ denotes p predictor scores and y_d denotes the outcome for subject d. Latent variables are represented by unshaded circles: z_{dn} denotes topic assignments for each word in each document; $\vec{\theta}_d$ denotes the K topic proportions for each document; $\vec{\beta}_k$ denotes the V topic-word probabilities for topic k; $\vec{\eta}$ denotes the regression coefficients relating $\vec{x}_d$ and $\vec{\bar{z}}_d$ to y_d; σ^2 denotes the residual variance of Y. Fixed parameters are represented by dots: $\vec{\alpha}$ denotes the hyperparameters of the topic probabilities; $\vec{\gamma}$ denotes the hyperparameters of the topic-word probabilities; $\vec{\mu}_0$ and $\vec{\Sigma}_0$ denote the prior mean vector and covariance matrix of $\vec{\eta}$, respectively; a_0 and b_0 are the shape and rate hyperparameters for σ^2. A set of (conditionally) independent replicates (i.e., words given topics; documents; word probabilities given a topic) is represented by a rectangle. SLDAX = supervised latent Dirichlet allocation with covariates.

Collapsed Gibbs Sampler Algorithm

We use an iterative Bayesian estimation algorithm—Gibbs sampling—to draw samples from the posterior. Within a Bayesian framework, model parameters and latent variables are viewed as random variables that have a joint distribution. The Gibbs sampler approximates this joint distribution by sampling sequentially from the full conditional distribution of each parameter and variable given the parameters and variables drawn in the previous iteration (e.g., Geman & Geman, 1984). We briefly describe the algorithm here. It is possible to sample from the joint posterior in Equation 5, but this requires sampling many parameters. In particular, there are $D(K-1)$ free parameters in $\vec{\Theta}$ and $K(V-1)$ free parameters in $\vec{B}$. As a result, mixing of the Markov chain from Gibbs sampling can be slow in practice. However, as Griffiths and Steyvers (2004) showed for LDA, the choice of Dirichlet priors for $\vec{\theta}_d$ and $\vec{\beta}_k$ allows $\vec{\Theta}$ and $\vec{B}$ to be integrated out of the posterior. For SLDAX, this yields a marginal posterior,

$$f(\vec{\eta}, \sigma^2, \vec{z}_1, \ldots, \vec{z}_D \mid \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D). \qquad (6)$$

Sampling from such a marginal posterior allows faster sampling with less autocorrelation between iterations (Liu, 1994). By collapsing out $\vec{\Theta}$ and $\vec{B}$, only $\vec{\eta}$, σ^2, and $\vec{z}_1, \ldots, \vec{z}_d, \vec{z}_D$ need to be sampled. After the Gibbs sampler converges, we can subsequently sample $\vec{\Theta}$ and $\vec{B}$ from their full conditional distributions. The computational steps are as follows. For iteration t, $t = 1, \ldots, T$:

1. Draw $\vec{\eta}^{(t)}$ from $f\left(\vec{\eta} \mid \sigma^{2^{(t-1)}}, \vec{z}_1^{(t-1)}, \ldots, \vec{z}_D^{(t-1)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D\right)$.
2. Draw $\sigma^{2^{(t)}}$ from $f\left(\sigma^2 \mid \vec{\eta}^{(t)}, \vec{z}_1^{(t-1)}, \ldots, \vec{z}_D^{(t-1)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D\right)$.
3. For n, $n = 1, \ldots, N_d$ and d, $d = 1, \ldots, D$:
   (a) Draw $z_{dn}^{(t)}$ from $f\left(z_{dn} \mid \vec{\eta}^{(t)}, \sigma^{2^{(t)}}, \vec{z}_1^{(t-1)}, \ldots, z_{d(-n)}^{(t-1)}, \ldots, \vec{z}_D^{(t-1)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D\right)$.
4. For document d, $d = 1, \ldots, D$:
   (a) Draw topic proportions $\vec{\theta}_d^{(t)}$ from $f\left(\vec{\theta}_d \mid \vec{\eta}^{(t)}, \sigma^{2^{(t)}}, \vec{z}_1^{(t)}, \ldots, \vec{z}_D^{(t)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D\right)$.
5. For topic k, $k = 1, \ldots, K$:
   (a) Draw topic–vocabulary distributions $\vec{\beta}_k^{(t)}$ from $f\left(\vec{\beta}_k \mid \vec{\Theta}^{(t)}, \vec{\eta}^{(t)}, \sigma^{2^{(t)}}, \vec{z}_1^{(t)}, \ldots, \vec{z}_D^{(t)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D\right)$.

when drawing $z_{dn}^{(t)}$, note that $z_{dn}^{(t-1)}$ is not conditioned upon. If $\vec{\Theta}$ and/or $\vec{B}$ are not of interest, Steps (4) and/or (5) can be omitted. The technical details of the derivations of the necessary conditional distributions are given in Appendix B.

Label Switching

Because the likelihood for SLDAX (as well as LDA and SLDA) is invariant to permutation of the topic labels, care is needed when summarizing the posterior samples for estimation and inference.[7] This permutation invariance leads to K! modes in the posterior that may be visited during sampling. This so-called label switching can result in parameter labels for the K topics (i.e., the topic proportions, topic-word probabilities, and topic regression coefficients) being permuted during sampling. As a consequence, ignoring label switching when obtaining posterior summaries can produce distorted posterior distributions and point estimates that do not correspond to any of the posterior modes.

Several strategies to resolve label switching have been proposed in the mixture modeling literature. The simplest approach in principle is to constrain some parameters that correspond to the mixture components such that a single posterior mode is identified during sampling (e.g., Richardson & Green, 1997). However, it is difficult to ensure that a particular set of constraints will guarantee the identification of a single mode and, worse, imposing such a constraint may even distort the sampled posterior distribution (Celeux et al., 2000). Clustering-type algorithms have been shown to provide better solutions to the label switching problem than ordered constraints (Celeux et al., 2000; Stephens, 2000). In particular, the algorithm proposed by Stephens (2000) has been shown to perform well (Cassiday et al., 2021; Dias & Wedel, 2004). Briefly, the Stephens algorithm seeks to align the draws at each iteration of the MCMC sampler of the matrix of K mixture proportions for the D observations by minimizing the Kullback-Leibler divergence between the classification matrix at each iteration and an average of the classification probabilities across the set of samples. In the context of the SLDAX model, we only need to supply the matrix of posterior topic proportions $\Theta^{(t)}$ at posterior samples t = 1, 2, . . . , T for the Stephens algorithm.

Another strategy to resolve the label switching problem is to change the prior specification. Instead of using symmetric priors on the topic parameters, asymmetric priors can be used. Research in this direction is limited, but Wallach et al. (2009) demonstrated empirically that asymmetrical priors on Θ combined with symmetric priors on B may improve model fit and interpretability for large corpora. This requires the specification of a more complicated hierarchical prior for Θ and is not readily available in most topic modeling software, so we do not pursue this approach further in this article.

Interpreting and Testing the Effects of Topics

Interpretation of the corpus parameters $\vec{\Theta}$ and $\vec{B}$ can be performed in the same way as for LDA. One nuance in interpreting the latent topics, much like the distinction between interpreting latent factors in factor analysis and structural equation modeling, is that the topics are conditioned on the outcome $\vec{y}$, meaning that $\vec{y}$ can essentially be seen as an additional indicator of each topic (see, e.g., Levy, 2017). Consequently, inclusion of $\vec{y}$ may lead to different topic-word probabilities $\vec{B}$ and topic proportions $\vec{\Theta}$ between LDA and SLDAX models.

Interpretation of the regression coefficients for the topics deserves particular attention. It is common, but incorrect, to compare the topic regression coefficients to zero for inference as is common practice in standard regression analysis; one may be tempted to interpret a positively signed coefficient as indicative of a positive relationship between the topic and the outcome and a negatively signed coefficient as indicative of a negative relationship (see, e.g., Packard & Berger, 2020; Rohrer et al., 2017; Schwartz et al., 2013). However, this is inappropriate and misleading because the topics are proportions. In this section, we briefly review key results from the mixture regression modeling literature and provide guidance on correct interpretation and testing of the topic regression coefficients in SLDAX and two-stage approaches.

Interpretation

In standard regression models, the absence of a relationship between a predictor and the outcome is commonly tested by assessing the corresponding regression coefficient η_k under the null hypothesis $H_0\colon \eta_k = 0$. However, when a set of K proportions $\{\bar{z}_{dk}\}_{k=1}^{K}$ are included as predictors, then the usual comparison of the regression coefficient for component k against zero is inappropriate because the set of proportions is constrained to sum to one. For simplicity, assume that no covariates are included (i.e., an SLDA model). The structural portion of the model is a regression equation where the empirical topic proportions are used as a set of K predictors,

$$\mathbb{E}[Y_d] = \sum_{k=1}^{K} \eta_k \bar{z}_{dk}. \tag{7}$$

Because the empirical topic proportions are constrained to sum to one, $\sum_k \bar{z}_{dk} = 1$, each coefficient η_k represents the expected value of Y when only component k is present (i.e., $\bar{z}_{dk}$ = 1 and $\bar{z}_{dk'} = 0$ for all $k' \neq k$; Cornell, 2002; Scheffé, 1958). To see why zero is not (generally) a meaningful point of comparison for η_k, consider the null hypothesis $H_0\colon \eta_k = 0, k =, \ldots, K$. Under this null hypothesis, $\mathbb{E}[Y_d] = \sum_{k=1}^{K} \eta_k \bar{z}_{dk} = \sum_{k=1}^{K} (0) \bar{z}_{dk} = 0$: the expected value of Y is zero when none of the empirical topic proportions are related to Y. Unless Y has been mean-centered, this will rarely be a meaningful hypothesis.

In a standard regression model, the comparable null hypothesis where all predictors have no relationship to Y is $H_0\colon \mathbb{E}[Y_d] = \eta_0$: the expected value of Y is a constant, but need not be equal to zero. Scheffé (1958) and Cox (1971) showed that the appropriate null hypothesis corresponding to a hypothesis that none of the K proportions are related to the outcome is $H_0\colon \eta_1 = \cdots = \eta_k = \cdots = \eta_K = \eta_0$ where η_0 is the population expected value of Y and η_0 need not be equal to zero. In an SLDAX model with additional predictors $\vec{X}$, η_0 will be the conditional expected value of Y when the other manifest predictors $\vec{X}$ are fixed at 0. The same reasoning holds for nonlinear

[7] Under symmetric Dirichlet priors, the posterior is also invariant to permutation of the topic labels.

regression models within the generalized linear model. Therefore, it is typically not meaningful to compare SLDAX (or SLDA) topic regression coefficients to zero for both interpretation and inference.

If, for example, the marginal sample mean of Y is equal to 5, we would expect the topic regression coefficient estimates from an SLDA model to be close to 5 when a topic is not related to Y, greater than 5 if the topic is positively associated with Y, and less than 5 if the topic is negatively associated with Y—a topic regression coefficient of 0 in this example would correspond to a topic that is negatively associated with Y, not unassociated with Y. In the case of a dichotomous outcome, topic regression coefficient estimates on the log-odds scale would be expected to vary around 0 (on the log-odds scale) when the marginal probability of $Y = 1$ is near .5. If the marginal probability is, say, .75, we would expect topic regression coefficients to vary around 1.1 instead of 0. Therefore, it is important to interpret topic regression coefficient estimates in the context of the marginal or conditional mean of Y rather than zero.

Inference

It follows from the discussion above that tests of the relationship between prevalence of a topic and the outcome differ from standard regression practice. Each coefficient η_k represents the average response when only topic k is present. One could compare $\hat{\eta}_k$ to η_0 to determine if that topic is related to the outcome, but this ignores the fact that increasing the prevalence of topic k corresponds to a reduction of the prevalence of the other $K - 1$ topics. For regression models with mixture components as predictors, Snee and Marquardt (1976) defined the effect of component k on Y as the change in the expected value of Y resulting from a change in the proportion of component k while holding the relative proportions of the remaining $K - 1$ components constant—this is in fact the slope of Equation 7 with respect to η_k along the $\bar{z}_k$ axis (for further discussion of mixture component slopes, see, e.g., Cornell, 2002). They define the effect as a contrast,

$$c_k = \eta_k - (K-1)^{-1}\sum_{k' \neq k}^{K} \eta_{k'}. \tag{8}$$

Throughout the rest of the article, we define the *effect of a topic* as the change in the expected value of Y resulting from a change in the empirical topic proportion $\bar{z}_k$ (for an SLDAX model) or the topic proportion θ_k (for a two-stage approach) of component k while holding the relative proportions of the remaining $K - 1$ topics constant; this effect will take the form of a contrast as in Equation 8, although two other approaches to defining this contrast will be discussed. The corresponding null hypothesis that component k has no unique relationship with Y is $H_0: c_k = 0$ (Park, 1978; Snee & Marquardt, 1976). In the context of SLDA and SLDAX models, this null hypothesis corresponds to the absence of a relationship between topic k and Y. Because we are using Bayesian estimation, we do not review details of the frequentist test statistics and distributional results (see Cornell, 2002; Snee & Marquardt, 1976). Instead, the posterior samples can be used directly to obtain posterior estimates of c_k and corresponding credible intervals for each topic. If the credible interval for c_k excludes 0, one could conclude that topic k is related to Y. Unlike the topic regression coefficients, the contrasts in Equation 8 can be directly compared to zero to determine whether and in what direction a topic is related to Y.

However, if the empirical topic proportions are range restricted (i.e., their range is less than 0 to 1), then Equation 8 is not an appropriate estimate of the topic effects. Because Equation 8 compares the expected value of Y when proportion k is equal to 0 (and the other proportions are equal to each other) to the expected value of Y when proportion k is equal to 1 (and all other proportions are 0), this comparison may not be meaningful if these points do not occur in the posterior empirical topic proportions with high probability—in other words, documents may not be completely explained by a single topic so proportions of 1 may never occur, particularly if it is a longer document; conversely, a topic may be used by all subjects to some extent so proportions of 0 may never occur. In this case, Equation 8 can be misleading. To account for range-restricted proportions, Snee and Marquardt (1976) defined a range-adjusted contrast by multiplying c_k by r_k,

$$c_k^{(adj)} = r_k c_k. \tag{9}$$

In the case of SLDAX, r_k is the range of the empirical topic proportions of topic k. This (nearly) restricts the test of topic effects to the observed space of empirical topic proportions.

However, Piepel (1982) showed that Equation 9 does not account for the relative size of the constrained region of the topic simplex and can still lead to tests where the location of the contrast proportions fall outside the observed empirical topic proportions or may even fall outside the range of 0 to 1. Consequently, the estimates of a given topic effect from Equation 9 can be of the wrong magnitude and even the wrong direction. Piepel (1982) proposed a third approach that resolves the problems associated with Equations 8 and 9. The algebraic steps are lengthy, so we refer readers to Piepel (1982) for details in order to save space. Generally, the approach compares the expected value of the outcome at the centroid of the constrained simplex to the expected value of the outcome at the maximum observed value of each proportion while holding the relative proportions of the other components constant. This accounts for any range restrictions in the empirical topic proportions and ensures that the estimated effect of changing a given empirical topic proportion is constrained to the observed space of proportions. We denote Piepel's contrasts as $c_k^{(P)}$.

Both Equation 9 and Piepel's method tend to yield more precise estimates (i.e., more powerful tests) of the contrasts of interest than Equation 8 because the latter approach does not adjust for range restriction. To ensure that the test of topic effects are consistent with the available data, we recommend Piepel's method instead of Equations 8 and 9, although Equation 9 is easier to use and may yield similar results to Piepel's method if range restriction is not extreme for any topic.[8] These methods are equally applicable to two-stage alternatives to SLDAX with the same limitations.

Finally, we note that the absence of statistical evidence for a nonzero topic effect does not imply that the topic should be removed from the model. Such a topic may provide useful

[8] If there is no range restriction in the empirical topic proportions, all three approaches are equivalent.

information about the measurement model of the corpus; refitting the model with fewer topics may deleteriously impact the quality of the measurement model of the responses. Rather, a topic effect near zero suggests that the topic may not be related to the outcome above and beyond the effects of the other topics and manifest predictors in the model. One might be interested in such a case in model simplification. One common approach in mixture regression modeling when several components have similar nonsignificant regression coefficient estimates is to refit the model under the constraint that these regression coefficients are equal (see, e.g., Cornell, 2002). In the case of SLDAX and SLDA, the structural model could be simplified by refitting the model using low-variance priors with a common (often nonzero) mean for the topic regression coefficients that are to be constrained.[9]

Simulation Study

This section describes a Monte Carlo simulation study designed to evaluate the performance of the Gibbs sampler for the SLDAX model and to compare its performance with a two-stage approach representative of current practice: (a) variational EM (VEM) is used to estimate an LDA model and (b) the estimated topic proportions from the first step are used as predictors along with another manifest predictor in a regression model. For this simulation, we used the SLDAX model with a normally distributed outcome Y as the population model. For D observations, the $D \times (K+1)$ matrix $\vec{R}$ included a single predictor X generated from a standard normal distribution and K empirical topic proportions $\vec{\vec{Z}}$. We set the number of topics K to 2 or 5. Without loss of generality, we set the average total variance of the outcome Y to 1. We fixed the hyperparameters as follows: α and γ were both set to 1. The magnitude of the regression coefficients was calculated so that the coefficient for X corresponded to a partial R^2 of .15 (a "medium" effect; Cohen, 1988) and the set of coefficients for the K topics jointly corresponded to a partial R^2 of .35: For $K = 2$, the coefficients were simulated to have equal magnitudes and opposite signs; for $K = 5$, the coefficients were simulated as $\{-2\eta^*, -\eta^*, 0, \eta^*, 2\eta^*\}$ for a constant η^* chosen to obtain the desired partial R^2. See Appendix D for details of the derivation of the regression coefficients. The residual variance was calculated as the difference between the total variance of the outcome Y (fixed to 1) and the variance explained by X and the topics. Using R (Version 3.6.2; R Core Team, 2019), we generated 100 data sets within each cell of a design with four crossed factors: (a) the number of topics K (2 and 5); (b) the number of subjects D (50, 200, 800, and 1,500); (c) the average document length $\bar{N}_d$ (15, 80, and 150); and (d) the vocabulary size V (500 and 1,000). The number of words in each document N_d were generated from a Poisson distribution with mean and variance equal to $\bar{N}_d$.

We used R to fit the SLDAX and two-stage models. SLDAX models were estimated with Gibbs sampling using the *psychtm* R package (Wilcox, 2021); the LDA models were estimated with the VEM algorithm proposed by Blei et al. (2003) using the *topicmodels* R package (Version 0.2-9; Grün & Hornik, 2011). When estimating the SLDAX models, we specified common diffuse priors: $\vec{\Sigma}_0$ was set to $10^4\vec{I}$; a_0 and b_0 were both set to .001; $\vec{\mu}_0$ was set to $\{-2, 2\}$ if $K = 2$ and $\{-5, -2.5, 0, 2.5, 5\}$ if $K = 5$ for the topic regression coefficients and 0 for the regression coefficient for X; α and γ were both set to 1. Possible label switching was handled using the relabeling algorithm proposed by Stephens (2000) which is implemented in the *label.switching* R package (Version 1.8; Papastamoulis, 2016). After examining trace plots and Geweke Z statistics to diagnose convergence on several artificial data sets, we generated 3,000 samples from the MCMC chain after a burn-in period of 5,000 iterations and a thinning period of 10.[10]

When estimating the LDA models, α was initialized to 1 and then either (a) estimated or (b) fixed to 1; γ is not used by the VEM algorithm when estimating the topic–word probabilities $\vec{\beta}$. The estimated α approach is commonly applied in practice and is the default in many topic modeling software packages, whereas the fixed α approach is more comparable to SLDAX because α is assumed to be fixed as a hyperparameter in the SLDAX algorithm. The maximum number of E steps was set to 500 with a convergence tolerance for the relative change in the log-likelihood of 10^{-6}; the maximum number of M steps was set to 1,000 with a convergence tolerance of 10^{-4}.[11] The topic proportion estimates $\hat{\vec{\vec{\Theta}}}$ were then used along with X in a linear regression model[12] of Y fit in R using (a) the *lm()* function for ordinary least squares regression or (b) the *gibbs_mlr()* function from the *psychtm* R package (Wilcox, 2021) for Bayesian regression using the same diffuse prior specifications for $\vec{\Sigma}_0$, $\vec{\mu}_0$, a_0, and b_0 as described above.[13] In summary, we compared five modeling approaches: (a) VEM LDA with α fixed followed by OLS regression; (b) VEM LDA with α estimated followed by OLS regression; (c) VEM LDA with α fixed followed by Bayesian regression; (d) VEM LDA with α estimated followed by Bayesian regression; and (e) SLDAX.

We compared the performance of the five modeling approaches in terms of estimation accuracy via relative bias and both accuracy and precision via the normalized root mean squared error (NRMSE). The relative bias for a parameter of interest, θ, was calculated by averaging $100\frac{\hat{\theta}_r-\theta}{\theta}$ (when $\theta \neq 0$) across replications where $\hat{\theta}_r$ is the point estimate from the rth replication.[14] We considered relative bias smaller than 10% ignorable (Hoogland & Boomsma, 1998; L. K. Muthén & Muthén, 2002). The NRMSE with R replications was

[9] The use of low-variance priors to constrain parameters has also been discussed in the context of Bayesian structural equation modeling by Muthén and Asparouhov (2012).

[10] Different random starting values were drawn from the prior distributions in each MCMC chain. Convergence was checked using the Geweke (1992) Z statistic for each parameter. Convergence rates were generally above 90% across conditions with one exception: When fitting a five-topic model with 200 subjects and an average document length of 80 words, convergence rates could decrease to 60%−70%.

[11] A different random initialization of the topic-word probabilities was used for each replication.

[12] As noted previously in the SLDAX model, no intercept was included to avoid perfect collinearity with $\hat{\vec{\vec{\Theta}}}$.

[13] Computational tasks were executed on the University of Notre Dame Center for Research Computing's supercomputing infrastructure using a Linux shell script to coordinate the simulation tasks. All simulation code is available upon request from the first author.

[14] When the population parameter was equal to zero, bias was calculated instead of relative bias by averaging $\hat{\theta}_r - \theta$ across replications.

calculated by $100\sqrt{\frac{\sum_{r=1}^{R}(\hat{\theta}-\theta)^2}{R}}/|\theta|$. For SLDAX models, $\hat{\theta}_r$ was calculated using the posterior mean. Smaller NRMSE indicates better estimation. We focus on the regression coefficients for X and $\vec{\vec{Z}}$ because these parameters are often of primary interest when incorporating topics and manifest predictors in a regression context. ANOVA models using sandwich-type standard error estimates (Huber, 1967; White, 1980) were fit for each regression coefficient to the relative bias and NRMSE in order to assess which simulation factors (modeling approach [five], number of topics [two], number of subjects [three], average document length [three], vocabulary size [two]) influenced estimation accuracy and precision. Given the large sample size (10,800 total samples), all factors for some regression coefficient metrics were significant including the five-way highest-order interaction, even though not all of these effects may be practically significant. Following standard practice in this scenario (e.g., Baird & Maxwell, 2016), factors (main effects and interactions) that were significant and contributed a partial R^2 of at least 2% (a "small" effect; Cohen, 1988) are emphasized below in the text, figures, and tables.[15]

Bias

The average relative biases of the estimated regression coefficients are shown in Tables 1, 2, and 3.

Regression Coefficient for the Manifest Predictor

For the regression coefficient of the manifest variable X, the total R^2 was small (1.0%) and the largest partial R^2 corresponding to a four-way interaction among modeling approach, number of subjects, average document length, and vocabulary size was negligible (.1%), so we do not discuss this model further. Bias for the estimates of the regression coefficient of X was ignorable for all five methods across all simulation conditions, $M = .3\%$, $SE = .1\%$, 95% CI [.0%, .6%].

Regression Coefficients for the Empirical Topic Proportions

The accuracy of the five modeling approaches differed, however, for the topic regression coefficients $\hat{\eta}_Z$—results were virtually identical across the topic regression coefficients so we present results for the first coefficient to save space. A three-way interaction among modeling approaches, the number of subjects, and the average document length, $F(24, 23145) = 431.40$, $p < .001$, $R^2_{partial} = 6.3\%$, and a three-way interaction between modeling approaches, the number of subjects, and the number of topics, $F(12, 23145) = 363.80$, $p < .001$, $R^2_{partial} = 2.5\%$, both explained more than 2% of the variance in relative bias. Including the associated lower-order factors with these factors explained 70.7% of the total variance. Relative bias for the first topic regression coefficient for the two-stage approaches using OLS regression and SLDAX is shown in Figure 3. Estimation bias was virtually identical across conditions using two-stage approaches when comparing OLS regression and Bayesian regression, so we present the two-stage results for OLS regression only for simplicity below.

For the two-stage approach with *estimated* α, the topic regression coefficient estimates were always biased. The magnitude of this bias increased as the number of subjects increased, as the average document length increased, and as the number of topics increased. In most cases, the topic regression coefficient were overestimated with relative bias often exceeding 100% and, in some cases, reaching approximately 2,000%. When the average document length was 15 words, bias was nonignorably negative with fewer than 1,500 subjects and nonignorably positive with 1,500 subjects. This suggests that using the VEM LDA algorithm with estimated α in a two-stage procedure can substantially inflate the magnitude of the estimated topic regression coefficients.[16]

Unlike the estimated α two-stage approach, *fixed* α two-stage estimates of the topic regression coefficients were not positively biased. However, this approach was still problematic because, unexpectedly, estimation bias did not consistently decrease as the average document length or the number of subjects increased. When the average document length was 15 words, the topic regression coefficient estimates were negatively biased: For a five-topic model, estimation bias was relatively unaffected by the number of subjects; for a two-topic model, estimation bias decreased in magnitude as the number of subjects increased. When the average document length was 80 words, estimation bias decreased in magnitude as the number of subjects increased for a five-topic model. However, estimation bias for a two-topic model unexpectedly became more negative as the number of subjects increased. When the average document length was 150 words, estimation bias decreased in magnitude as the number of subjects increased for two- and five-topic models. With one exception[17], estimation bias was only ignorable when the average document length was 150 words with at least 800 or 1,500 subjects depending on the number of topics.

Like the fixed α two-stage approach, SLDAX estimates of the topic regression coefficients were not positively biased. However, the bias of the SLDAX estimates was consistent with expectations that it would decrease with more subjects and longer documents. For two-topic models, the topic coefficient estimates were negatively biased when the average document length was 15 words but approached the population values as the number of subjects increased; for an average document length of 80 words or more, the bias of the estimated topic regression coefficients was ignorable for a two-topic model with at least 200 subjects. For five-topic models, the topic coefficients were always negatively biased, but the magnitude of the bias depended on both the number of subjects and the average document length; for an average document length of 15 words, estimates were negatively biased ($M = -98\%$) even with 1,500 subjects. Holding the number of subjects fixed, increasing the average document length decreased the magnitude of the bias. Similarly, holding the average document length fixed, increasing the number of subjects decreased the magnitude of the bias, although increasing the number of subjects from 800 to 1,500 reduced the magnitude

[15] Complete results are available from the first author upon request.

[16] Inspection of the α estimates in these conditions showed that the estimated α was often one to two orders of magnitude larger than the population value.

[17] Bias was ignorable when the average document length was 80 words for a two-topic model only with 200 subjects.

Table 1
Relative Bias (%) for the Two-Stage Approach With Estimated α and OLS Regression

K	D	$\overline{N}_d$	η_X	$\eta_{\overline{Z}_1}$	$\eta_{\overline{Z}_2}$	$\eta_{\overline{Z}_4}$	$\eta_{\overline{Z}_5}$	σ^2
2	50	15	1.2	−87.8	−88.1			67.3
2	50	80	6.5	−59.5	−60.0			45.2
2	50	150	2.8	42.0	40.3			37.0
2	200	15	−0.8	−86.5	−85.5			68.4
2	200	80	−0.6	142.2	141.9			59.7
2	200	150	1.4	366.5	367.0			59.0
2	800	15	−0.2	−19.5	−19.2			69.0
2	800	80	0.9	428.0	427.8			65.0
2	800	150	0.4	546.5	545.7			62.1
2	1,500	15	−0.5	27.3	27.6			69.6
2	1,500	80	0.7	440.5	440.9			65.5
2	1,500	150	0.1	615.0	614.6			59.7
5	50	15	−2.5	−80.3	−82.0	−84.3	−79.9	71.3
5	50	80	0.3	−77.2	−80.4	−81.2	−78.7	64.4
5	50	150	0.7	−72.2	−75.2	−75.8	−71.8	59.3
5	200	15	−0.3	−87.7	−90.1	−88.6	−86.4	73.0
5	200	80	−0.0	312.2	261.0	248.4	301.2	66.5
5	200	150	−0.3	946.3	801.9	743.4	946.0	62.9
5	800	15	−0.6	124.6	86.0	84.8	125.1	71.3
5	800	80	0.9	1,123.0	987.5	940.4	1,186.9	67.4
5	800	150	0.1	1,774.6	1,595.1	1,500.4	1,828.2	65.9
5	1,500	15	−0.0	284.9	221.9	213.0	279.7	71.5
5	1,500	80	0.1	1,168.5	922.1	904.0	1,171.5	67.7
5	1,500	150	0.6	1,948.5	1,648.7	1,740.5	1,980.2	64.0

Note. OLS = ordinary least squares; K = number of topics; D = number of documents; $\overline{N}_d$ = average document length (words); η_X = regression coefficient for predictor X; $\eta_{\overline{Z}_1}$ = regression coefficient for topic 1; $\eta_{\overline{Z}_2}$ = regression coefficient for topic 2; $\eta_{\overline{Z}_4}$ = regression coefficient for topic 4; $\eta_{\overline{Z}_5}$ = regression coefficient for topic 5; σ^2 = residual variance. Note that for $\eta_{\overline{Z}_3} = 0$, relative bias is undefined so it is not included.

of the bias less than the improvement in bias from 200 to 800 subjects. In comparison, increasing the average document length noticeably reduced the magnitude of the bias. Comparison of Tables 2 and 3 shows that SLDAX generally yielded more accurate estimates than the fixed α two-stage approach except when the average document length was 15 words for which SLDAX yielded slightly more negatively biased estimates.[18]

NRMSE

NRMSE, which measures both bias and variance of the estimator, for the three approaches is shown in Tables 4, 5, and 6.

Regression Coefficient for the Manifest Predictor

For the regression coefficient of the manifest variable X, the total R^2 was 29.4%, but was almost completely accounted for by the main effect of the number of subjects, $F(3, 23145) = 3108.97$, $p < .001$, $R^2_{partial} = 28.4\%$. As expected, NRMSE decreased significantly as the number of subjects increased for all approaches. No other factors explained more than .1% of the variance.

Regression Coefficients for the Empirical Topic Proportions

NRMSE results were virtually identical across the topic regression coefficients, so we focus on results for the first coefficient to save space. For the topic proportion regression coefficient, the total R^2 was 85.7%. A four-way interaction among modeling approaches, the number of subjects, the number of topics, and the average document length, $F(24, 23145) = 226.29$, $p < .001$, $R^2_{partial} = 3.4\%$, explained more than 2% of the variance in NRMSE. Including the associated lower-order factors with these factors explained 76.2% of the total variance. NRMSE was virtually identical across conditions using two-stage approaches when comparing OLS regression and Bayesian regression, so we present the two-stage results for OLS regression only for simplicity below.

For the estimated α two-stage approach, the NRMSE was greater across all conditions for the topic regression coefficients than both the fixed α two-stage approach or SLDAX, so we do not discuss results from this method further.

NRMSE for the fixed α and SLDAX methods is shown in Figure 4. In all conditions except when the average document length was 15 words and a five-topic model was fit, the NRMSE of the SLDAX estimates was generally smaller than or comparable to the NRMSE of the fixed α two-stage estimates. Furthermore, the NRMSE from the fixed α two-stage estimates increased for two-topic models when the average document length was 80 and 150 as the number of subjects increased, which is consistent with the inflated bias in these conditions discussed previously. The NRMSE of the SLDAX estimates behaved as expected by consistently decreasing as the amount of data increases. The advantage of SLDAX can be seen clearly for conditions where the estimation bias from both the fixed α two-stage approach and SLDAX was ignorable because NRMSE then corresponds to estimation standard error: Estimation bias from both approaches was ignorable in two conditions ($K = 2$, $D = 200$, $\bar{N}_d = 80$ and $K = 2$, $D = 1, 500$,

[18] Two exceptions occurred for a five-topic model when the average document length was 150 words with 800 and 1,500 subjects where the fixed α two-stage approach yielded slightly less biased estimates than SLDAX.

Table 2
Relative Bias (%) for the Two-Stage Approach With Fixed α and OLS Regression

K	D	$\bar{N}_d$	η_X	$\eta_{\bar{Z}_1}$	$\eta_{\bar{Z}_2}$	$\eta_{\bar{Z}_4}$	$\eta_{\bar{Z}_5}$	σ^2
2	50	15	1.5	−85.4	−85.9			67.4
2	50	80	5.6	−55.0	−55.2			42.3
2	50	150	2.5	−24.2	−25.8			13.2
2	200	15	−0.8	−83.5	−82.5			67.9
2	200	80	−1.2	−8.1	−8.9			12.4
2	200	150	1.1	−15.5	−14.8			20.7
2	800	15	−0.2	−68.8	−68.4			64.1
2	800	80	0.4	−25.2	−25.2			33.1
2	800	150	−0.0	−19.1	−20.9			55.8
2	1,500	15	−0.6	−43.1	−42.8			53.9
2	1,500	80	0.5	−43.6	−43.9			56.9
2	1,500	150	−0.2	−7.8	−8.8			56.6
5	50	15	−0.8	−71.5	−75.3	−74.8	−73.0	72.0
5	50	80	1.4	−73.0	−78.2	−78.0	−74.5	64.3
5	50	150	2.5	−66.0	−71.0	−70.7	−66.4	57.1
5	200	15	−0.1	−80.5	−83.6	−82.1	−80.1	73.3
5	200	80	0.5	−67.5	−71.8	−73.5	−67.2	63.0
5	200	150	−0.4	−43.8	−50.2	−50.7	−43.4	45.8
5	800	15	−0.5	−81.3	−85.5	−83.7	−81.3	71.2
5	800	80	0.9	−41.8	−47.4	−51.0	−40.1	53.3
5	800	150	0.0	−8.7	−14.7	−16.9	−7.9	27.4
5	1,500	15	−0.0	−81.7	−85.3	−84.3	−81.6	71.3
5	1,500	80	0.2	−22.8	−28.9	−27.7	−22.7	45.6
5	1,500	150	0.4	2.5	−9.4	−10.2	4.4	20.7

Note. OLS = ordinary least squares; K = number of topics; D = number of documents; $\bar{N}_d$ = average document length (words); η_X = regression coefficient for predictor X; $\eta_{\bar{Z}_1}$ = regression coefficient for topic 1; $\eta_{\bar{Z}_2}$ = regression coefficient for topic 2; $\eta_{\bar{Z}_4}$ = regression coefficient for topic 4; $\eta_{\bar{Z}_5}$ = regression coefficient for topic 5; σ^2 = residual variance. Note that for $\eta_{\bar{Z}_3} = 0$, relative bias is undefined so it is not included.

$\bar{N}_d = 150$). For these conditions, NRMSE for SLDAX was always smaller than NRMSE for the fixed α two-stage approach, suggesting again that SLDAX yielded more efficient estimates than the two-stage approach.

Empirical Example

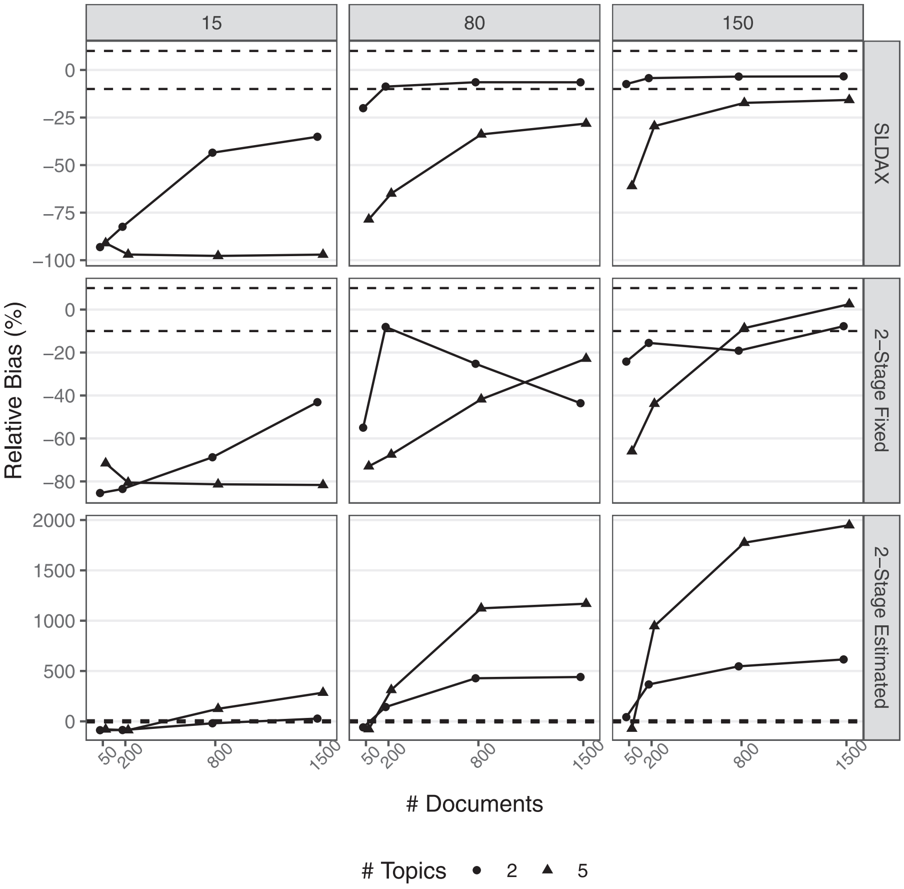
To illustrate the estimation and interpretation of a SLDAX model, we apply SLDAX to interpersonal interview data from a study of

Table 3
Relative Bias (%) for SLDAX

K	D	$\bar{N}_d$	η_X	$\eta_{\bar{Z}_1}$	$\eta_{\bar{Z}_2}$	$\eta_{\bar{Z}_4}$	$\eta_{\bar{Z}_5}$	σ^2
2	50	15	0.7	−93.2	−93.9			76.6
2	50	80	5.9	−20.1	−18.9			26.9
2	50	150	3.4	−7.4	−8.6			13.4
2	200	15	0.0	−82.6	−82.9			68.9
2	200	80	−0.6	−8.7	−9.7			13.4
2	200	150	0.5	−4.3	−4.5			7.7
2	800	15	−0.2	−43.3	−42.9			48.1
2	800	80	−0.3	−6.5	−6.4			9.2
2	800	150	−0.2	−3.5	−3.7			4.6
2	1,500	15	−0.4	−35.0	−34.8			40.7
2	1,500	80	0.0	−6.5	−6.6			8.8
2	1,500	150	0.2	−3.4	−3.3			4.9
5	50	15	−3.5	−91.0	−92.5	−91.4	−90.6	79.9
5	50	80	1.5	−78.3	−81.9	−82.7	−79.4	71.9
5	50	150	1.1	−61.0	−68.1	−70.5	−60.2	62.1
5	200	15	−0.5	−97.0	−97.4	−97.7	−96.9	74.7
5	200	80	−1.5	−65.5	−73.8	−73.6	−65.0	61.7
5	200	150	0.1	−29.4	−28.8	−30.1	−28.1	34.3
5	800	15	−0.1	−97.7	−97.7	−98.5	−98.2	71.6
5	800	80	0.8	−33.7	−32.7	−34.3	−33.3	39.3
5	800	150	−0.8	−17.4	−15.9	−17.4	−16.6	21.5
5	1,500	15	1.0	−97.1	−97.7	−97.8	−97.4	70.6
5	1,500	80	0.3	−28.3	−29.2	−30.0	−28.3	34.3
5	1,500	150	−0.5	−15.8	−15.5	−16.2	−15.3	20.0

Note. K = number of topics; D = number of documents; $\bar{N}_d$ = average document length (words); η_X = regression coefficient for predictor X; $\eta_{\bar{Z}_1}$ = regression coefficient for topic 1; $\eta_{\bar{Z}_2}$ = regression coefficient for topic 2; $\eta_{\bar{Z}_4}$ = regression coefficient for topic 4; $\eta_{\bar{Z}_5}$ = regression coefficient for topic 5; σ^2 = residual variance; SLDAX = supervised latent Dirichlet allocation with covariates. Note that for $\eta_{\bar{Z}_3} = 0$, relative bias is undefined so it is not included.

Figure 3
Relative Bias for the First Topic Regression Coefficient



Note. The left, center, and right columns correspond to average document lengths of 15, 80, and 150 words, respectively. Horizontal dashed lines provide reference for ±10% relative bias. Note that the scale of the ordinate axis varies across the three modeling approaches. Two-stage results using ordinary least squares regression are shown. Two-stage results using Bayesian regression were virtually identical. SLDAX = supervised latent Dirichlet allocation with covariates.

interpersonal stress and nonsuicidal self-injury (NSSI; Ammerman et al., 2021). NSSI is the deliberate, self-inflicted damage of body tissue without suicidal intent (Nock, 2009). We also compare results from a two-stage approach with fixed α and OLS regression. Participants were 41 undergraduate students with ages between 18 and 39 ($M = 20.8$, $SD = 3.9$); the majority (84%) identified as female; 23 participants (56%) reported a history of NSSI. Participants completed a semistructured interview regarding a recent upsetting ("negative") interpersonal interaction that occurred with someone with whom they have an ongoing relationship—the same procedure has been previously utilized on research examining interpersonal distress among high-risk populations (Gratz et al., 2011). Participants were asked to describe details of the interaction such as the other person(s) involved, the environment, and their feelings and thoughts. History of NSSI was determined using the Inventory of Statements about Self-Injury (ISAS; Klonsky & Glenn, 2009). Emotional dysregulation was measured by the Difficulties in Emotion Regulation Scale (DERS; Gratz & Roemer, 2004) where higher scores indicate greater dysregulation ($M = 85.9$, $SD = 22.6$). Participants were also asked to rate how upsetting or distressing the described interaction was on a Likert scale from 1 (*not at all upsetting or distressing*) to 10 (*most upset or distressed I've been;* $M = 6.5$, $SD = 1.5$).

The narratives were preprocessed before further analysis following standard practices in computational linguistics (e.g., Manning et al., 2008; M. E. Roberts et al., 2014). First, punctuation, numbers, and other nonletter characters were removed. Second, in order to focus the analysis on informative content words instead of highly common but topic-irrelevant words (e.g., "and," "the"), stop words from the NLTK Python library list available in the *stopwords* (Version 2.0; Benoit et al., 2020) R package were removed. Finally, any words that only occurred once in the corpus were removed. After preprocessing, the median narrative length was 63 words ($M = 64$, $SD = 17$, range = 40–114) and included 318 unique words.

The number of topics for the measurement model of the narratives was chosen by evaluating model fit for a sequence of LDA models with two to five topics. Each model was fit with α fixed equal to 1. Model fit was assessed using two metrics, *coherence* and *exclusivity*. Coherence (Mimno et al., 2011) measures how frequently the most probable words for a given topic co-occur within each document relative to their marginal prevalence in the corpus (i.e., how often the

Table 4
NRMSE (%) for the Two-Stage Approach With Estimated α and OLS Regression

K	D	$\overline{N}_d$	η_X	$\eta_{\overline{Z}_1}$	$\eta_{\overline{Z}_2}$	$\eta_{\overline{Z}_4}$	$\eta_{\overline{Z}_5}$	σ^2
2	50	15	37.75	89.29	86.43			75.94
2	50	80	32.79	65.68	67.28			56.60
2	50	150	30.96	154.16	157.54			54.54
2	200	15	16.55	87.37	86.43			69.89
2	200	80	17.11	211.35	211.27			64.68
2	200	150	18.18	451.25	451.05			61.96
2	800	15	8.13	84.17	84.53			69.42
2	800	80	8.43	532.18	531.87			65.97
2	800	150	8.13	634.81	634.50			63.17
2	1,500	15	6.63	102.80	102.61			69.84
2	1,500	80	6.21	544.37	544.89			65.97
2	1,500	150	6.31	709.11	708.75			60.73
5	50	15	36.10	81.15	84.13	86.47	80.71	80.77
5	50	80	34.89	78.04	82.58	83.38	79.59	71.50
5	50	150	37.15	73.39	78.35	78.90	73.05	68.67
5	200	15	16.58	87.91	90.63	89.24	86.68	74.94
5	200	80	15.21	444.65	421.77	413.85	424.00	68.33
5	200	150	16.96	1,095.52	1,069.10	1,004.01	1,096.82	64.89
5	800	15	7.85	223.38	205.36	201.65	221.05	71.85
5	800	80	8.80	1,242.54	1,174.10	1,207.70	1,295.82	67.97
5	800	150	8.41	1,926.22	1,951.85	1,836.34	1,985.50	66.48
5	1,500	15	6.39	326.60	308.01	294.20	326.23	71.71
5	1,500	80	5.80	1,289.53	1,155.55	1,158.40	1,287.04	68.01
5	1,500	150	6.35	2,112.34	2,004.36	2,098.87	2,131.45	64.32

Note. NRMSE = normalized rootmean squared error; OLS = ordinary least squares regression; K = number of topics; D = number of documents; $\overline{N}_d$ = average document length (words); η_X = regression coefficient for predictor X; $\eta_{\overline{Z}_1}$ = regression coefficient for topic 1; $\eta_{\overline{Z}_2}$ = regression coefficient for topic 2; $\eta_{\overline{Z}_4}$ = regression coefficient for topic 4; $\eta_{\overline{Z}_5}$ = regression coefficient for topic 5; σ^2 = residual variance. Note that for $\eta_{\overline{Z}_3} = 0$, NRMSE is undefined so it is not included.

most representative words for each topic co-occur relative to their overall frequency). Coherence is a popular metric because it has been shown to be positively associated with higher human ratings of topic quality (Mimno et al., 2011). However, M. E. Roberts et al. (2014) pointed out that coherence does not consider how different topics are from one another and, therefore, suggest simultaneously examining exclusivity. Exclusivity (M. E. Roberts et al., 2019) measures the extent to which the probabilities of words in a given topic

Table 5
NRMSE (%) for the Two-Stage Approach With Fixed α and OLS Regression

K	D	$\overline{N}_d$	η_X	$\eta_{\overline{Z}_1}$	$\eta_{\overline{Z}_2}$	$\eta_{\overline{Z}_4}$	$\eta_{\overline{Z}_5}$	σ^2
2	50	15	37.75	87.24	87.75			75.77
2	50	80	32.22	61.47	62.86			54.28
2	50	150	29.59	33.27	35.63			28.71
2	200	15	16.67	84.68	83.80			69.37
2	200	80	13.96	20.12	20.34			19.00
2	200	150	15.29	27.13	26.88			31.17
2	800	15	8.05	72.92	72.72			65.02
2	800	80	7.43	44.68	44.29			44.78
2	800	150	7.97	49.71	49.40			60.22
2	1,500	15	6.16	54.57	54.32			55.69
2	1,500	80	5.94	55.72	55.26			61.53
2	1,500	150	6.54	49.43	50.01			59.59
5	50	15	36.84	73.13	78.69	78.39	74.30	81.32
5	50	80	36.69	74.09	80.90	80.29	75.48	71.39
5	50	150	36.77	67.61	74.95	75.38	68.14	66.64
5	200	15	16.18	80.98	84.78	83.03	80.53	75.33
5	200	80	15.16	68.88	74.73	76.04	68.64	65.20
5	200	150	16.53	47.28	59.24	59.92	47.61	49.34
5	800	15	7.86	81.78	86.15	84.64	81.83	71.74
5	800	80	9.04	47.02	58.84	60.89	46.06	54.29
5	800	150	7.28	23.86	43.50	44.46	23.80	30.52
5	1,500	15	6.39	82.00	85.95	84.90	82.10	71.55
5	1,500	80	5.38	34.15	47.51	44.90	32.38	47.06
5	1,500	150	5.35	20.22	39.23	42.22	19.58	23.27

Note. NRMSE = normalized rootmean squared error; OLS = ordinary least squares; K = number of topics; D = number of documents; $\overline{N}_d$ = average document length (words); η_X = regression coefficient for predictor X; $\eta_{\overline{Z}_1}$ = regression coefficient for topic 1; $\eta_{\overline{Z}_2}$ = regression coefficient for topic 2; $\eta_{\overline{Z}_4}$ = regression coefficient for topic 4; $\eta_{\overline{Z}_5}$ = regression coefficient for topic 5; σ^2 = residual variance. Note that for $\eta_{\overline{Z}_3} = 0$, NRMSE is undefined so it is not included.

Table 6
NRMSE (%) for the SLDAX Model

K	D	$\bar{N}_d$	η_X	$\eta_{\bar{Z}_1}$	$\eta_{\bar{Z}_2}$	$\eta_{\bar{Z}_4}$	$\eta_{\bar{Z}_5}$	σ^2
2	50	15	37.26	94.26	94.88			82.44
2	50	80	29.44	29.01	28.97			36.26
2	50	150	29.36	21.21	22.28			26.69
2	200	15	16.35	83.21	83.67			70.88
2	200	80	13.15	13.18	13.15			17.09
2	200	150	13.34	10.92	10.97			12.96
2	800	15	7.20	44.11	43.71			48.75
2	800	80	6.99	8.26	8.02			10.88
2	800	150	5.94	6.12	6.24			6.86
2	1,500	15	5.29	35.51	35.31			41.17
2	1,500	80	4.81	7.60	7.43			9.63
2	1,500	150	4.88	5.01	4.88			5.99
5	50	15	36.98	91.48	93.67	92.64	90.98	88.45
5	50	80	34.00	79.17	83.63	85.17	80.34	77.88
5	50	150	35.48	63.30	71.47	74.94	62.83	70.67
5	200	15	14.68	97.22	98.08	97.81	96.90	75.91
5	200	80	15.81	66.92	75.33	74.55	67.15	63.06
5	200	150	15.04	31.44	37.77	38.72	31.31	36.34
5	800	15	7.98	97.74	97.71	98.70	98.24	72.16
5	800	80	7.51	34.72	35.05	36.80	34.17	39.92
5	800	150	6.59	18.45	20.83	21.94	17.94	22.19
5	1,500	15	6.43	97.06	97.64	97.90	97.39	70.96
5	1,500	80	5.53	28.83	30.03	31.27	28.90	34.67
5	1,500	150	4.92	16.55	17.82	18.23	16.10	20.54

Note. NRMSE = normalized rootmean squared error; SLDAX = supervised latent Dirichlet allocation with covariates; K = number of topics; D = number of documents; $\bar{N}_d$ = average document length (words); η_X = regression coefficient for predictor X; $\eta_{\bar{Z}_1}$ = regression coefficient for topic 1; $\eta_{\bar{Z}_2}$ = regression coefficient for topic 2; $\eta_{\bar{Z}_4}$ = regression coefficient for topic 4; $\eta_{\bar{Z}_5}$ = regression coefficient for topic 5; σ^2 = residual variance. Note that for $\eta_{\bar{Z}_3} = 0$, NRMSE is undefined so it is not included.

differ from those in other topics. Taken together, an optimal model should maximize both coherence and exclusivity without favoring one metric over the other (i.e., high coherence with low exclusivity or low coherence with high exclusivity). For these data, a four-topic LDA model optimized coherence and exclusivity—this is consistent with the dimensionality of the topic space in other applications of topic modeling in psychology (e.g., three-topic solutions were found to be optimal by Finch et al., 2018; Kim, Kwak, Cardozo-Gaibisso, et al., 2017; M. E. Roberts et al., 2014).

The most strongly representative words measured by term score (i.e., a more interpretable summary of topics than the original word probabilities that reweights the word probabilities per topic to emphasize words that "load" more strongly on a given topic over words that have similar probabilities across all topics; Blei & Lafferty, 2009) associated with each topic are shown in Figure 5—because the word probabilities were similar for LDA and the SLDAX model we discuss below, we only present the SLDAX-based word term scores. Topic 1 corresponded to interpersonal conflict regarding romantic relationships; Topic 2 corresponded to interpersonal conflict with and/or concern regarding family members; Topic 3 corresponded to interpersonal conflict with friends and peers; and Topic 4 corresponded to interpersonal conflict regarding shared living spaces. Here, we consider whether these topics are significant predictors of emotional dysregulation above and beyond the participant self-ratings of distress and NSSI history.

Having chosen the number of topics for the measurement model of the free response, we fit an SLDAX model with subjective rating (mean-centered) and NSSI history (coded as "no" = −.5, "yes" = .5) and four topics to model emotional dysregulation. For comparison, we used the topic proportion estimates $\hat{\Theta}$ from the four-topic LDA model as predictors in an ordinary least squares regression model along with the subjective ratings and NSSI history to model emotional dysregulation. We used the same diffuse priors for the SLDAX model as were used in the simulation study.[19] Convergence was evaluated using trace plots and the Geweke (1992) and Heidelberger and Welch (1983) statistics for each parameter. We used a burn-in of 15,000 iterations and a thinning period of 10, resulting in a total of 18,500 posterior samples. The posterior samples were permuted using Stephens (2000) algorithm and trace plots were examined to address possible label switching before posterior summary statistics were computed.

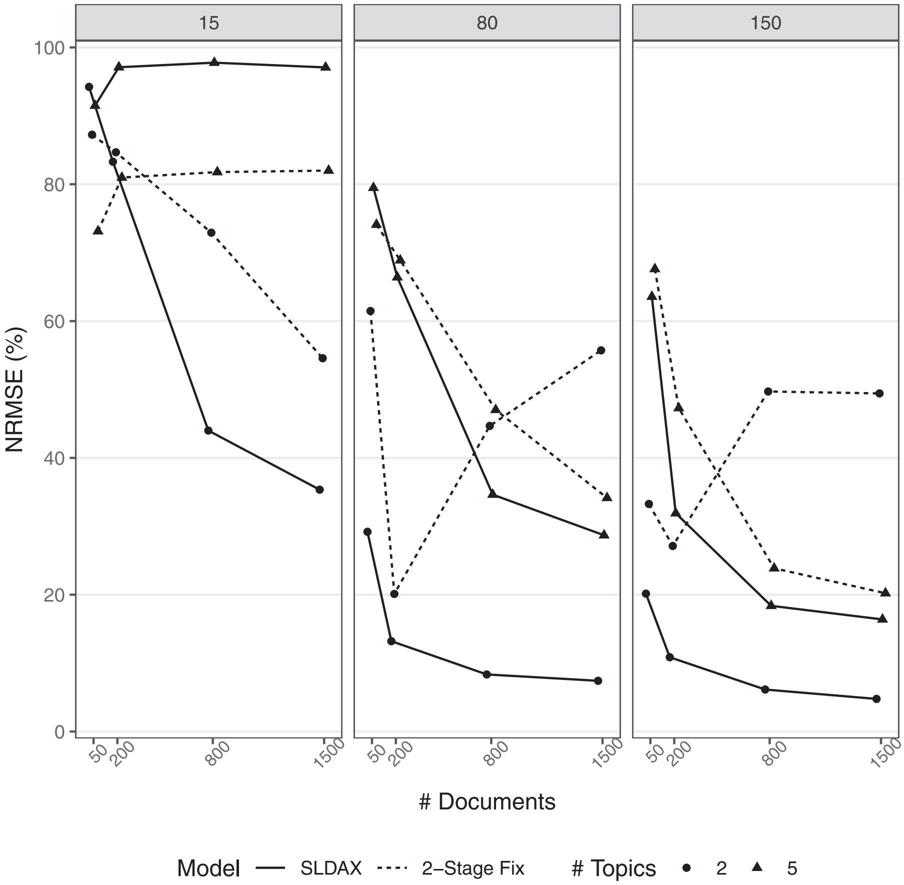
Example excerpts from participant narrative transcripts are provided in Table 7 along with their corresponding topic proportion estimates to further aid interpretation of the topics. Each narrative is assigned its own set of topic proportions, allowing different parts of a given narrative to be modeled by an appropriate topic.

The results of the SLDAX and two-stage approaches are summarized in Table 8. First, participants with a history of NSSI reported significantly higher emotional dysregulation than those with no history using both methods. While the estimated topic regression coefficients varied around the marginal mean of 84.6, the two-stage estimates tended to be further from the marginal mean than the SLDAX estimates, which is consistent with our simulation results, and the corresponding standard errors were smaller for SLDAX, as expected. As a consequence of the difference in coefficient estimates, the Piepel (1982) contrast estimates $\hat{c}_k^{(P)}$ were smaller using SLDAX than the two-stage approach. The associated standard errors for the Piepel contrasts were larger using SLDAX because

[19] Results were robust to other prior specifications.

Figure 4

NRMSE (%) for the First Topic Regression Coefficient



Note. The left, center, and right columns correspond to average document lengths of 15, 80, and 150 words, respectively. Two-stage results using ordinary least squares regression are shown. Two-stage results using Bayesian regression were virtually identical. SLDAX = supervised latent Dirichlet allocation with covariates; NRMSE = normalized rootmean squared error.

SLDAX uses the full posterior of the empirical topic proportions whereas the two-stage approach treats the topic proportion estimates as fixed. For SLDAX, this manifested as a larger range of proportions which can increase the standard error estimates of the Piepel topic effects. Taken together, the smaller estimates and larger standard errors of the topic effects for SLDAX yielded nonsignificant individual relationships among the topics and emotional dysregulation whereas the two-stage approach yielded two significant relationships although the direction of the effects is the same.

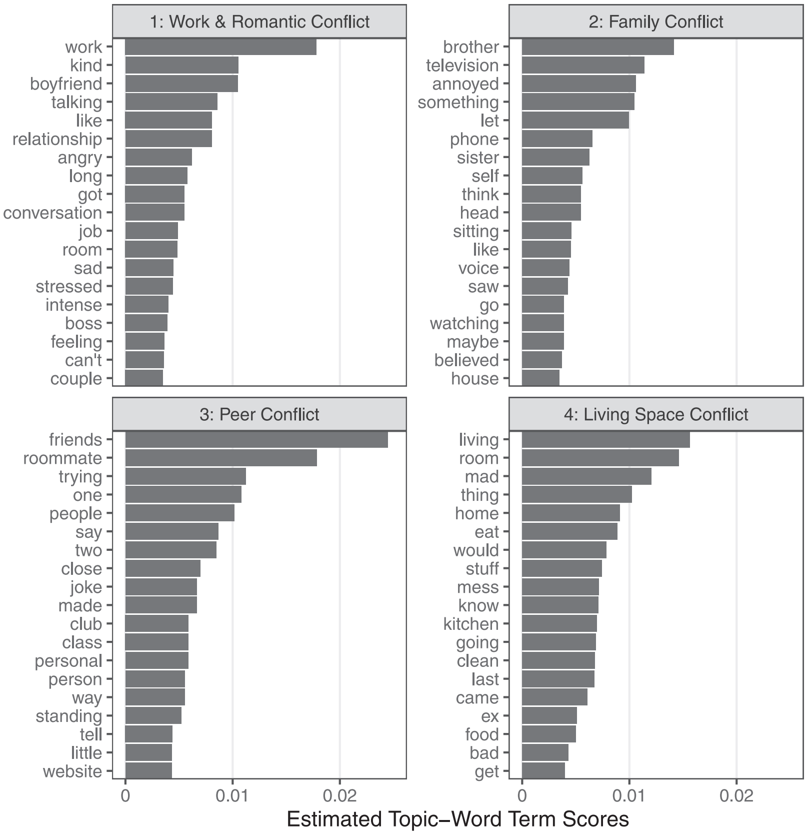
Although we do not know the "truth" in this empirical example, the direction of both sets of estimates are consistent. Because our simulation results suggested that the two-stage fixed α approach can yield inconsistent estimates of the topic regression coefficients, it is possible that the two-stage estimates here may be incorrect. Given that the simulation results for similar conditions suggest that the SLDAX topic regression coefficient estimates may be attenuated, we suspect that the SLDAX topic effect estimates in this case may be underestimated; the number of participants is limited, so the credible intervals are wide. However, we see that narratives whose language had higher prevalence of Topic 1 (T1)—work conflict—and Topic 3 (T3)—peer conflict—were associated with higher emotional dysregulation whereas higher prevalence of Topic 2 (T2)—family conflict—and Topic 4 (T4)—living space conflict—was associated with lower emotional dysregulation. Finally, we examined the joint contribution of the topics to modeling emotional dysregulation above and beyond the other predictors. The self-ratings and NSSI history variables only explained 24% of the variability in emotional dysregulation while the four topics explained an additional 15%, with an overall R^2 (Gelman et al., 2019) of 39% for the SLDAX model.

Finally, as suggested by a reviewer, we fit (a) a SLDA model with only four topics as predictors and (b) a regression model with only subjective rating and NSSI history as predictors. This allowed us to empirically demonstrate differences in estimates, standard errors, and inferences that can occur if a simpler model (SLDA or regression with manifest variables only) is used instead of the full SLDAX model. Results are shown in Table 8.

First, suppose a researcher only used an SLDA model in this example. The topic regression coefficients estimates (T1–T4) were comparable with the inclusion of the two manifest predictors in the SLDAX model yielding slightly smaller (i.e., closer to the

Figure 5

Fifteen Largest Estimated Word Term Scores per Topic From the SLDAX Model of the Interpersonal Narratives



Note. SLDAX = supervised latent Dirichlet allocation with covariates.

overall mean of the outcome) coefficient estimates by .2%−2.8%. However, the more important and interpretable topic effects (T1 effect–T4 effect) differed more substantially between the two approaches. The adjusted effects of Topics 2, 3, and 4 were 1.7%−21.6% smaller using the SLDAX model compared with the unadjusted topic effects of the SLDA model; the adjusted effect of Topic 1 from the SLDAX model was 1.6% larger than the unadjusted effect from the SLDA model. The standard error estimates of the topic coefficient and topic effect estimates also differed between the SLDAX and SLDA models. Both the standard error estimates of the topic coefficient estimates and the topic effect estimates were smaller (9.9%−13.7% for the topic coefficients and 18.9%−24.5% for the topic effects) using the SLDAX model. These results illustrate that the inclusion or omission of manifest predictors can impact the estimated topic coefficients, effects, and standard errors as SLDAX yields adjusted estimates and standard errors whereas SLDA yields unadjusted estimates and standard errors. Consequently, the SLDAX model can yield greater power than SLDA as in this example.

Second, suppose a researcher only used the manifest predictors without including information from the narratives. As in the previous case, the manifest-only regression coefficient estimates differ from the corresponding estimates from the SLDAX model (which are adjusted for the topics). SLDAX provides greater power than the manifest-only model by yielding 32.1% and .5% smaller standard error estimates.

The differences between the SLDA, manifest-only, and SLDAX estimates may be explained by the correlation among topics and manifest variables or the reduction of residual variance. The absolute value of correlations between subject rating and topics ranged from .02 to .14 while absolute correlations between NSSI history and topics ranged from .01 to .06, which could account for the difference between sets of regression coefficient estimates and standard errors. If these two sets of predictors were perfectly uncorrelated, standard regression results show that the regression coefficient estimates will be asymptotically equivalent (see, e.g., Rencher & Schaalje, 2008). In this empirical example, even relatively small correlations between manifest predictors and topics yielded noticeable differences in coefficient and standard error estimates. It is important to note that the interpretation of these regression relationships also changes between model specifications. SLDA yields topic regression effects that are unadjusted for manifest predictors, whereas SLDAX yields topic regression effects that are adjusted for the manifest predictors. Similarly, a manifest-only regression model yields estimates that are unadjusted for the topic effects whereas SLDAX yields estimates that are adjusted for the topic effects.

Table 7
Representative Responses for SLDAX Topics

NSSI	DERS	Self-rating	Narrative transcript excerpts	T1 Pr.	T2 Pr.	T3 Pr.	T4 Pr.
No	73	7	Dad came to visit her at work at retail place; didn't tell her, just showed up; . . . had a conversation with her boss about politics for 30 min; embarrassed and angry he did that at work . . . she tried to be distant stay engrossed in the work.	0.71	0.11	0.06	0.12
Yes	107	6	Her and boyfriend had argument about being in a long distance relationship; . . . for his job he will need to be away for months at a time; . . . she wants to move after school and he has to stay for his job.	0.60	0.11	0.09	0.20
Yes	59	2	Watching TV with brother; he saw something on the TV he didn't like; let that get inside his head and upset him; annoyed . . . that he let TV change his mood; . . . too wrapped up in his head to listen; . . . at their parents' house, brother started talking about it.	0.10	0.81	0.04	0.06
No	85	9	Mom and dad just got a divorce; brother was put in an awkward situation with another guy her mom is seeing; brother talked to her and then she had a discussion/argument with mom; . . . mom wasn't really reacting, was pretty upset by it with mom.	0.17	0.65	0.06	0.11
No	83	7	Friend will always make jokes about her [body]; . . . female friend said this to two of their mutual friends . . . in class; . . . don't like when people comment on bodies, class was starting so didn't say anything . . . a little sad and hurt.	0.11	0.09	0.70	0.11
Yes	108	7	Hanging out with roommate and best friend and other friends; all trying to give advice to other friend about her roommate; [friend] . . . cracked a joke that felt more like a jab; . . . felt very insulting.	0.09	0.05	0.80	0.07
Yes	63	7	Came home and his roommates had friends over and they . . . left a mess and never cleaned it; . . . mess still in the kitchen; sort of a typical thing.	0.07	0.07	0.08	0.78
No	53	6	Her ex-roommate was really messy; . . . he trashed the house and she was pissed; the bathroom was really dirty and there was a lot of bad food in the fridge; fridge smelled bad.	0.06	0.04	0.06	0.83

Note. Partial excerpts from participant narrative transcripts with high probabilities (>.6) for one of the four topics are shown along with participants' NSSI history (NSSI), emotional dysregulation (DERS), self-rating (Self-rating) of distress, and the estimated topic proportions (e.g., T1 Pr.) are shown. The first two rows correspond to narratives with a high proportion for Topic 1, the third and fourth rows correspond to narratives with a high proportion for Topic 2, the fifth and sixth rows correspond to narratives with a high proportion of Topic 3, and the last two rows correspond to narratives with a high proportion of Topic 4. SLDAX = supervised latent Dirichlet allocation with covariates; NSSI = nonsuicidal self-injury; DERS = Difficulties in Emotion Regulation Scale.

Discussion

Despite the growing interest in and availability of text data in psychological research (e.g., Garten et al., 2018; Obeid et al., 2019; Popping, 2015; Schwartz et al., 2013), the development of appropriate statistical models to connect text data to psychological theory has been limited. Recently, topic models have been particularly popular, especially when researchers are interested in using the latent topic estimates as predictors of other psychological constructs (Finch et al., 2018; Kim, Kwak, Cardozo-Gaibisso, et al., 2017; Kim, Kwak, & Cohen, 2017; Packard & Berger, 2020; Rohrer et al., 2017; Schwartz et al., 2013). However, the two-stage approach used in these applications has not been systematically evaluated and alternative modeling approaches have been unavailable. This article demonstrated through a simulation study that two common two-stage approaches can have undesirable statistical behavior: If the α hyperparameter governing the topic proportions is estimated, estimates of the topic regression coefficients can be substantially biased, whereas if the α hyperparameter is fixed, the estimates of the topic regression coefficients can be unpredictably biased depending on model complexity and sample size. Further, the choice of OLS and Bayesian regression with diffuse priors in the two-stage framework did not significantly impact the accuracy or precision of the two-stage methods. To resolve these problems, we introduced a novel generalization of the LDA topic model, SLDAX, as a one-stage alternative to the two-stage approaches. We described a Gibbs sampling algorithm to estimate the SLDAX model for continuous and dichotomous outcomes. The SLDAX model has several advantages over two-stage procedures: It (a) has a fully Bayesian foundation; (b) can be easily extended to model other data characteristics and prior information; (c) yields Bayesian posterior estimates of model parameters and functions of model parameters; (d) accounts for the uncertainty in the topic proportions used to predict the outcome; and (e) simultaneously provides a measurement model of text and a structural model of relationships between an outcome and both the latent topics and manifest covariates. Furthermore, the computation time for SLDAX is comparable with the two-stage procedures because the computational complexity of both approaches is dominated by estimating the topic proportion and topic-word probability matrices, whereas the estimation of the regression parameters is negligible in comparison.

Table 8

SLDAX, Two-Stage, SLDA, and Manifest Variable Model Estimates of Emotional Dysregulation

Method	Variable	$\hat{\eta}_j$	*SE*	95% CI
SLDAX	Self-rating	0.66	2.22	[−3.73, 5.01]
	NSSI history	21.66	6.48	[8.84, 34.45]
	T1	92.31	10.42	[71.45, 112.58]
	T2	66.98	11.94	[43.19, 90.72]
	T3	100.95	10.30	[80.74, 121.76]
	T4	75.39	11.52	[52.09, 97.83]
	T1 effect	9.72	12.29	[−15.12, 33.71]
	T2 effect	−20.44	13.31	[−46.32, 6.39]
	T3 effect	19.48	11.57	[−3.47, 42.33]
	T4 effect	−10.80	13.06	[−36.75, 14.87]
Two-stage	Self-rating	0.25	1.91	[−3.64, 4.13]
	NSSI history	21.30	5.61	[9.91, 32.69]
	T1	109.70	12.06	[85.21, 134.20]
	T2	49.70	12.51	[24.32, 75.09]
	T3	108.02	9.96	[87.81, 128.23]
	T4	65.18	11.98	[40.85, 89.50]
	T1 effect	32.42	9.96	[12.20, 52.64]
	T2 effect	−20.78	11.39	[−43.91, 2.35]
	T3 effect	35.24	9.13	[16.72, 53.77]
	T4 effect	−6.20	11.33	[−29.19, 16.80]
Manifest only	Self-rating	1.20	3.27	[−3.17, 5.54]
	NSSI history	21.94	6.51	[9.11, 34.91]
SLDA	T1	92.46	11.96	[68.17, 115.89]
	T2	67.71	13.53	[41.17, 94.34]
	T3	103.91	11.94	[80.14, 127.05]
	T4	77.04	12.78	[51.30, 101.80]
	T1 effect	9.57	16.06	[−23.43, 40.44]
	T2 effect	−23.42	16.75	[−55.62, 10.36]
	T3 effect	24.84	15.05	[−5.54, 53.82]
	T4 effect	−10.99	16.10	[−42.90, 20.59]

Note. Parameter posterior mean estimates ($\hat{\eta}_j$), posterior standard deviations (*SE*), 95% credible intervals (BCI) for supervised latent Dirichlet allocation with covariates (SLDAX) or 95% confidence intervals for two-stage estimation. Topic effect estimates $c_k^{(P)}$ were obtained using Piepel's (1982) method with empirical topic proportions for SLDAX and the topic proportion estimates for two-stage estimation. SLDA = supervised latent Dirichlet allocation model; NSSI = nonsuicidal self-injury.

We showed by simulation that the SLDAX estimates were less biased and more efficient than the two-stage approaches. Furthermore, the SLDAX estimates were ignorably biased when the average document length and number of subjects were both "large" when the two-stage estimates could be unacceptably biased. We also describe methodology for correctly interpreting, estimating, and testing relationships between latent topics and an outcome that differ from standard applied practice (see, e.g., Blei & McAuliffe, 2008; Packard & Berger, 2020; Rohrer et al., 2017; Schwartz et al., 2013). Our empirical example demonstrated the use of the SLDAX model as a means of jointly assessing several research questions. First, we obtained a concise person-specific measurement model of free responses. Second, we were able to estimate and test the relationships between the latent topics from interpersonal participant narratives and emotional dysregulation while controlling for other clinical factors and vice versa. Importantly, the use of the SLDAX model allows both tasks to be completed simultaneously, avoiding the limitations of a two-stage procedure. As an added advantage, we demonstrated that the unexplained variance in the outcome can be reduced by incorporating the topics.

This article was a first step in developing the SLDAX model and evaluating the statistical performance of SLDAX and two-stage approaches, so it was necessarily limited in scope. First, particularly for more complex models and shorter documents, we found that the topic regression coefficient estimates can be attenuated using either SLDAX or the fixed α two-stage approach. The attenuation of the topic regression coefficients is an interesting phenomenon that has gone unstudied within the topic modeling literature. This suggests that the reliability of the topic regression coefficient estimates is governed by the average document length. Consistent with our simulation results, we expect reliability to improve as the amount of available information in a given document grows larger, thereby improving the estimation of the topic regression relationships. We suspect that the attenuation of the topic regression coefficients may be a result of latent classification error as seen in other latent class models (Croon, 2002). We conjecture that when the average document contains few construct-relevant words, it is difficult to accurately estimate the underlying topics, which can lead to misclassification of the topics.

Second, it is difficult to quantify what constitutes "small" text data. In the cross-sectional context explored in this article, we emphasized two features of data quantity: the number of subjects and the length of the documents. We considered a reasonable set of values for both factors consistent with empirical text data in psychological applications, but the performance of SLDAX and other models for text should be further evaluated under a wider variety of data and model conditions. Our results suggested that both factors of data size are important. When faced with temporal and financial constraints, however, researchers may need to optimize one or the other. For the SLDAX model, our simulation suggests that the average document length may be the more critical of the two for obtaining accurate regression estimates. However, although not emphasized in this paper, power for testing these regression relationships will depend on the number of subjects. Whether SLDAX and related models are appropriate in practice will depend on the length of documents or responses and the complexity of the underlying latent topic structure. In low-stakes applications (e.g., survey responses), responses may be too short to support a topic model, so analysis of a small subset of word frequencies may be preferable. In high-stakes applications like standardized testing or clinical interviews, responses may be longer and could support a richer analysis by topic modeling. Therefore, it is crucial during study design to have a sense of the length and complexity of text that can be collected given the population and item prompts of interest before a particular statistical method is chosen. One potential solution to these small-sample limitations is readily available within the SLDAX framework. Bayesian methods have been advocated for handling small-sample problems (Baldwin & Fellingham, 2013; Depaoli, 2014; Miočević et al., 2017; van de Schoot et al., 2015; Zondervan-Zwijnenburg et al., 2017) because well-chosen prior distributions can improve model estimation. In our simulation study, we only studied diffuse priors in order to (a) match the default choice in most Bayesian software applications and (b) focus on model performance as a function of the data characteristics, not the prior specification. Some studies have evaluated the impact of alternative prior specifications for topic models (Blei & McAuliffe, 2008; Magnusson et al., 2020; Perotte et al., 2011; Zhu et al., 2013), but further study is needed.

Third, we note that the proposed Gibbs sampling algorithm may not scale well to massive text data sets. Our experience in fitting SLDAX models to several psychological data sets suggests that the proposed MCMC algorithm can be fit in minutes on a personal computer for the data characteristics explored in this article. However, the computational complexity scales directly with both the number of subjects and, in particular, the document lengths. For massive data sets, alternative strategies such as parallelization of the sampler or algorithms such as variational EM may be preferable.

Fourth, much like a distal outcome mixture model or structural equation model, the choice of a one-stage versus. two-stage estimation approach can affect the latent topic space that is estimated. Including an outcome predicted by the topics in a one-stage model like SLDAX can affect the estimated topic measurement parameters, potentially resulting in interpretational confounding (Burt, 1973). This can be avoided by fixing the measurement model of the topics by estimating an unsupervised topic model and using a two-stage approach. However, as our simulation results demonstrated, currently available two-stage approaches for topic models can lead to biased regression coefficient estimates and incorrect standard errors. It would be useful to develop adjustments to correct the two-stage approach estimates, for example, along the lines of Bakk et al. (2013) and Vermunt (2010).

Finally, it is important to carefully consider the choice of model within the topic modeling family. In this article, we discussed several extensions of the original LDA model including SLDA, STM, the model proposed by Ansari et al. (2018), and SLDAX. While model selection approaches such as Bayes factors or information criterion could potentially be developed to assist with model selection among these models, this is a methodological line of research beyond the scope of this article. Notably, these models answer fundamentally different research questions. Focusing on STM and SLDAX, SLDAX is a cross-sectional model in which both covariates and topics (whose measurement by text is not assumed to be affected by the covariates) are used to predict an outcome. In effect, the research questions afforded by SLDAX involve the contribution of the topics above and beyond that of the covariates when trying to model an outcome and vice versa, not whether the covariates predict the topical content of the text data. STM is comparable with an MIMIC model (Joreskog & Goldberger, 1975) in which multiple covariates are used to predict latent topics and the topics, in turn, predict or cause multiple words. The two models also imply a different temporal or causal order of relationships. In SLDAX, the covariates and the topics are assumed to act concurrently as antecedents of the outcome. In STM, the predictors enter the model as antecedents of the topics. In psychological research, it may be difficult to obtain an appropriate temporal order to justify a causal model like STM as it is common for multiple measures to be collected concurrently. In this case, an SLDAX model may be more appropriate. If, however, text is collected after the antecedent covariates, then the STM would be preferable. Ultimately, the choice of model should reflect the research questions and design.

Conclusion

Our article proposed a new measurement and structural modeling approach to allow researchers model an outcome using both qualitative and quantitative data as predictors. Our simulation study suggested that the proposed model offers more accurate and efficient estimates than conventional two-stage approaches, and, consequently, better inferences to psychological researchers interested in studying the predictive relationships from text-based latent topics and an outcome while controlling for other covariates of interest and vice versa. The *psychtm* R package (Wilcox, 2021) provides a free, open-source software implementation of the SLDAX model. We hope that this article stimulates further development of statistical methodology for mixed-methods research that is tailored to the unique challenges and goals of psychological science.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899–917. https://doi.org/10.1037/amp0000190

Ammerman, B. A., Sorgi, K. M., Fahlgren, M. K., Puhalla, A. A., & McCloskey, M. S. (2021). An experimental examination of interpersonal problem-solving in nonsuicidal self-injury: A pilot study. *Journal of Psychiatric Research*, *144*, 146–150. https://doi.org/10.1016/j.jpsychires.2021.09.005

Ammerman, B. A., Wilcox, K. T., O’Loughlin, C. M., & McCloskey, M. S. (2021). Characterizing the choice to disclose nonsuicidal self-injury. *Journal of Clinical Psychology*, *77*(3), 683–700. https://doi.org/10.1002/jclp.23045

Ansari, A., Li, Y., & Zhang, J. Z. (2018). Probabilistic topic model for hybrid recommender systems: A stochastic variational Bayesian approach. *Marketing Science*, *37*(6), 987–1008. https://doi.org/10.1287/mksc.2018.1113

Baird, R., & Maxwell, S. E. (2016). Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. *Psychological Methods*, *21*(2), 175–188. https://doi.org/10.1037/met0000070

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272–311. https://doi.org/10.1177/0081175012470644

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*(2), 151–164. https://doi.org/10.1037/a0030642

Benoit, K., Muhr, D., & Watanabe, K. (2020). *Stopwords: Multilingual stopword lists* (Version 2.0). https://CRAN.R-project.org/package=stopwords

Blei, D. M., & Lafferty, J. D. (2006). Correlated topic models. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 1511–1518). MIT Press. https://proceedings.neurips.cc/paper/2005/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71–94). Chapman and Hall/CRC.

Blei, D. M., & McAuliffe, J. D. (2008). Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 121–128). Curran Associates. https://proceedings.neurips.cc/paper/2007/file/d56b9fc4b0f1be8871f5e1c40c0067e7-Paper.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27. https://doi.org/10.1093/pan/mph001

Burt, R. S. (1973). Confirmatory factor-analytic structures and the theory construction process. *Sociological Methods & Research*, *2*(2), 131–190. https://doi.org/10.1177/004912417300200201

Cassiday, K. R., Cho, Y., & Harring, J. R. (2021). A comparison of label switching algorithms in the context of growth mixture models. *Educational and Psychological Measurement*, *81*(4), 668–697. https://doi.org/10.1177/0013164420970614

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*(451), 957–970. https://doi.org/10.2307/2669477

Clemens, W. V. (1966). *An analytical and empirical examination of some properties of ipsative measures*. Psychometric Society. https://www.psychometricsociety.org/sites/main/files/file-attachments/mn14.pdf

Cohen, J. (1988). Multiple regression and correlation analysis. In *Statistical power analysis for the behavioral sciences* (2nd ed., pp. 407–467). Erlbaum.

Cornell, J. A. (2002). *Experiments with mixtures: Designs, models, and the analysis of mixture data* (3rd edition). Wiley.

Cox, D. R. (1971). A note on polynomial response functions for mixtures. *Biometrika*, *58*(1), 155–159. https://doi.org/10.1093/biomet/58.1.155

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195–224). Erlbaum.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 239–252. https://doi.org/10.1080/10705511.2014.882686

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770. https://doi.org/10.1177/0013164415607618

Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement*, *79*(6), 1017–1037. https://doi.org/10.1177/0013164419844552

Dias, J. G., & Wedel, M. (2004). An empirical comparison of EM, SEM and MCMC performance for problematic Gaussian mixture likelihoods. *Statistics and Computing*, *14*(4), 323–332. https://doi.org/10.1023/B:STCO.0000039481.32211.5a

Dickey, J. M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, *78*(383), 628–637. https://doi.org/10.2307/2288131

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*(2), 137–154. https://doi.org/10.1111/j.1745-3984.1998.tb00531.x

Finch, W. H., Finch, M. E. H., McIntosh, C. E., & Braun, C. (2018). The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science*, *4*(4), 403–424. https://doi.org/10.1037/tps0000173

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods*, *50*(1), 344–361. https://doi.org/10.3758/s13428-017-0875-9

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd). Chapman & Hall.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, *6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (4th ed., pp. 169–193). Oxford University Press.

Gratz, K. L., Hepworth, C., Tull, M. T., Paulson, A., Clarke, S., Remington, B., & Lejuez, C. (2011). An experimental investigation of emotional willingness and physical pain tolerance in deliberate self-harm: The moderating role of interpersonal distress. *Comprehensive Psychiatry*, *52*(1), 63–74. https://doi.org/10.1016/j.comppsych.2010.04.009

Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the Difficulties in Emotion Regulation Scale. *Journal of Psychopathology and Behavioral Assessment*, *26*(1), 41–54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*. Advance online publication. https://doi.org/10.18637/jss.v040.i13

Hayes, T., & Usami, S. (2020). Factor score regression in connected measurement models containing cross-loadings. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 1–10. https://doi.org/10.1080/10705511.2020.1729160

He, Q. (2013). *Text mining and IRT for psychiatric and psychological assessment* [Doctoral dissertation]. University of Twente. https://doi.org/10.3990/1.9789036500562

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*(6), 1109–1144. https://doi.org/10.1287/opre.31.6.1109

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–233). University of California Press.

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*(2), 265–290. https://doi.org/10.1017/langcog.2014.30

Jacobucci, R., Ammerman, B. A., & Wilcox, K. T. (2021). The use of text-based responses to improve our understanding and prediction of suicide risk. *Suicide and Life-Threatening Behavior*, *51*(1), 55–64. https://doi.org/10.1111/sltb.12668

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639. https://doi.org/10.2307/2285946

Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics*, *1*(1), 82–102. https://doi.org/10.37514/JWA-J.2017.1.1.05

Kim, S., Kwak, M., & Cohen, A. S. (2017). A mixture partial credit model analysis using language-based covariates. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology* (Vol. 196, pp. 321–333). Springer International Publishing. https://doi.org/10.1007/978-3-319-56294-0_28

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115. https://doi.org/10.1037/met0000191

Klonsky, E. D., & Glenn, C. R. (2009). Assessing the functions of non-suicidal self-injury: Psychometric properties of the Inventory of Statements About Self-injury (ISAS). *Journal of Psychopathology and Behavioral Assessment*, *31*(3), 215–219. https://doi.org/10.1007/s10862-008-9107-z

Kovacs, B., & Kleinbaum, A. M. (2020). Language-style similarity and social networks. *Psychological Science*, *31*(2), 202–213. https://doi.org/10.1177/0956797619894557

Levy, R. (2017). Distinguishing outcomes from indicators via Bayesian modeling. *Psychological Methods*, *22*(4), 632–648. https://doi.org/10.1037/met0000114

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, *89*(427), 958–966. https://doi.org/10.1080/01621459.1994.10476829

Lu, I. R. R., & Thomas, D. R. (2008). Avoiding and correcting bias in score-based latent variable regression with discrete manifest items. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 462–490. https://doi.org/10.1080/10705510802154323

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science+Business Media.

Magnusson, M., Jonsson, L., & Villani, M. (2020). DOLDA: A regularized supervised topic model for high-dimensional multi-class regression. *Computational Statistics*, *35*(1), 175–201. https://doi.org/10.1007/s00180-019-00891-1

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *An introduction to information retrieval*. Cambridge University.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman and Hall/CRC.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language* (pp. 262–272). Association for Computational Linguistics. https://doi.org/10.5555/2145432.2145462

Miočević, M., MacKinnon, D. P., & Levy, R. (2017). Power in Bayesian mediation analysis for small sample research. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 666–683. https://doi.org/10.1080/10705511.2017.1312407

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904

Nock, M. K. (2009). Why do people hurt themselves?: New insights into the nature and functions of self-injury. *Current Directions in Psychological Science*, *18*(2), 78–83. https://doi.org/10.1111/j.1467-8721.2009.01613.x

Obeid, J. S., Weeda, E. R., Matuskowitz, A. J., Gagnon, K., Crawford, T., Carr, C. M., & Frey, L. J. (2019). Automated detection of altered mental status in emergency department clinical notes: A deep learning approach. *BMC Medical Informatics and Decision Making*, *19*(1), 164. https://doi.org/10.1186/s12911-019-0894-9

Packard, G., & Berger, J. (2020). Thinking of you: How second-person pronouns shape cultural success. *Psychological Science*, *31*(4), 397–407. https://doi.org/10.1177/0956797620902380

Papastamoulis, P. (2016). Label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software*, *69*(1), 1–24. https://doi.org/10.18637/jss.v069.c01

Park, S. H. (1978). Selecting contrasts among parameters in Scheffe's mixture models: Screening components and model reduction. *Technometrics*, *20*(3), 273–279. https://doi.org/10.2307/1268136

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Perotte, A. J., Wood, F., Elhadad, N., & Bartlett, N. (2011). Hierarchically supervised latent Dirichlet allocation. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24, pp. 2608–2617). Curran Associates. https://proceedings.neurips.cc/paper/2011/file/489d0396e6826eb0c1e611d82ca8b215-Paper.pdf

Piepel, G. F. (1982). Measuring component effects in constrained mixture experiments. *Technometrics*, *24*(1), 29–39. https://doi.org/10.2307/1267575

Popping, R. (2015). Analyzing open-ended questions by means of text analysis procedures. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *128*(1), 23–39. https://doi.org/10.1177/0759106315597389

R Core Team. (2019). *R: A language and environment for statistical computing* (Version 3.6.2). R Core Development Team. https://www.R-project.org/

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). Wiley-Interscience.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*(4), 731–792. https://doi.org/10.1111/1467-9868.00095

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, *7*(1), 110–120. https://doi.org/10.1214/aoap/1034625254

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rohrer, J. M., Brümmer, M., Schmukle, S. C., Goebel, J., & Wagner, G. G. (2017). What else are you worried about?" Integrating textual responses into quantitative social science research. *PLoS ONE*, *12*(7), e0182156. https://doi.org/10.1371/journal.pone.0182156

Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 344–360. https://doi.org/10.1111/j.2517-6161.1958.tb00299.x

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. https://doi.org/10.1007/BF02296196

Snee, R. D., & Marquardt, D. W. (1976). Screening concepts and designs for experiments with mixtures. *Technometrics*, *18*(1), 19–29. https://doi.org/10.2307/1267912

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(4), 795–809. https://doi.org/10.1111/1467-9868.00265

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*(1), 25216. https://doi.org/10.3402/ejpt.v6.25216

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing* *systems* (Vol. 22, pp. 1973–1981). Curran Associates. https://proceedings.neurips.cc/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*(3), 263–311. https://doi.org/10.1207/s15327906mbr2803_1

Wilcox, K. T. (2021). *Psychtm: Text mining methods for psychological research* (Version 2021.1.0). https://cran.r-project.org/package=psychtm

Zhu, J., Zheng, X., & Zhang, B. (2013). Improved Bayesian logistic supervised topic models with data augmentation. *arXiv*. https://arxiv.org/pdf/1310.2408.pdf

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development*, *14*(4), 305–320. https://doi.org/10.1080/15427609.2017.1370966

Appendix A

Metropolis-in-Gibbs Algorithm for SLDAX With a Dichotomous Outcome

When y is dichotomous, we use the canonical logit link function to specify a linear relationship between the predictor and the log-odds for the SLDAX model,

$$log\left(\frac{\pi_d}{1-\pi_d}\right) = \vec{r}_d{}'\vec{\eta}, \tag{A1}$$

where $\pi_d = Pr[Y_d = 1 \mid \cdot]$ and $\vec{r}_d = (\vec{x}_d, \vec{\bar{z}}_d)$ is a $(p + K) \times 1$ vector of the p predictor values for observation d and the K empirical topic frequencies for observation/document d.

Assuming that the outcomes y_d, documents, and the words are conditionally independent, the likelihood function is

$$L(\vec{\Theta}, \vec{B}, \vec{\eta}) = \prod_{d=1}^{D} exp\{y_d \vec{r}_d{}'\vec{\eta}\} \times (1 + exp\{\vec{r}_d{}'\vec{\eta}\})^{-1} \prod_{n=1}^{N_d} \theta_{dz_{dn}} \beta_{z_{dn}w_{dn}}. \tag{A2}$$

Combining the priors and the likelihood, the posterior distribution is

$$f(\vec{\eta}, \vec{\Theta}, \vec{B}, \vec{z}_1, \ldots, \vec{z}_D \mid \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D) = \frac{L(\vec{\Theta}, \vec{B}, \vec{\eta}) f(\vec{\eta}) \prod_{d=1}^{D} f(\vec{\theta}_d) \prod_{k=1}^{K} f(\vec{\beta}_k)}{f(\vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D)}. \tag{A3}$$

Collapsed Metropolis-in-Gibbs Sampler Algorithm

As in the case of a normally distributed outcome, we marginalize over $\vec{\Theta}$ and $\vec{B}$ to obtain a collapsed sampling algorithm. With a dichotomous outcome, however, we can no longer directly sample the regression coefficients $\vec{\eta}$ from a known full conditional distribution. Instead, we use a Metropolis step to sample the regression coefficients (Gelman et al., 2014; Metropolis et al., 1953). We use independent normal distributions as proposal distributions for the regression coefficients,

$$\eta_j \sim \mathrm{N}(\mu_j, \tau_j), \tag{A4}$$

where the proposal variance τ_j for each coefficient is tuned during the burn-in period of sampling to yield desirable acceptance ratios (G. O. Roberts et al., 1997). Alternative proposal distributions can be used (e.g., t distributions) if desired.

To sample each coefficient $\eta_j, j = 1, \ldots, p + K$, a candidate draw $\eta_j^{(c)}$ at iteration t is drawn from $\mathrm{N}\left(\eta_j^{(t-1)}, \tau_j\right)$ where $\eta_j^{(t-1)}$ is the previous draw for that coefficient. Second, we compute the acceptance ratio $R = \frac{f(\vec{y} \mid \vec{Z}, \vec{X}, \vec{\eta}^{(c)}) f(\vec{\eta}^{(c)})}{f(\vec{y} \mid \vec{Z}, \vec{X}, \vec{\eta}^{(t-1)}) f(\vec{\eta}^{(t-1)})}$. Third, we sample $u \sim \mathrm{U}(0, 1)$. Finally, if $R > u$, we accept $\eta_j^{(c)}$ as a draw from the desired full conditional distribution and set $\eta_j^{(t)} = \eta_j^{(c)}$. If $R \leq u$, we set $\eta_j^{(t)} = \eta_j^{(t-1)}$. While the tuning parameters τ_j can be initialized to any positive number, we initialize $\tau_j = \tau = 2.38$ for all proposal variances as this has been shown to yield optimal acceptance ratios near .25 (G. O. Roberts et al., 1997).

The computational steps of the Metropolis-in-Gibbs algorithm are as follows. For iteration t, $t = 1, \ldots, T$:

1. Draw $\vec{\eta}^{(t)}$ from $f(\vec{\eta} \mid \sigma^{2^{(t-1)}}, \vec{z}_1^{(t-1)}, \ldots, \vec{z}_D^{(t-1)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D)$ using the Metropolis algorithm.
2. For n, $n = 1, \ldots, N_d$ and d, $d = 1, \ldots, D$:
    (a) Draw $z_{dn}^{(t)}$ from $f(z_{dn} \mid \vec{\eta}^{(t)}, \vec{z}_1^{(t-1)}, \ldots, z_{d(-n)}^{(t-1)}, \ldots, \vec{z}_D^{(t-1)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D)$.
3. For document d, $d = 1, \ldots, D$:
    (a) Draw topic proportions $\vec{\theta}_d^{(t)}$ from $f(\vec{\theta}_d \mid \vec{\eta}^{(t)}, \vec{z}_1^{(t)}, \ldots, \vec{z}_D^{(t)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D)$.
4. For topic k, $k = 1, \ldots, K$:
    (a) Draw topic–vocabulary distributions $\vec{\beta}_k^{(t)}$ from $f(\vec{\beta}_k \mid \vec{\Theta}^{(t)}, \vec{\eta}^{(t)}, \vec{z}_1^{(t)}, \ldots, \vec{z}_D^{(t)}, \vec{y}, \vec{X}, \vec{w}_1, \ldots, \vec{w}_D)$.

If $\vec{\Theta}$ and/or $\vec{B}$ are not of interest, Steps (3) and/or (4) can be omitted. The derivations of the necessary conditional distributions are provided in Appendix C.

(*Appendices continue*)

Appendix B

Derivation of the Gibbs Sampler for SLDAX With a Normally Distributed Outcome

If the outcome y_d, $d = 1, \ldots, D$ is assumed to follow a normal distribution, then the full-data joint posterior distribution for the SLDAX model has the form of

$$\begin{aligned} f(\vec{\eta}, \sigma^2, \vec{\Theta}, \vec{B}, \vec{Z} \mid \vec{W}, \vec{X}, \vec{y}) \propto\ & f(\vec{y} \mid \vec{Z}, \vec{X}, \vec{\eta}, \sigma^2) \\ & \times \prod_{d=1}^{D} f(\vec{w}_d \mid \vec{z}_d, \vec{B}) f(\vec{z}_d \mid \vec{\theta}_d) \times \prod_{k=1}^{K} f(\vec{\beta}_k) \\ & \times \prod_{d=1}^{D} f(\vec{\theta}_d) \times f(\vec{\eta} \mid \sigma^2) \times f(\sigma^2), \end{aligned} \tag{B1}$$

which can be written explicitly as

$$\begin{aligned} & f(\vec{\eta}, \sigma^2, \vec{\Theta}, \vec{B}, \vec{Z} \mid \vec{W}, \vec{X}, \vec{y}) \\ & \propto\ \sigma^{-D} exp\left\{ -\frac{1}{2\sigma^2} (\vec{y} - \vec{R}\vec{\eta})'(\vec{y} - \vec{R}\vec{\eta}) \right\} \\ & \times \prod_{d=1}^{D} \prod_{n=1}^{N_d} \beta_{z_{dn} w_{dn}} \theta_{d z_{dn}} \times \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\gamma - 1} \times \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha - 1} \\ & \times exp\left\{ -\frac{1}{2} (\vec{\eta} - \vec{\mu}_0)' \vec{\Sigma}_0^{-1} (\vec{\eta} - \vec{\mu}_0) \right\} \\ & \times (\sigma^2)^{-\dfrac{a_0}{2} - 1} exp\left\{ -\frac{b_0}{2\sigma^2} \right\}, \end{aligned} \tag{B2}$$

where $\vec{R} = (\vec{X}, \vec{\bar{Z}})$ is a $D \times (p + K)$ augmented matrix, $\beta_{z_{dn} w_{dn}}$ is the probability of observing word n in document d given the corresponding topic, $\theta_{d z_{dn}}$ is the probability of drawing topic z_{dn} for word n in document d, β_{kv} is the probability of observing the v-th element of the vocabulary, $v = 1, \ldots, V$ from topic k, $k = 1, \ldots, K$, and θ_{dk} is the probability of observing the k-th topic in document d.

We obtain samples from the posterior by repeatedly sampling in sequence from the full conditional distributions obtained from the full-data joint posterior:

(1) Draw the regression coefficients $\vec{\eta}$ from $f(\vec{\eta} \mid \cdot)$:

$$\begin{aligned} f(\vec{\eta} \mid \cdot) & \propto f(\vec{y} \mid \vec{\eta}, \sigma^2, \vec{Z}, \vec{X}) \times f(\vec{\eta} \mid \sigma^2) \\ & \propto\ exp\left\{ -\frac{1}{2\sigma^2} (\vec{y} - \vec{R}\vec{\eta})'(\vec{y} - \vec{R}\vec{\eta}) \right\} \\ & \times exp\left\{ -\frac{1}{2} (\vec{\eta} - \vec{\mu}_0)' \vec{\Sigma}_0^{-1} (\vec{\eta} - \vec{\mu}_0) \right\} \\ & \propto exp\left\{ -\frac{1}{2} (\vec{\eta} - \vec{\eta}_1)' \vec{\Sigma}_1^{-1} (\vec{\eta} - \vec{\eta}_1) \right\} \end{aligned} \tag{B3}$$

where $\vec{\Sigma}_1 = \left(\vec{\Sigma}_0^{-1} + \vec{R}'\vec{R}(\sigma^2)^{-1} \right)^{-1}$ assuming $\vec{\Sigma}_0$ and $\vec{\Sigma}_1^{-1}$ are invertible. A generalized inverse can be used if $\vec{\Sigma}_1^{-1}$ is singular. Let $\vec{\eta}_1 = \vec{\Sigma}_1 \left(\vec{\Sigma}_0^{-1} \vec{\mu}_0 + \vec{R}'\vec{y}(\sigma^2)^{-1} \right)$. Therefore, the full conditional distribution of $\vec{\eta}$ is a multivariate normal distribution,

$$\vec{\eta} \mid \cdot\ \sim\ \mathrm{N}(\vec{\eta}_1, \vec{\Sigma}_1). \tag{B4}$$

(2) Draw the residual variance σ^2 from $f(\sigma^2 \mid \cdot)$:

$$\begin{aligned} f(\sigma^2 \mid \cdot) & \propto f(\vec{y} \mid \vec{Z}, \vec{X}, \vec{\eta}, \sigma^2) \times f(\sigma^2) \\ & \propto (\sigma^2)^{-\frac{D}{2}} exp\left\{ -\frac{1}{2\sigma^2} (\vec{y} - \vec{R}\vec{\eta})'(\vec{y} - \vec{R}\vec{\eta}) \right\} \\ & \times (\sigma^2)^{-\frac{a_0}{2} - 1} exp\left\{ -\frac{b_0}{2\sigma^2} \right\} \\ & \propto (\sigma^2)^{-\frac{a_0 + D}{2} - 1} \\ & exp\left\{ -\frac{1}{2\sigma^2} \left(b_0 + (\vec{y} - \vec{R}\vec{\eta})'(\vec{y} - \vec{R}\vec{\eta}) \right) \right\} \end{aligned} \tag{B5}$$

Therefore, the full conditional distribution of σ^2 is an inverse-gamma distribution,

$$\sigma^2 \mid \cdot\ \sim\ \mathrm{IG}\left(\frac{a_0 + D}{2}, \frac{1}{2} [b_0 + (\vec{y} - \vec{R}\vec{\eta})'(\vec{y} - \vec{R}\vec{\eta})] \right) \tag{B6}$$

(3) Draw topic assignment z_{dn} for $d = 1, \ldots, D; n = 1, \ldots, N_d$ from $f(z_{dn} \mid \cdot)$: Griffiths and Steyvers (2004) derived the full conditional distribution[20] $f(z_{dn} \mid \cdot)$ when Θ and B are marginalized out of the joint posterior distribution of an LDA model,

[20] We omit a constant term in the denominator for simplicity because it does not affect the sampling of z_{dn}.

(*Appendices continue*)

$$f\left(z_{dn} = k \mid \vec{Z}^{(-dn)}, \vec{W}\right) \propto \frac{\left(n_{w_{dn}k}^{(-dn)} + \gamma\right)\left(n_{dk}^{(-dn)} + \alpha\right)}{n_{k}^{(-dn)} + V\gamma}, \tag{B7}$$

where $\vec{Z}^{(-dn)}$ denotes the topic assignments for the corpus excluding the topic assignment for position n in document d, $n_{w_{dn}k}^{(-dn)}$ is the number of co-occurrences of word w_{dn} with topic k excluding position n in document d, $n_{dk}^{(-dn)}$ is the number of assignments of topic k in document d excluding the topic assignment for position n in document d, $n_{k}^{(-dn)}$ is the number of assignments of topic k in the corpus excluding position n in document d, and $n_{d}^{(-dn)}$ is the number of words in document d excluding position n. Using Equation B7, the full conditional distribution $f(z_{dn} \mid \cdot)$ is

$$\begin{aligned} f(z_{dn} = k \mid \cdot) &\propto \; f(y_d \mid \vec{\bar{z}}_d, \vec{x}_d, \vec{\eta}, \sigma^2) f(z_{dn} = k \mid \vec{Z}^{(-dn)}, \vec{W}) \\ &\propto \; exp\left\{-\frac{1}{2\sigma^2}(y_d - \vec{r}_d{}'\vec{\eta})^2\right\} \times \frac{\left(n_{w_{dn}k}^{(-dn)} + \gamma\right)\left(n_{dk}^{(-dn)} + \alpha\right)}{n_{k}^{(-dn)} + V\gamma}, \end{aligned} \tag{B8}$$

where $\vec{r}_d = (\vec{x}_d, \vec{\bar{z}}_d)$ is a $(p + K) \times 1$ vector of the p predictor values for observation d and the K empirical topic frequencies for observation/document d.

(4) Optionally, draw the topic proportions $\vec{\theta}_d$ for d = 1, . . . , D from $f(\vec{\theta}_d \mid \cdot)$:

$$\begin{aligned} f(\vec{\theta}_d \mid \cdot) &\propto \; f(\vec{z}_d \mid \vec{\theta}_d) f(\theta_d \mid \alpha) \\ &\propto \; \prod_{k=1}^{K}\left[\prod_{n=1}^{N_d}\theta_{dk} I(z_{dn} = k)\right]\theta_{dk}^{\alpha-1} = \prod_{k=1}^{K}\theta_{dk}^{n_{dk}+\alpha-1}, \end{aligned} \tag{B9}$$

where $I(\cdot)$ is in an indicator function equal to one if its predicate is true and zero otherwise and n_{dk} is the number of assignments of topic k in document d. Therefore, the full conditional distribution of $\vec{\theta}_d$ is a Dirichlet distribution,

$$\vec{\theta}_d \mid \cdot \;\sim\; \text{Dir}(n_{d1} + \alpha, \ldots, n_{dK} + \alpha). \tag{B10}$$

(5) Optionally, draw the topic-word probabilities $\vec{\beta}_k$ for k = 1, . . . , K from $f(\vec{\beta}_k \mid \cdot)$:

$$\begin{aligned} f(\vec{\beta}_k \mid \cdot) &\propto \; f(\vec{W} \mid B, \vec{Z}) f(\vec{\beta}_k \mid \gamma) \\ &\propto \; \prod_{d=1}^{D}\prod_{n=1}^{N_d}\beta_{z_{dn}w_{dn}}\prod_{v=1}^{V}\beta_{kv}^{\gamma-1} \\ &= \prod_{v=1}^{V}\beta_{kv}^{n_{kv}+\gamma-1}, \end{aligned} \tag{B11}$$

where n_{kv} is the number of co-occurrences of topic k and word v in the corpus. Therefore, the full conditional distribution of $\vec{\beta}_k$ is a Dirichlet distribution,

$$\vec{\beta}_k \mid \cdot \;\sim\; \text{Dir}(n_{k1} + \gamma, \ldots, n_{kV} + \gamma). \tag{B12}$$

Appendix C

Derivation of the Gibbs Sampler for SLDAX With a Bernoulli-Distributed Outcome

The Gibbs sampling algorithm in appendix can be modified to handle a dichotomous outcome. In this case, the outcome $y_d \in \{0, 1\}, d = 1, \ldots, D$ is assumed to follow a Bernoulli distribution. The full-data joint posterior distribution for the SLDAX model becomes

$$\begin{aligned} f(\vec{\eta}, \vec{\Theta}, \vec{B}, \vec{Z} \mid \vec{W}, \vec{X}, \vec{y}) &\propto f(\vec{y} \mid \vec{Z}, \vec{X}, \vec{\eta}) \\ &\times \prod_{d=1}^{D} f(\vec{w}_d \mid \vec{z}_d, \vec{B}) f(\vec{z}_d \mid \vec{\theta}_d) \times \prod_{k=1}^{K} f(\vec{\beta}_k) \\ &\times \prod_{d=1}^{D} f(\vec{\theta}_d) \times f(\vec{\eta}), \end{aligned} \tag{C1}$$

which can be written explicitly as

$$\begin{aligned} &f(\vec{\eta}, \vec{\Theta}, \vec{B}, \vec{Z} \mid \vec{W}, \vec{X}, \vec{y}) \\ &\quad \propto \prod_{d=1}^{D}\left(\frac{exp\{y_d \vec{r}_d{}'\vec{\eta}\}}{1 + exp\{\vec{r}_d{}'\vec{\eta}\}}\right)^{y_d}\left(\frac{1}{1 + exp\{\vec{r}_d{}'\vec{\eta}\}}\right)^{1-y_d} \\ &\quad \times \prod_{d=1}^{D}\prod_{n=1}^{N_d}\beta_{z_{dn}w_{dn}}\theta_{dz_{dn}} \times \prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\gamma-1} \times \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{dk}^{\alpha-1} \quad . \\ &\quad \times exp\left\{-\frac{1}{2}(\vec{\eta} - \vec{\mu}_0)'\vec{\Sigma}_0^{-1}(\vec{\eta} - \vec{\mu}_0)\right\} \end{aligned} \tag{C2}$$

We obtain samples from the posterior by repeatedly sampling in sequence from the full conditional distributions obtained from the full-data joint posterior:

(1) Draw the regression coefficients $\vec{\eta}$ from $f(\vec{\eta} \mid \cdot)$:

(*Appendices continue*)

$$\begin{aligned} f(\vec{\eta} \mid \cdot) &\propto f(\vec{y} \mid \vec{\eta}, \vec{Z}, \vec{X}) \times f(\vec{\eta}) \\ &\propto \prod_{d=1}^{D} \left(\frac{exp\{y_d \vec{r}_d{}' \vec{\eta}\}}{1 + exp\{\vec{r}_d{}' \vec{\eta}\}} \right)^{y_d} \left(\frac{1}{1 + exp\{\vec{r}_d{}' \vec{\eta}\}} \right)^{1-y_d} \\ &\times exp\left\{ -\frac{1}{2} (\vec{\eta} - \vec{\mu}_0)' \vec{\Sigma}_0^{-1} (\vec{\eta} - \vec{\mu}_0) \right\} \end{aligned} \tag{C3}$$

This proportional density function does not have a known distributional form, so we instead use the Metropolis algorithm (Gelman et al., 2014; Metropolis et al., 1953) to sample from this full conditional distribution. We use independent normal proposal distributions to draw candidates $\eta_j^{(c)}$ for the regression coefficients at iteration t,

$$\eta_j^{(c)} \sim \mathrm{N}\left(\eta_j^{(t-1)}, \tau_j\right), \tag{C4}$$

where $\eta_j^{(t-1)}$ is the jth regression coefficient drawn in the previous iteration of the sampler and the proposal variance τ_j for each coefficient is tuned during the burn-in period of sampling to yield desirable acceptance ratios. We have found that initializing $\tau_j = \tau = 2.38$ for all proposal variances (G. O. Roberts et al., 1997) and then tuning the proposal variances yields good acceptance ratios around .25 (e.g., Gelman et al., 2014). We use normal proposal distributions because they are a common choice in Bayesian modeling (Gelman et al., 2014; Lynch, 2007), but alternative proposal distributions can be used.

Sampling proceeds for each coefficient $\eta_j, j = 1, \ldots, p + K$ by first sampling a candidate draw $\eta_j^{(c)}$ at iteration t according to Equation C4. Second, we compute $R = \frac{f(\vec{y} \mid \vec{Z}, \vec{X}, \vec{\eta}^{(c)}) \times f(\vec{\eta}^{(c)})}{f(\vec{y} \mid \vec{Z}, \vec{X}, \vec{\eta}^{(t-1)}) \times f(\vec{\eta}^{(t-1)})}$. Third, we sample $u \sim \mathrm{U}(0, 1)$. Finally, if $R > u$, we accept $\eta_j^{(c)}$ as a draw from the desired full conditional distribution and set $\eta_j^{(t)} = \eta_j^{(c)}$. If $R \leq u$, we set $\eta_j^{(t)} = \eta_j^{(t-1)}$.

(2) Draw topic assignment z_{dn} for $d = 1, \ldots, D; n = 1, \ldots, N_d$ from $f(z_{dn} \mid \cdot)$: As discussed in Appendix B, a collapsed Gibbs sampling step for z_{dn} can be obtained by marginalizing over Θ and B (Griffiths & Steyvers, 2004). Similar to Equation B8, the full conditional distribution $f(z_{dn} \mid \cdot)$ is

$$\begin{aligned} f(z_{dn} = k \mid \cdot) &\propto f(y_d \mid \vec{\bar{z}}_d, \vec{x}_d, \vec{\eta}, \sigma^2) f(z_{dn} = k \mid \vec{Z}^{(-dn)}, \vec{W}) \\ &\propto \prod_{d=1}^{D} \left(\frac{exp\{y_d \vec{r}_d{}' \vec{\eta}\}}{1 + exp\{\vec{r}_d{}' \vec{\eta}\}} \right)^{y_d} \left(\frac{1}{1 + exp\{\vec{r}_d{}' \vec{\eta}\}} \right)^{1-y_d} \\ &\times \frac{\left(n_{w_{dn}k}^{(-dn)} + \gamma\right)\left(n_{dk}^{(-dn)} + \alpha\right)}{n_k^{(-dn)} + V\gamma}. \end{aligned} \tag{C5}$$

(3) Optionally, draw the topic proportions $\vec{\theta}_d$ for $d = 1, \ldots, D$ from $f(\vec{\theta}_d \mid \cdot)$ given in Equation B10.

(4) Optionally, draw the topic-word probabilities $\vec{\beta}_k$ for $k = 1, \ldots, K$ from $f(\vec{\beta}_k \mid \cdot)$ given in Equation B12.

Appendix D

Derivation of the Regression Coefficients for the Simulation Study

This appendix describes the calculation of the regression coefficients used in the simulation study using variance decomposition of the total variance of the outcome Y.

As described in the Simulation Study, the data generation model was an SLDAX model with one manifest predictor X and K topics with a normally distributed outcome. The corresponding regression model is

$$y_d = \eta_X x_d + \sum_{k=1}^{K} \eta_k \bar{z}_{dk} + \epsilon_d, \tag{D1}$$

where the residuals ϵ_d, $d = 1, \ldots, D$ are independent of x_d and $\vec{\bar{z}}_d$, and identically distributed as $\epsilon_d \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$. Without loss of generality, we set the population marginal variance of Y to 1. For simplicity, we generated X independently of the topic assignments Z according to $X \sim \mathrm{N}(0, 1)$. In order to set the regression coefficients using partial R^2 effect size measures, we decomposed the marginal variance of Y,

$$\begin{aligned} \mathbb{V}[Y] &= \mathbb{V}[\eta_X X + \sum_{k=1}^{K} \eta_k \bar{Z}_k + \epsilon] \\ &= \eta_X^2 \mathbb{V}[X] + \mathbb{V}\left[\sum_{k=1}^{K} \eta_k \bar{Z}_k\right] + \sigma^2. \end{aligned} \tag{D2}$$

Let the model-explained variance be given by

$$\mathbb{V}[f] = \mathbb{V}[Y] - \sigma^2. \tag{D3}$$

We can define the proportion of model-explained variance as

(*Appendices continue*)

$$R^2 = \frac{\mathbb{V}[f]}{\mathbb{V}[Y]}, \tag{D4}$$

which simplifies to $R^2 = \mathbb{V}[f]$ because $\mathbb{V}[Y] = 1$. Because X and $\vec{Z}$ are generated independently, we can decompose R^2 into two orthogonal components associated uniquely with X and $\vec{Z}$, respectively,

$$R^2 = R_X^2 + R_Z^2. \tag{D5}$$

Regression Coefficient for X

The decomposition in Equation D5 implies that for a desired partial correlation R_X^2, the regression coefficient η_X can be obtained by

$$|\eta_X| = \frac{R_X}{\sqrt{\mathbb{V}[X]}}, \tag{D6}$$

which simplifies to $|\eta_X| = R_X$ because $\mathbb{V}[X] = 1$. Without loss of generality, we let $\eta_X = R_X$. For the desired effect size R_X^2 of .15 in this simulation study, $\eta_X = \sqrt{0.15}$.

Regression Coefficients for Empirical Topic Proportions

The decomposition in Equation D5 implies that for a desired partial correlation R_Z^2, the regression coefficients η_k, $k = 1, \ldots, K$ for the empirical topic proportions can be obtained from

$$R_Z^2 = \mathbb{V}\left[\sum_{k=1}^{K} \eta_k \bar{Z}_k\right]. \tag{D7}$$

Without loss of generality, we let $\mathbb{E}[Y] = 0$. Because $\mathbb{E}[X] = 0$, then $\mathbb{E}\left[\sum_{k=1}^{K} \eta_k \bar{Z}_k\right] = 0$. It is useful to recognize that generative distribution of $\vec{Z}_d$ is Dirichlet-multinomial. Because we specified the hyperparameter α to be 1 when generating data, then the populating mean and variance of the kth empirical topic proportion $\bar{Z}_{dk}$ for a document of length N_d are $\mathbb{E}[\bar{Z}_{dk}] = 1/K$ and $\mathbb{V}[\bar{Z}_{dk}] = \frac{K-1}{K^2(K+1)}\left(1 + \frac{K}{N_d}\right)$, and the covariance between two empirical topic proportions ($k \neq k'$) is $\text{Cov}[\bar{Z}_{dk}, \bar{Z}_{dk'}] = -\frac{1}{K^2(K+1)}\left(1 + \frac{K}{N_d}\right)$. We use these results in the following subsections to derive effect-size based regression coefficients for the empirical topic proportions.

Condition 1: K = 2 Topics

Combining the previous assumption that $\mathbb{E}\left[\sum_{k=1}^{K} \eta_k \bar{Z}_k\right] = 0$ with the expected value of $\bar{Z}_{dk}$ given above,

$$\begin{aligned} \mathbb{E}\left[\sum_{k=1}^{K} \eta_k \bar{Z}_k\right] &= 0 \\ \sum_{k=1}^{K} \eta_k K^{-1} &= 0 \\ \eta_1 &= -\eta_2, \end{aligned} \tag{D8}$$

we find that the topic regression coefficients must be equal in magnitude and opposite in sign with $K = 2$ topics.

Using the variance and covariance results for the empirical topic proportions given above, Equation D7 can be simplified,

$$\begin{aligned} R_Z^2 &= \mathbb{V}[\sum_{k=1}^{K} \eta_k \bar{Z}_k] \\ &= \eta_1^2 \mathbb{V}[\bar{Z}_{d1}] + (-\eta_1) 2 \mathbb{V}[\bar{Z}_{d2}] + 2\eta_1(-\eta_1)\text{Cov}[\bar{Z}_{dk,\bar{Z}_{dk}}], \end{aligned} \tag{D9}$$

For simplicity, we assume that $N_d = N, d = 1, \ldots, D$.[21] Equation D9 can then be shown to be equal to

$$R_Z^2 = \frac{1}{3}\left(1 + \frac{2}{N}\right)\eta_1^2, \tag{D10}$$

which yields

$$|\eta_1| = \frac{\sqrt{3R_Z^2 N}}{\sqrt{N+2}}. \tag{D11}$$

For the desired effect size R_Z^2 of .35 in this simulation study, we set $\eta_1 = -\eta_2 = \frac{\sqrt{(0.35)3N}}{\sqrt{N+2}}$ where N corresponds to $\bar{N}_d$ using the notation in the Simulation study section.

Condition 2: K = 5 Topics

In addition to the assumption that $\mathbb{E}\left[\sum_{k=1}^{K} \eta_k \bar{Z}_k\right] = 0$, we further assume that for a positive constant η^*, the topic regression coefficients are given by

[21] We observed little difference in a pilot simulation study between results using this assumption and results for which equal document lengths were not assumed.

(*Appendices continue*)

$$\vec{\eta}_Z = (-2\eta^*, -\eta^*, 0, \eta^*, 2\eta^*)'. \tag{D12}$$

Using the variance and covariance results for the empirical topic proportions given above and again assuming for simplicity equal document lengths N, it can be shown that Equation D7 yields

$$\begin{aligned} R_Z^2 &= \mathbb{V}\left[\sum_{k=1}^{K} \eta_k \bar{Z}_k\right] \\ &= \frac{2\eta^{*^2}}{6}\left(1 + \frac{5}{N}\right). \end{aligned} \tag{D13}$$

Solving for η^* yields

$$|\eta^*| = \frac{\sqrt{3R_Z^2 N}}{\sqrt{N+5}}. \tag{D14}$$

For the desired effect size R_Z^2 of .35 in this simulation study, we set $\eta_* = \frac{\sqrt{(0.35)3N}}{\sqrt{N+5}}$ and calculate the topic regression coefficients according to Equation D12 where N corresponds to $\bar{N}_d$ using the notation in the Simulation study section.

Received January 26, 2021
Revision received August 21, 2022
Accepted September 14, 2022 ■