Psychological Test and Assessment Modeling, Volume 65, 2023 (1), 3-35

Model Selection for Latent Dirichlet Allocation In Assessment Data

Constanza Mardones-Segovia[1], Jordan M. Wheeler[1], Hye-Jeong Choi[2], Shiyu Wang[1], Allan S. Cohen[1]

[1] Department of Educational Psychology, The University of Georgia, Athens, GA, USA
[2] The Human Resources Research Organization, Louisville, KY, and USA

Abstract

Latent Dirichlet Allocation (LDA) is a probabilistic topic model that has been used as a tool to detect the latent thematic structure in a body of text. In the context of classroom testing, LDA has been used to detect the latent themes in examinees' responses to constructed-response (CR) items. There is a growing body of evidence that latent themes detected by LDA have been found to reflect the kinds of reasoning examinees use in their responses to CR items. The use of the information from a model such as LDA requires that the model fit the data. To this end, a number of different model selection indices have been used with LDA to determine the best model fit. There does not as yet appear to be clear evidence, however, as to which of these indices is most accurate in conditions common with measurement data. In this study, we evaluated the performance of several model selection indices, including similarity measures and perplexity using 5-fold cross-validation. Their performance for model selection was compared from two commonly used algorithms for estimation of the LDA model, Gibbs sampling and variational expectation-maximization. Data were simulated with different numbers of topics, documents, average lengths of answers, and numbers of unique words typical of practical measurement conditions. Results suggested that the average cosine similarity and perplexity using 5-fold cross-validation were most accurate for model selection over the conditions simulated in this study.

Keywords: latent Dirichlet allocation, model selection, perplexity, k-fold cross-validation, similarity measures, Gibbs sampling, variational expectation maximization

Author Note

Correspondence concerning this article should be addressed to Constanza Mardones-Segovia, 125P Aderhold Hall, Department of Educational Psychology, Mary Frances Early, The University of Georgia, Athens, GA, 30602. E-mail: cam04214@uga.edu.

Introduction

Constructed-response (CR) items provide a useful format for measuring examinees' inquiry skills on complex tasks (Attali, 2014). Typically, the text of answers to CR items is scored using a rubric and then the scores are analyzed. Algorithmic scoring has automated the scoring process, improving the speed at which test results can be returned to examinees (e.g., Lockwood, 2014). Responses to CR items also have been shown to provide information about examinees' thinking and reasoning in addition to that provided by rubric-based scores (e.g., Buxton et al., 2014). Recent research on algorithmic analysis of the text of responses to CR items, for example, has suggested that information from topic modeling of the text of examinees' responses can provide information that indicates the thinking and reasoning underlying both, correct and incorrect answers in item response data (e.g., Copur-Gencturk et al., 2022).

Topic models are a family of statistical algorithms, designed to detect clusters in a corpus of textual data, that are assumed to reflect the latent thematic structure in the corpus. Latent Dirichlet allocation (LDA; Blei et al., 2003) is a commonly used topic model in educational measurement due to its utility for providing additional information about students beyond the scores from rubrics. For example, LDA has been shown to provide insight into students' thinking and reasoning, when responding to CR items (Cardozo-Gaibisso et al., 2019; Wheeler, Raczynski, et al., 2022). These studies illustrate how LDA can be used to illuminate the writing strategies used by students. In addition, Shin et al. (2019) used LDA to identify students' misconceptions in order to create item distractors and Basu et al. (2013) described a use of LDA for short answering grading of CR items. Besides providing information about scores, the LDA model has also been found to help improve the estimates of ability over traditional psychometric modeling of the rubric-based scores (Wheeler, Wang, et al., 2022).

LDA is a mixed membership model designed to detect latent clusters in a corpus of text. These clusters are assumed to reflect the latent thematic structure in the corpus (Wesslen, 2018). LDA, described more completely below, is typically used as an exploratory tool in order to detect the latent thematic structure in a corpus of textual data. Due to the exploratory nature of LDA, an important consideration in applying LDA is to select the model with the number of topics that best fits the data. A number of indices have been reported to inform topic model selection, including similarity indices and model perplexity using cross-validation (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014). Although LDA has shown useful and promising results when analyzing students' textual responses to CR items, there appears to be little consistency across studies on model selection indices used. Additionally, there are limited studies that have investigated the performance of different model selection indices on corpora that contain data typically found in practical educational settings. That is, the textual responses to CR items are generally more constrained and contain fewer unique words due to the requirements indicated in the prompts, and the texts are often shorter in length. Kim et al. (2017), for example, used LDA to detect three latent topics in the answers of 243 middle-grade students on a test of science inquiry knowledge. The average answer length was 98.6 words with the number of unique words as only 532 words.

In this study, the performances of three model selection indices were investigated and compared: the average cosine similarity (CS; Deveaud et al., 2014), the average Jensen-Shannon divergence (JSD; Deveaud et al., 2014), and perplexity using 5-fold cross-validation (Refaeilzadeh et al., 2009). Specifically, these three model selection indices were studied using a simulation study that compared their accuracy in selecting the correct LDA model. The conditions of the simulation study were designed to reflect corpora that are often found in applications of LDA within educational contexts. Additionally, the accuracy of these model selection indices was compared using two estimator methods: Gibbs sampling and variational expectation maximization (VEM).

Latent Dirichlet Allocation

LDA is a probabilistic model designed to detect the latent topic structure present in a corpus of textual documents (Blei, 2012; Blei et al., 2003). LDA assumes that a corpus is a collection of D documents and each document $d \in \{1, \ldots, D\}$ is a collection of N_d observed words. The vocabulary is the set of V unique words that appear in the corpus. Document d is denoted by $\vec{w}_d$ = $\left[w_1, \ldots, w_{n_d}\right]$, where $w_{d,n}$ for $d = 1, \ldots, D$ and $n = 1, \ldots, N_d$, represents the *n*th observed word in the *d*th document.

Assuming K topics a priori, LDA estimates three parameters: topics, topic proportions, and topic assignments. Topics are assumed to follow a Dirichlet distribution over the V words in the vocabulary. Topic $k \in \{1, \ldots, K\}$ is denoted by $\vec{\phi}_k = [\phi_1, \ldots, \phi_V]$ where $\phi_{k,v}$, for $k = 1, \ldots, K$ and $v = 1, \ldots, V$, is the probability of the *v*th word in the vocabulary occurring in the *k*th topic. Topic proportions are assumed to follow a Dirichlet distribution over the K topics and indicate the mixture proportions in a document. The topic proportions for document $d \in \{1, \ldots, D\}$ is denoted by $\vec{\theta}_d$ = $[\theta_1, \ldots, \theta_k]$ where $\theta_{d,k}$, for $d = 1, \ldots, D$ and $k = 1, \ldots, K$, is the proportion of the *k*th topic appearing in the *d*th document. Topic assignments are discrete values that indicate the topic membership of each word in a document. The topic assignments for document d is denoted by $\vec{z}_d$ = $\left[z_1, \ldots, z_{n_d}\right]$ where $z_{d,n}$, for $d = 1, \ldots, D$ and $n = 1, \ldots, N_d$, is the topic membership of the *n*th word in the *d*th document.

As mentioned above, LDA assumes topics, $\vec{\phi}$ and topic proportions, $\vec{\theta}$, follow a Dirichlet distribution with hyperparameters $\vec{\beta}$ and $\vec{\alpha}$, respectively. Additionally, the topic assignments for document d are assumed to follow a multinomial distribution with parameters $\vec{\phi}_k$ and N_d. LDA estimates these parameters using the observed words (Ponweiser, 2012). The conditional distribution for the topics, topic proportions, and topic assignments, given the observed words is (Blei, 2012):

$$p(\vec{\theta}_{1:D}, \vec{\phi}_{1:K}, \vec{z}_{1:D} | \vec{w}_{1:D}) = \frac{p(\vec{\theta}_{1:D}, \vec{\phi}_{1:K}, \vec{z}_{1:D}, \vec{w}_{1:D})}{p(\vec{w}_{1:D})} \quad (1)$$

where the numerator, $p(\vec{\theta}_{1:D}, \vec{\phi}_{1:K}, \vec{z}_{1:D}, \vec{w}_{1:D})$, is the joint probability distribution of the latent and observed variables, and the denominator, $p(\vec{w}_{1:D})$, is the marginal probability distribution of the observed words. The joint probability distribution for the latent and observed variables is:

$$p(\vec{\theta}_{1:D}, \vec{\phi}_{1:K}, \vec{z}_{1:D}, \vec{w}_{1:D}) = \prod_{k=1}^{K} p(\vec{\phi}_k | \vec{\beta}) \prod_{d=1}^{D} p(\vec{\theta}_d | \vec{\alpha}) \left(\prod_{n=1}^{N_d} p(z_{d,n} | \vec{\theta}_d)\, p(w_{d,n} | \vec{\phi}_{k=z_{d,n}}) \right), \quad (2)$$

where $p(\vec{\phi}_k | \vec{\beta})$ is the conditional probability distribution for topic k with hyperparameter $\vec{\beta}$; $p(\vec{\theta}_d | \vec{\alpha})$ is the conditional probability distribution of the topic proportions for document *d* with hyperparameter $\vec{\alpha}$; $p(z_{d,n} | \vec{\theta}_d)$ is the conditional probability distribution of the topic assignments for the *n*th word in the *d*th document; and $p(w_{d,n} | \vec{\phi}_{k=z_{d,n}})$ is the conditional probability distribution of the *n*th word in the *n*th document given the topic and topic assignment (Ponweiser, 2012).

The LDA computes the marginal probability distribution of words by integrating over the topics, $\vec{\phi}_{1:K}$, and topic proportions, $\vec{\theta}_{1:D}$, and summing over the topic assignments, $\vec{z}_{1:D}$, such that:

$$p(\vec{w}_{1:D}) = \int_{\vec{\phi}} \int_{\vec{\theta}} p(\vec{\theta}_d | \vec{\alpha})\ p(\vec{\phi}_k | \vec{\beta}) \prod_{n=1}^{N_d} \sum_{\vec{z}_{1:D}} \prod_{n=1}^{N_d} p(z_{d,n} | \vec{\theta}_d)\, p(w_{d,n} | \vec{\phi}_{k=z_{d,n}})\, d\vec{\theta} d\vec{\phi}. \quad (3)$$

Estimating Parameters

Since summing over the combination of all the topic assignments is computationally intractable (Blei et al., 2003; Ponweiser, 2012), different methods have been proposed to estimate the LDA parameters, including the collapsed Gibbs sampling (Griffiths & Steyvers, 2004) estimation method and the variational expectation maximization (VEM; Blei et al., 2003) estimation method.

Gibbs sampling is a Markov chain Monte Carlo algorithm (MCMC; Blei et al., 2003; Griffiths & Steyvers, 2004) that approximates a complex posterior distribution. The Gibbs sampling algorithm randomly selects samples from a conditional distribution of a lower dimension until the sampled values reach a stationary distribution (Steyvers

& Griffiths, 2007). The sampled values obtained at the beginning of the Markov chain tend to be a poor approximation of the posterior. Thus, a burn-in period is used in which the estimates from this period are discarded, and the subsequent estimates, i.e., after discarding the samples from the beginning of the Markov chain, are used to obtain the posterior estimates. Although the Gibbs sampling does provide an accurate estimation method to obtain the posterior, it often requires a long process to reach convergence to the target distribution, therein requiring more time and computational resources to obtain the posterior (for more information, see Steyvers & Griffiths, 2007).

VEM, another algorithm sometimes used to estimate topic models, is a deterministic algorithm which uses optimization to approximate the posterior distribution. Instead of sampling the posterior, VEM approximates the lower bound of the log-likelihood. Thus, it is an estimation method that tends to be faster and less computationally intensive than MCMC methods (Blei et al., 2017). VEM consists of two steps: the E-step, which finds the optimizing values of the topic proportions and topic assignments, and the M-step, which maximizes the lower bound on the log-likelihood by minimizing the Kullback-Liebler Divergence (KLD) between the approximation of the true and the estimated distributions (Taddy, 2012). Although VEM is often faster than Gibbs sampling, it may underestimate the variance of the posterior (Blei et al., 2017).

There are various factors that influence the accuracy of the parameter estimates for the LDA model, including the estimation algorithm, the number of documents in the corpus, the number of unique words in the corpus (i.e., the vocabulary size), the length of each document (i.e., number of words in each document), and the selection of hyperparameters (Syed & Spruit, 2018; Wallach et al., 2009; Wheeler et al., 2021; Wheeler, Xiong, et al., 2022).

When the text length is short, LDA has been reported to lack word co-occurrences sufficient for detecting latent clusters in the corpus (e.g., Chen et al., 2016; Hu et al., 2009; Tang et al., 2014; Zuo et al., 2016). Recent evidence suggests, however, that LDA tends to improve its performance, when the distance among topics is large, that is, when topics are semantically different (Tang et al., 2014). In particular, the hyperparameters $\vec{\beta}$ and $\vec{\alpha}$ have been found to influence the estimation of the topics and topic proportions (Ponweiser, 2012). When their values are equal to 1, the Dirichlet distribution resembles a uniform distribution, meaning that all words are equally likely to occur in each topic and topics are equally likely to appear in each document. When their values are less than 1, the Dirichlet distribution tends to be asymmetric, meaning a smaller subset of words are likely to occur in each topic and a smaller subset of topics are likely to appear in a document (Syed & Spruit, 2018; Wallach et al., 2009).

Mardones-Segovia et al. (2022) compared the performance of Gibbs sampling and VEM estimation algorithms for LDA. The study used simulation conditions that reflected corpora sizes often found in educational testing data. Results suggested that the accuracy of the topic proportions was better when the set of hyperparameters was

$\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$, that is, when topics were semantically different. This effect of the hyperparameters appeared to decrease as the length of each document increased. When the documents had a length of 20 words, the Gibbs sampling tended to estimate the topic proportions more accurately than using the VEM algorithm. However, for documents with at least 50 words, differences in the accuracy of the topics proportion were lower between both estimation methods.

Model Selection Indices

As noted earlier, the LDA model is an exploratory model where the number of topics is specified a priori (Blei et al., 2003). Therefore, the use of LDA often requires running a set of candidate models with differing the number of topics. Each candidate model is compared using a variety of model selection indices and a final model is chosen based on the performance of the model selection indices and the interpretability of the final model (Cohen & Cho, 2017; Myung & Pitt, 2004).

The objective of model selection indices is to help evaluate which candidate model produces the smallest discrepancy between the probability distributions of the true model and the estimated candidate model. There are a variety of model selection indices used throughout the LDA literature, including information criterion indices (e.g., Lau et al., 2013; Schröder et al., 2017; Wang et al., 2016), topic coherence measures (e.g., Mimno et al., 2011; Newman et al., 2010; Nikolenko, 2016; Röder et al., 2015), similarity measures (e.g., Anderson et al., 2020; Roque et al., 2019), and perplexity (e.g., Blei et al., 2003; Vu et al., 2019). Each of these types of indices is described below. The information criterion indices and topic coherence measures are briefly described, and the similarity measures and perplexity measures are described in detail.

Model Selection Based on Information Criterion Indices

Information criterion (IC) indices are indicators of relative fit among candidate models (Kang & Cohen, 2007; Li et al., 2009; Sen & Cohen, 2019). That is, the fit of the models is compared only to those models in the set of candidate models. In this context, the smallest value for an IC index represents the best fit among the set of candidate models. The IC indices involve two terms: a goodness of fit term that represents how well a model fits the data and a penalty term that indicates the complexity of the model. Commonly used information criterion indices for LDA include the Akaike Information Criterion (AIC; Akaike, 1998), the Bayesian Information Criterion (BIC; Raftery, 1995), the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), and the sample size adjusted BIC (SABIC; Sclove, 1987).

Model Selection Based on Topic Coherence Measures

Topic coherence measures were developed specifically for topic models in order to address model selection issues. These measures select the candidate model in which individual topics are more semantically coherent and exclusive (Boyd-Graber et al., 2014; Chang et al., 2009). There are many topic coherence measures used for topic models (see Nikolenko, 2016; Röder et al., 2015). A few commonly used topic coherence measures include the semantic coherence measure (SC; Mimno et al., 2011), the pairwise pointwise mutual information metric (PMI; Newman et al., 2010), and the frequency and exclusivity metric (FREX; Bischof & Airoldi, 2012).

Model Selection Based on Similarity Measures

The use of similarity measures assumes that the best-fitting topic model is the one that produces high within-topics similarity and low between-topics similarity. Commonly used similarity-based model selection methods include the methods that utilize the cosine similarity measure (CS; Cao et al., 2009) and the Jensen-Shannon Divergence measure (JSD; Deveaud et al., 2014).

Cao et al. (2009) proposed selecting the best topic model by computing the average cosine similarity between every possible combination of topics. This method first requires calculating the cosine similarity between each possible combination of topics. The cosine similarity between two topics, Topic *i* and Topic $_j$, where $i \neq j$ and $i, j \in \{1, \dots, K\}$, is given by:

$$CS(\vec{\beta}_i, \vec{\beta}_j) = \frac{\sum_{v=1}^{V} \beta_{i,v}\beta_{j,v}}{\sqrt{\sum_{v=1}^{V}(\beta_{i,v})^2}\sqrt{\sum_{v=1}^{V}(\beta_{j,v})^2}} \tag{4}$$

where $\beta_{i,v}$and $\beta_{j,v}$are the probabilities of the *v*th word in the vocabulary occurring in the *i*th and *j*th topics, respectively. Cosine similarity is a type of correlation measure, therefore, a cosine similarity closer to 1 indicates that two topics are highly similar and a cosine similarity closer to 0 indicates that two topics are independent. Once the cosine similarity between each topic within a candidate model is calculated, the model selection method computes the average cosine similarity, such that:

$$\overline{CS} = \frac{\sum_{i=1}^{K}\sum_{j=i+1}^{K} CS(\vec{\beta}_i, \vec{\beta}_j)}{K\left(\frac{K-1}{2}\right)} \tag{5}$$

where $K\left(\frac{K-1}{2}\right)$ is the number of combinations of topics. A smaller average cosine similarity value indicates that the topics are more dissimilar and independent whereas a larger average cosine similarity value indicates that the topics are more similar and correlated. Cao et al. (2009) suggests that the best-fitting topic model is the one that minimizes the average cosine similarity, suggesting that the topics are more stable. Deveaud et al. (2014) proposed a model selection method that derives the best-fitting model through topic distributions. This model selection method uses the *JSD* measure to estimate the optimal number of topics. The *JSD* measure is a symmetric version of *KLD* that evaluates the similarity of pairs of words within topics, as shown by:

$$JSD = \frac{1}{2} \sum_{v \in V_k \cap V_k'} p(\phi_{k,v}) \times \log\left(\frac{p(\phi_{k,v})}{p(\phi_{k',v})}\right) + \frac{1}{2} \sum_{v \in V_k \cap V_k'} p(\phi_{k',v}) \times \log\left(\frac{p(\phi_{k',v})}{p(\phi_{k,v})}\right). \tag{6}$$

When its value approaches zero, the more similar are the words. Conversely, the closer its value to one, the more dissimilar are the words. Deveaud et al. (2014) expanded this for calculating the semantic similarity of words between topics. This metric requires computing the average JSD measure between all possible pairs of topics. Therefore, assume that T_k is the set of K candidate topics. Then, the best candidate model can be obtained as follows:

$$\overline{JSD} = \frac{1}{K(K-1)} \sum_{(k,k') \in T_k} JSD. \tag{7}$$

In this context, the model with the more distinguishable terms between topics, and the more coherent words within topics, is the most accurate topic structure. In this way, the model that maximizes Equation 7 represents the best candidate model.

Model Selection based on Perplexity Measures

Perplexity is another model selection technique that has vastly been applied in the framework of machine learning (Ding et al., 2018). It is a statistical method that evaluates how well a model can predict a new corpus of documents. In this regard, the candidate model with the lower perplexity value represents the best topic structure (Blei et al., 2003). Formally, perplexity is defined as the probability of the observed words in a document, $\vec{w}_d$, normalized by the total number of words in the document, N_d, such that:

$$perplexity = exp\left\{- \frac{\sum_{d=1}^{D} \log p(\vec{w}_d)}{\sum_{d=1}^{D} N_d}\right\} \quad (8)$$

In general, perplexity is evaluated using a cross-validation technique (CV; Neishabouri & Desmarais, 2020). CV is a re-sampling method that evaluates the generalizability of a model on a training set and a test set (Berrar, 2019). There are different strategies to split the corpus of documents. A commonly method in machine learning is the k-fold CV. This method randomly splits the original corpus into k folds or subsets. Then, CV fits a set of candidates LDA models using a $k-1$ folds as a training set and uses the remaining subset to evaluate the performance of those LDA models in an unseen corpus (Refaeilzadeh et al., 2009).

Method Design of Simulation Study

The simulation study investigates the performance of model selection indices for LDA under practical testing conditions. Practical testing conditions are defined by properties of corpora often found in applications of LDA to educational data. The corpora factors manipulated in the simulation include the number of documents in the corpus (four levels: $D = 200, D = 300, D = 500, D = 1000$), the number of unique words in the corpus (i.e., vocabulary size; two levels: $V = 350$ and $V = 650$), the average number of words in each document (i.e., average document length; five levels: $\lambda = \overline{N}_{1:D} = 5, \lambda = \overline{N}_{1:D} = 20, \lambda = \overline{N}_{1:D} = 50, \lambda = \overline{N}_{1:D} = 100,$ and $\lambda = \overline{N}_{1:D} = 200$), and the number of topics (three levels: $K = 3, K = 4,$ and $K = 5$).

The levels for the number of documents in the corpus were selected to reflect analysis of responses at the individual school and school district levels, that is, $D = 200, 300$ represents the number of students at the school level whereas $D = 500, 1000$ represents the number of students at the school district level. The levels for the average number of words in each document were selected to reflect the different types of CR items that can be analyzed with LDA. That is, $\lambda = \overline{N}_{1:D} = 5, 20$ represents short responses (e.g., one-line responses on an assessment), and $\lambda = \overline{N}_{1:D} = 50, 100, 200$ represents long responses (e.g., one-paragraph or one-page responses on an assessment). The number of unique words in the corpus and the number of topics were selected based on previous applications of LDA to educational data (e.g., Buxton et al., 2014; Cardozo-Gaibisso et al., 2019; Wheeler, Engelhard, et al., 2022).

In addition to the different corpora factors, the study also investigated two different estimation algorithms (Gibbs sampling and VEM) and three hyperparameters for $\vec{\alpha}$ (three levels: $\vec{\alpha} = 0.5,\ \vec{\alpha} = 1,\ \vec{\alpha} = \frac{1}{K}$, where K is the number of topics). Mardones-Segovia et al. (2022) found that LDA accurately estimated the topics regardless of the $\vec{\beta}$ values or the estimation algorithm used. Therefore, $\vec{\beta}$ values were not manipulated. All factors were crossed with 50 replications of each condition.

Data Generation

The data used in the simulation study were generated according to the generative process described in Blei et al. (2003). To generate the topics, this study utilized the mean of a topic word distribution from a real data set with conditions like those in this simulation study ($\vec{\beta} = 0.004$). The data were generated as follows:

1. Generate each topic, $\vec{\phi}_k \sim Dirichlet(\vec{\beta})\ \forall\ k = 1, \dots, K$
2. Generate topic proportions for each document, $\vec{\theta}_d \sim Dirichlet(\vec{\alpha})\ \forall\ d = 1, \dots, D$
3. Generate the observed words for each document $d = 1, \dots, D$:
    i. Choose the number of words in the document, $N_d \sim Poisson(\lambda)$
    ii. For each word in the document $n = 1, \dots, N_d$:
        - Select the topic member of the word, $z_n \sim Multinomial(\ \vec{\theta}_d)$
        - Select the observed word, $w_n \sim Multinomial(\vec{\beta}_{k=z_n})$

The generated data were then used to estimate a variety of candidate models with different numbers of topics. See the Appendix for visualizing the R functions used to simulate the data.

Parameter Estimation

For each condition in the simulation study, nine candidate LDA models were estimated, $T_K = \{2,3, \dots, 10\}$ using the Gibbs sampling and VEM algorithms as implemented in the R package *topicmodels* (Grün & Hornik, 2011). In the case of the Gibbs sampling, LDA estimated the parameters using three sets of hyperparameters: $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$, $\vec{\alpha} = 1$ & $\vec{\beta} = 1$, and $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$, where K is the number of topics. Samples at the beginning of a Markov chain were discarded as burn-in, and parameters were estimated from the posterior distribution using the post-burn-in iterations. For this purpose, the first 10,000 iterations were discarded as burn-in, and the next 5,000 iterations were used to estimate the LDA parameters from the posterior distribution. Two approaches were used to estimate the VEM algorithm: (1) setting the $\vec{\alpha}$ hyperparameter to the same values used for generating the data, and (2) setting the $\vec{\alpha}$ hyperparameter to $\vec{\alpha} = \frac{50}{K}$ as suggested by Griffiths and Steyvers (2004). Both VEM approaches freely estimated the hyperparameter $\vec{\beta}$. See the Appendix for an R code example for estimating LDA models using both estimation methods.

Label Switching

Label switching is an important concept to address when estimating a mixture model. For LDA, label switching occurs when a topic changes its membership association label within iterations in a single chain or between chains during the estimation procedure (Cho et al., 2013; Stephens, 2000). Addressing label switching is important for evaluating parameter recovery in simulation studies because the generating topic labels need not be the same as the estimated topic labels. For example, *Topic 1* in the generating model may not be *Topic 1* in the estimated model because the topic labels from the estimated model depend on the initialization of the estimation algorithm. Therefore, label switching identifies which topics from the generating model belong to which topics in the estimated model. In this study, label switching was detected by computing the cosine similarity between the generated and estimated LDA parameters using the *lsa* R package (Wild, 2020). A cosine similarity value close to one suggests that the generating topic is associated with the estimated topic, whereas a value close to zero suggests that the generating topic is not associated with the estimated topic. In this study, we were able to successfully identify which generating topics were associated with which estimated topics, thus allowing us to evaluate parameter recovery.

Model Selection Indices

This study investigated the performance of the cosine similarity measure presented in Equation (5), the JSD measure presented in Equation (6), and the perplexity measure presented in Equation (8). The similarity measures were obtained using the R package *ldatuning* (Nikita, 2019) and the perplexity measure was obtained using the R package *topicmodels* (Grün & Hornik, 2011).

In the context of machine learning, researchers tend to calculate perplexity using k-fold CV (Xiong et al., 2020). This study evaluated the performance of perplexity using 5-fold CV (perplexity 5-CV; Hasan et al., 2021; Refaeilzadeh et al., 2009). Perplexity 5-CV fits the set of candidate models five times. Each time, it chooses $k - 1$ subsets of the document-term matrix and trains the LDA models. Next, perplexity 5-CV tests the previous LDA models using the remaining subset. As the topics were given by these previously fitted models, LDA only estimates the topic distribution. Finally, this study computed perplexity for each k-fold and candidate model. For interpretation purposes, this study computed the average of the five perplexity scores. The model with the lowest average perplexity was taken as the best candidate LDA model.

The accuracy for model selection of the similarity measures and perplexity were determined by counting the number of times that each indicator selected the correct topic model across replications. A frequency of zero indicates that the model selection index did not select the simulated topic model in any of the 50 replications, while a frequency of 50 indicates that the model selection index selected the generated topic model in each replication.

Recovery of Model Parameters

The performance of an estimated model can be evaluated by comparing the generating parameters to the estimated parameters. One such metric is the root mean squared error (RMSE) that measures the error between the estimated parameters and the generating parameters. A smaller value indicates that the estimated parameters are close to the generating parameters, and the model accurately estimated the parameters. A larger value indicates that the estimated parameters are not close to the generating parameters, and the model did not accurately estimate the parameters (Hübner & Pelzer, 2020).

In this study the RMSE was calculated between the estimated topics and the generating topics, and between the estimated topic proportions and the generating topic proportions. The RMSE for topic K is calculated by:

$$RMSE\left(\vec{\phi}_k\right) = \sqrt{\frac{\sum_{v=1}^{V}\left(\hat{\phi}_{k,v} - \phi_{k,v}\right)^2}{V}}, \tag{9}$$

where $\hat{\phi}_{k,v}$ is the estimated probability of the *v*th word in the vocabulary to occur in the *k*th topic and $\phi_{k,v}$ is the true probability of the *v*th word in the vocabular to occur in the *k*th topic. The RMSE for the topic proportions for document d is calculated by:

$$RMSE\left(\vec{\theta}_d\right) = \sqrt{\frac{\sum_{k=1}^{K}\left(\hat{\theta}_{d,k} - \theta_{d,k}\right)^2}{K}}, \tag{10}$$

where $\hat{\theta}_{d,k}$ is the estimated proportion of the *k*th topic appearing in the *d*th document and $\theta_{d,k}$ is the true proportion of the *k*th topic appearing in the *d*th document.

The average RMSE values for each topic and each topic proportion were calculated for each condition within a replication, indicating the average RMSE value calculated for the recovery of the topic parameters and the average RMSE value calculated for the recovery of the topic proportions. These two values indicate the overall performance of the estimated model for each condition. Additionally, the average RMSE are reported across all 50 replications for each condition.

Results

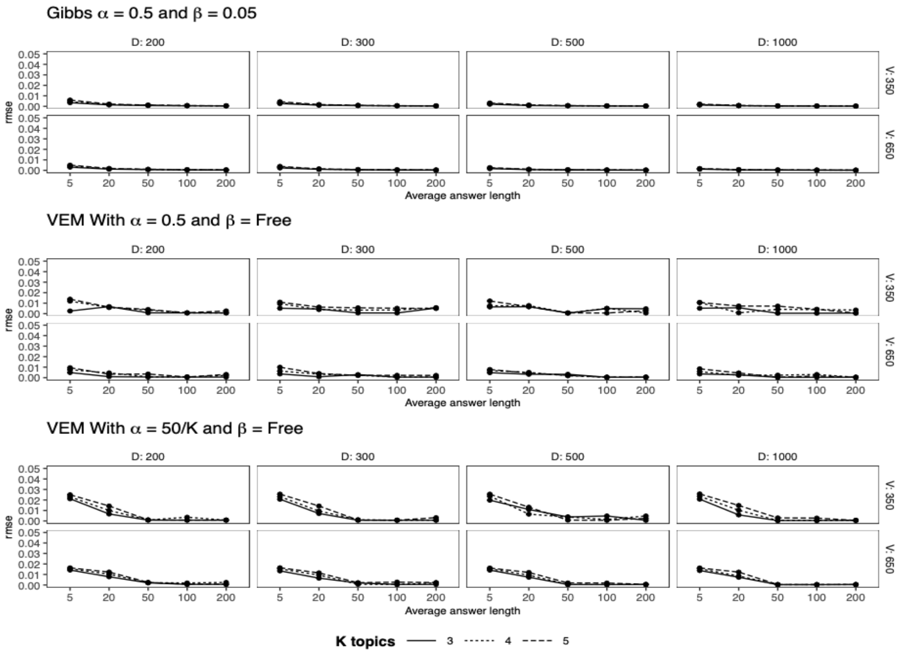
Recovery of Parameters

Recovery results for the topic parameters are plotted in Figure 1 and the recovery for the topic proportions parameters are plotted in Figures 2, 3, 4. Generally, the recovery of the topics and topic proportions appeared to vary mainly by the average answer length, the hyperparameters, and the type of estimation algorithm. Regardless of the vocabulary of unique words and the document size, RMSE values tended to be higher for corpora with average document lengths of 5 and 20 words (i.e., $\lambda = 5$ and $\lambda = 20$). However, as the average document lengths increased, the variability between the estimated and generated parameters tended to become smaller. Further, the influence of average answer length on the recovery of the topic proportion appeared to decrease when LDA estimated the parameters using $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$, regardless of the estimation algorithm used.

In addition, the variability between the estimated and generated parameters tended to be smaller for parameters estimated using the Gibbs sampling than using VEM. In particular, RMSE values were higher, when $\vec{\alpha}$ was $\frac{50}{K}$ and corpora had a very small average answer length. However, differences in RMSE values between both algorithms tended to be minimal for corpora with average document lengths of 50, 100, and 200 words.

Figure 1

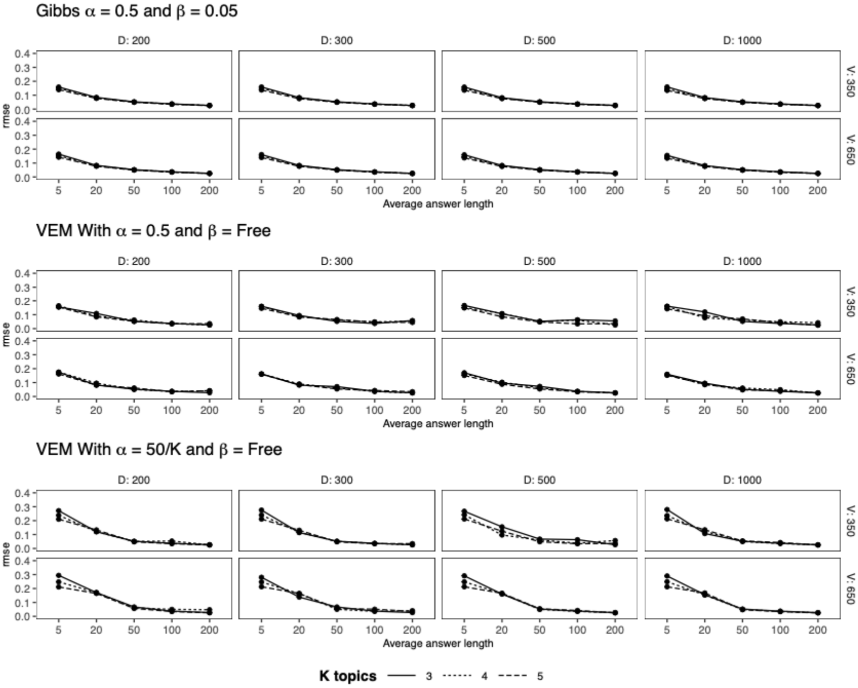
RMSE results for topics for data simulated using $\vec{\alpha} = 0.5$



Note: RMSE = Root Mean Squared Error, $D =$ number of doc number of documents in corpus, $V =$ vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, Gibbs = Gibbs sampling algorithm, VEM = variational-expectation maximization algorithm, $\vec{\alpha} = 0.5$ hyperparameter for topic proportions used to generate the data, $\vec{\alpha} = \frac{50}{K}$ hyperparameter for topic proportions used to estimate the model, and $\vec{\beta} =$ hyperparameter for topics used to estimate the model.

Figure 2

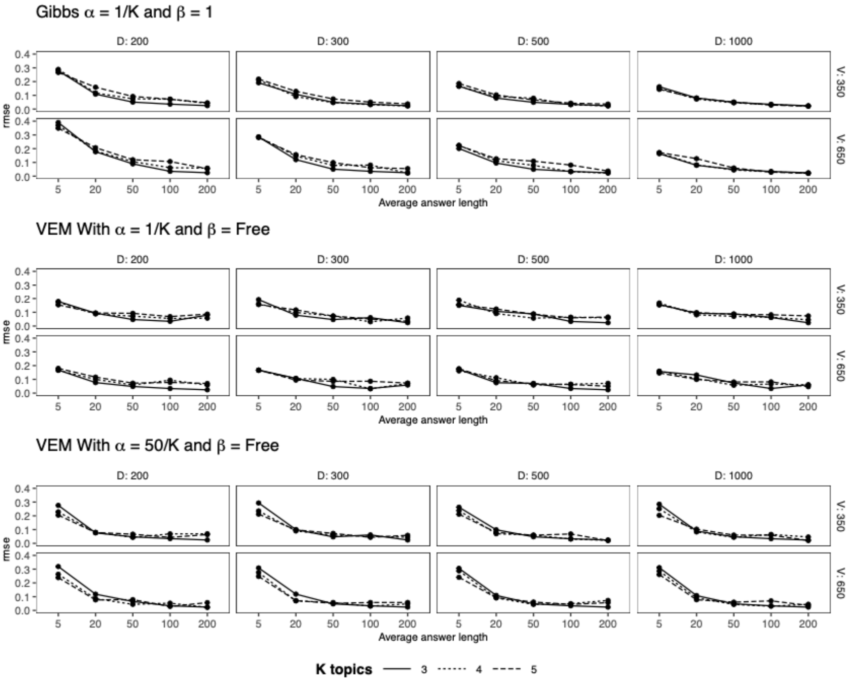
RMSE results for topic proportion for data simulated using $\vec{\alpha} = 0.5$



Note: RMSE = Root Mean Squared Error, $D =$ number of doc number of documents in corpus, $V =$ vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, Gibbs = Gibbs sampling algorithm, VEM = variational-expectation maximization algorithm, $\vec{\alpha} = 0.5$ hyperparameter for topic proportions used to generate the data, $\vec{\alpha} = \frac{50}{K}$ hyperparameter for topic proportions used to estimate the model, and $\vec{\beta} =$ hyperparameter for topics used to estimate the model.

Figure 3

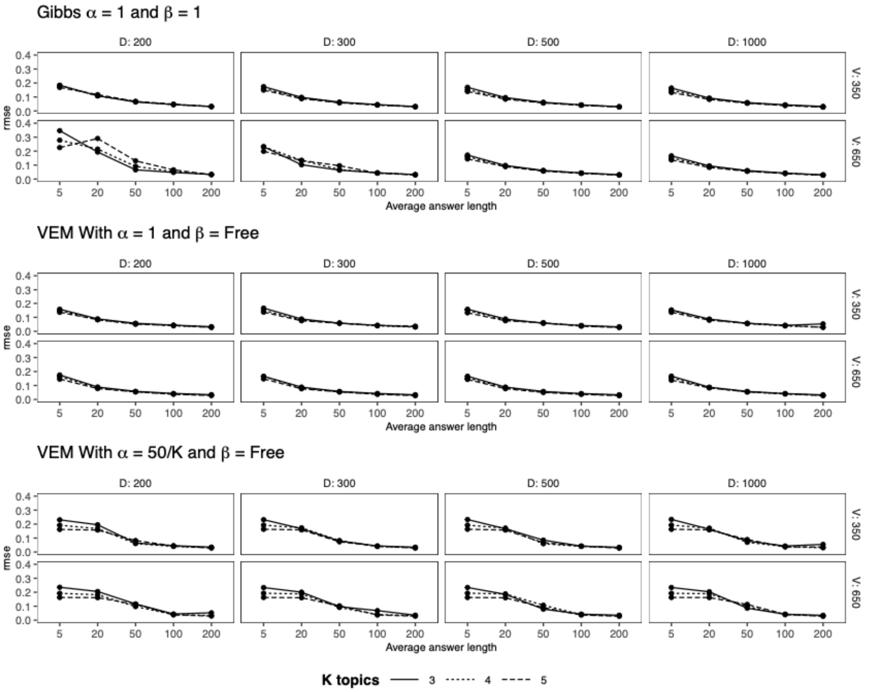
RMSE results for topic proportion for data simulated using $\vec{\alpha} = \frac{1}{K}$



Note: RMSE = Root Mean Squared Error, $D =$ number of doc number of documents in corpus, $V =$ vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, Gibbs = Gibbs sampling algorithm, VEM = variational-expectation maximization algorithm, $\vec{\alpha} = \frac{1}{K}$ hyperparameter for topic proportions used to generate the data, $\vec{\alpha} = \frac{50}{K}$ hyperparameter for topic proportions used to estimate the model, and $\vec{\beta} =$ hyperparameter for topics used to estimate the model.

Figure 4

RMSE results for topic proportion for data simulated using $\vec{\alpha} = 1$



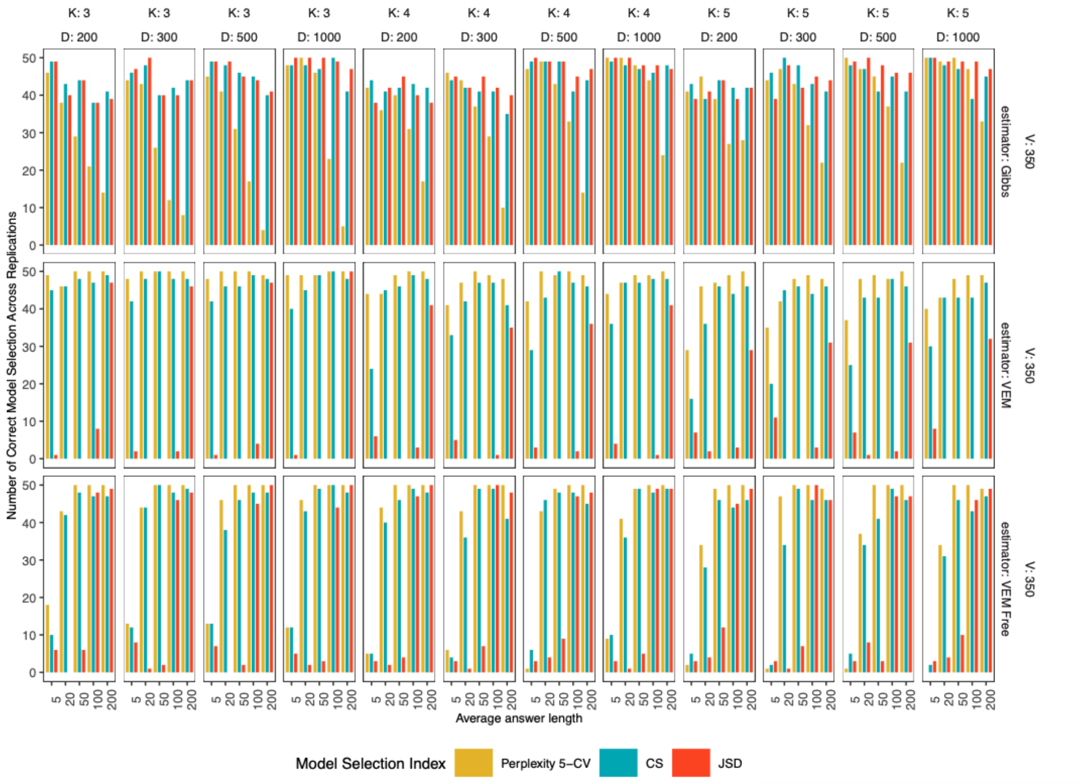
Note: RMSE = Root Mean Squared Error, $D =$ number of doc number of documents in corpus, $V =$ vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, Gibbs = Gibbs sampling algorithm, VEM = variational-expectation maximization algorithm, $\vec{\alpha} = 1$ hyperparameter for topic proportions used to generate the data, $\vec{\alpha} = \frac{50}{K}$ hyperparameter for topic proportions used to estimate the model, and $\vec{\beta} =$ hyperparameter for topics used to estimate the model.

Model Selection

Figures 5, 6, 7, 8, 9, and 10 show the number of correct model selections for each of the three indices for each algorithm and testing condition. Results suggested that the performance of similarity measures and perplexity using 5-CV vary for the estimation algorithm, the average document lengths (i.e., λ values), the number of topics, and the corpus size.

Figure 5

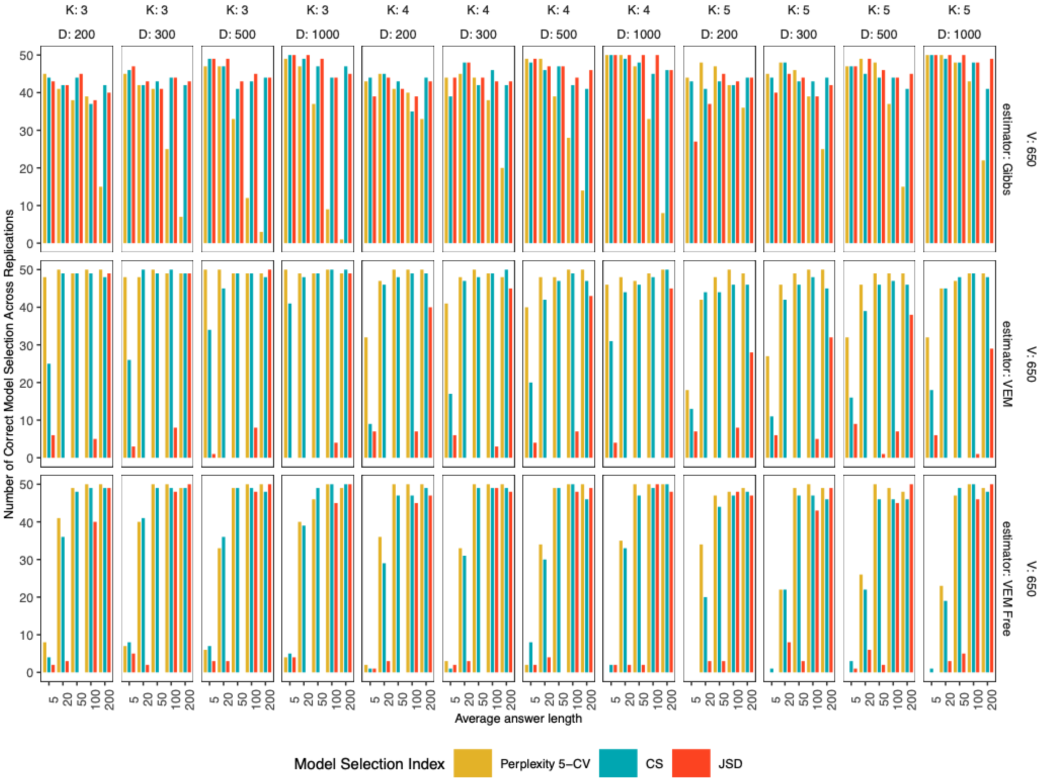
Performance model selection indices for data simulated using $\vec{\alpha} = 0.5$ *and* $V = 350$



Note: D = number of doc number of documents in corpus, V = vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, number of replications = 50, *perplexity 5-CV* = perplexity using 5-fold cross-validation, *CS* = average cosine similarity, *JSD* = average Jensen-Shannon divergence, Gibbs = Gibbs sampling algorithm estimating LDA with $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05,$ VEM = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$, VEM free = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free.$

Figure 6

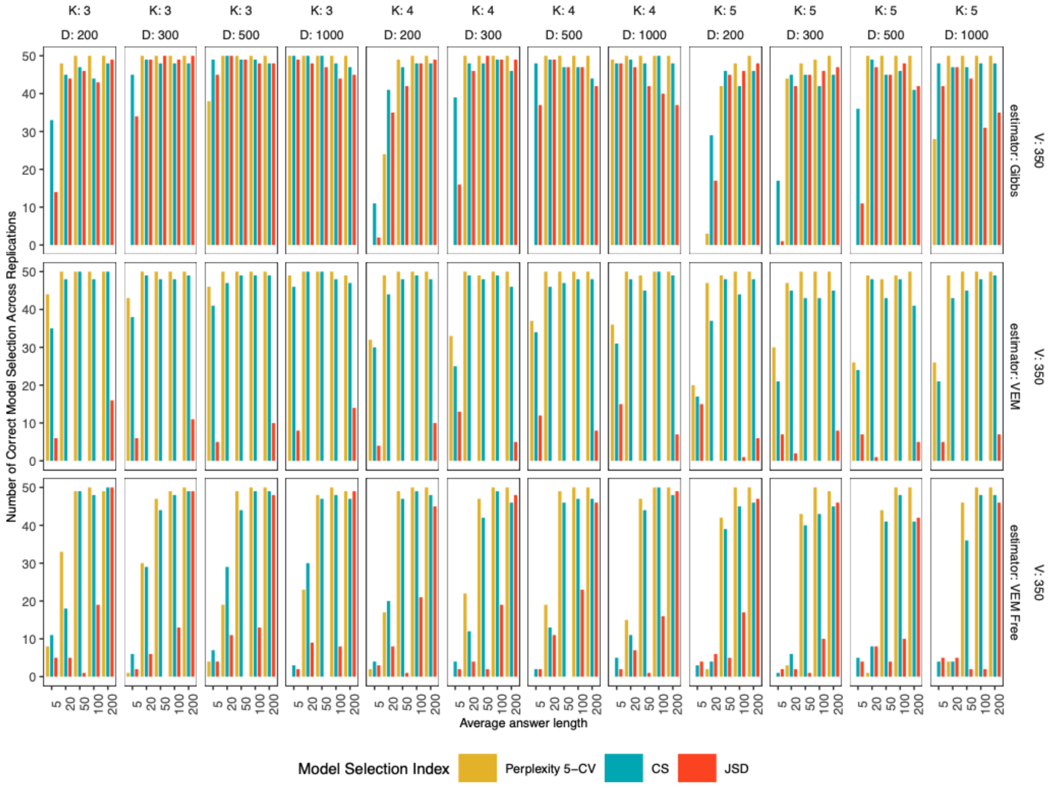
Performance model selection indices for data simulated using $\vec{\alpha} = 0.5$ *and* $V = 650$



Note: $D =$ number of doc number of documents in corpus, $V =$ vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, number of replications = 50, *perplexity 5-CV* = perplexity using 5-fold cross-validation, *CS* = average cosine similarity, *JSD* = average Jensen-Shannon divergence, Gibbs = Gibbs sampling algorithm estimating LDA with $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05,$ VEM = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$, VEM free = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free.$

Figure 7

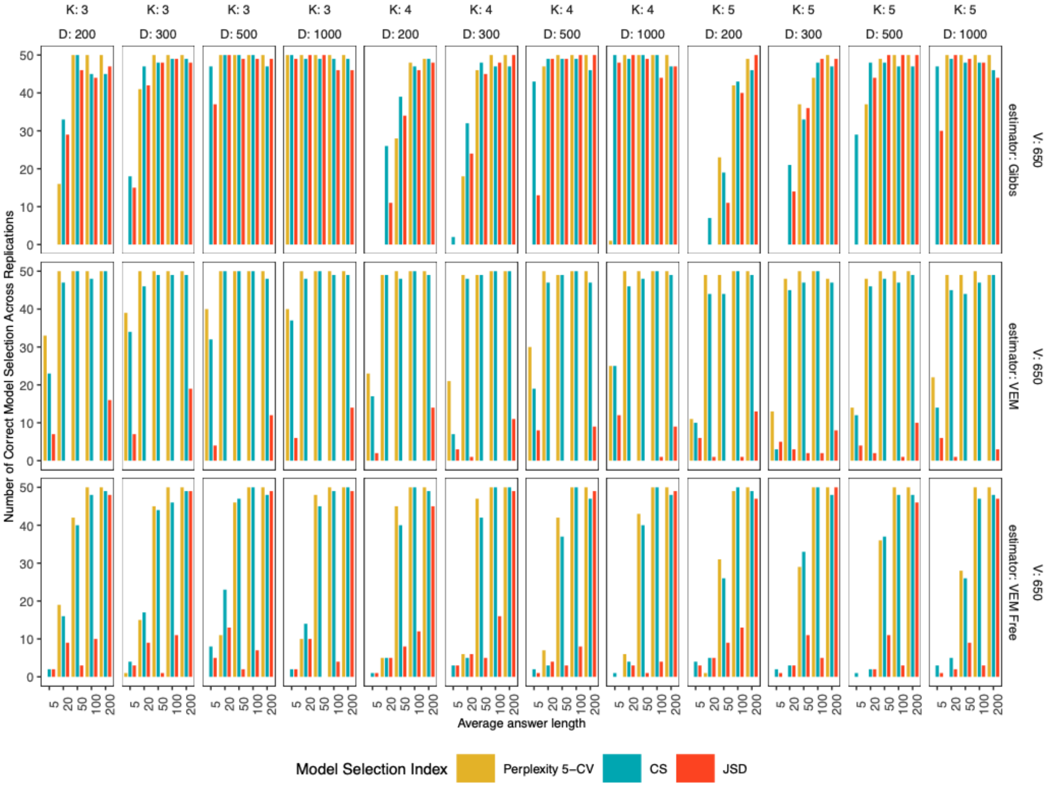
Performance model selection indices for data simulated using $\vec{\alpha} = 1$ *and* $V = 350$



Note: $D =$ number of doc number of documents in corpus, $V =$ vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, number of replications = 50, *perplexity 5-CV* = perplexity using 5-fold cross-validation, *CS* = average cosine similarity, *JSD* = average Jensen-Shannon divergence, Gibbs = Gibbs sampling algorithm estimating LDA with $\vec{\alpha} = 1$ & $\vec{\beta} = 1,$ VEM = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = 1$ & $\vec{\beta} = 1$, VEM free = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free$.

Figure 8

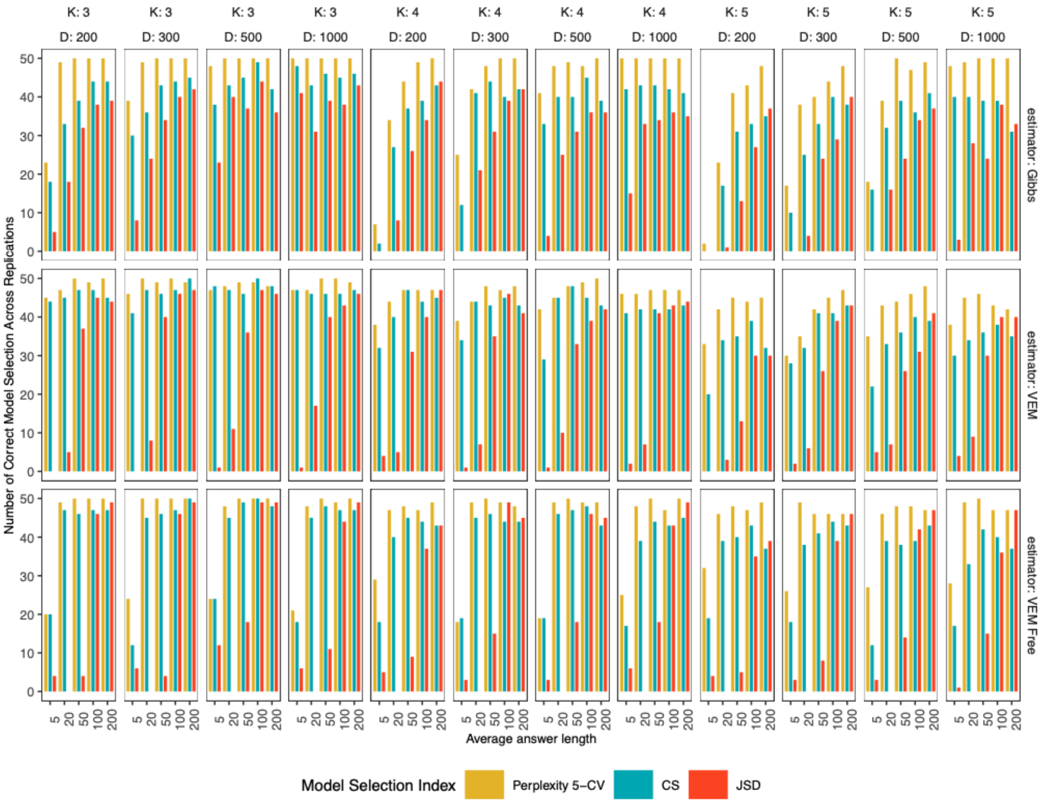
Performance model selection indices for data simulated using $\vec{\alpha} = 1$ *and* $V = 650$



Note: D = number of doc number of documents in corpus, V = vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), $K =$ true topic structure, number of replications = 50, *perplexity 5-CV* = perplexity using 5-fold cross-validation, *CS* = average cosine similarity, *JSD* = average Jensen-Shannon divergence, Gibbs = Gibbs sampling algorithm estimating LDA with $\vec{\alpha} = 1$ & $\vec{\beta} = 1,$ VEM = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = 1$ & $\vec{\beta} = 1$, VEM free = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free.$

Figure 9

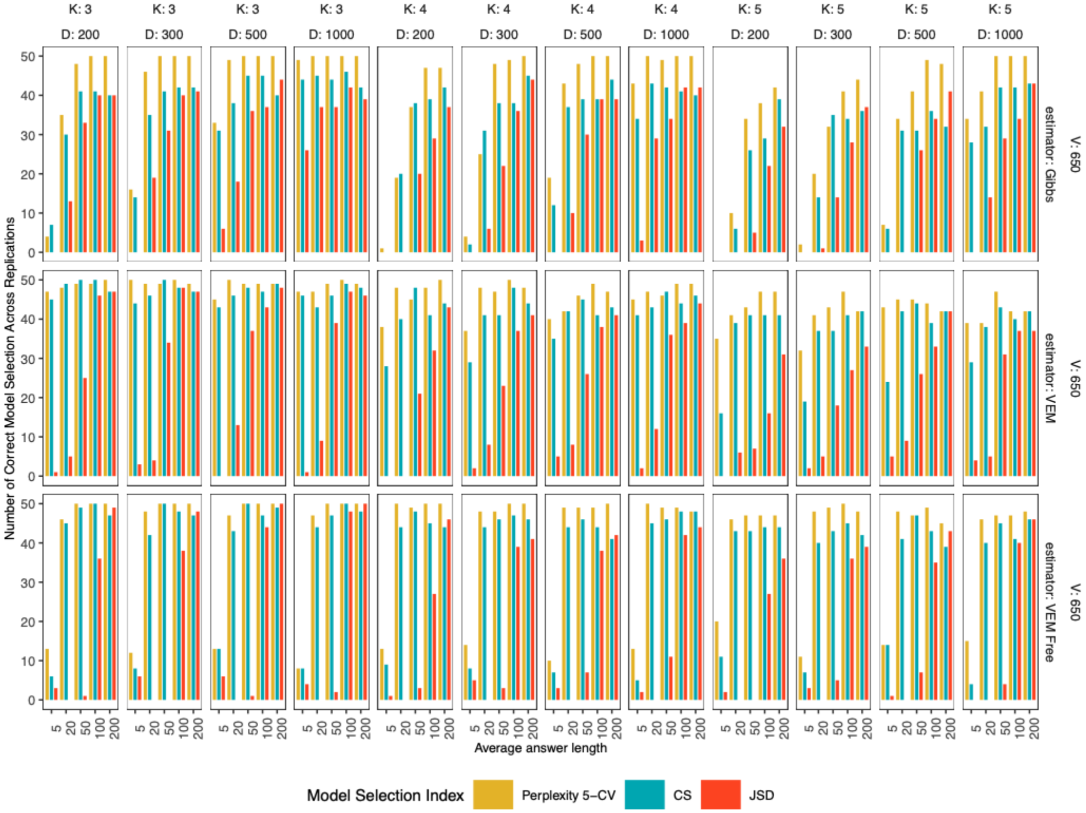
Performance model selection indices for data simulated using $\vec{\alpha} = \frac{1}{k}$ *and* $V = 350$



Note: D = number of doc number of documents in corpus, V = vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), K = true topic structure, number of replications = 50, *perplexity 5-CV* = perplexity using 5-fold cross-validation, *CS* = average cosine similarity, *JSD* = average Jensen-Shannon divergence, Gibbs = Gibbs sampling algorithm estimating LDA with $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$, VEM = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$, VEM free = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free$.

Figure 10

Performance model selection indices for data simulated using $\vec{\alpha} = \frac{1}{k}$ *and* $V = 650$



Note: D = number of doc number of documents in corpus, V = vocabulary of unique words in the corpus (i.e., number of words in the vocabulary), K = true topic structure, number of replications = 50, *perplexity 5-CV* = perplexity using 5-fold cross-validation, *CS* = average cosine similarity, *JSD* = average Jensen-Shannon divergence, Gibbs = Gibbs sampling algorithm estimating LDA with $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$, VEM = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$, VEM free = variational-expectation maximization algorithm estimating LDA with $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free$.

Results for Gibbs sampling

When LDA estimated the parameters using the Gibbs sampling algorithm, perplexity 5-CV and $\overline{CS}$ appeared to be better for model selection than $\overline{JSD}$. The accuracy for model selection of these three indices seemed to be influenced by the set of hyperparameters, the average document length, and on some occasions, the number of topics. Below we explain the performance of each model selection index under the tested conditions.

Overall, perplexity 5-CV tended to be more accurate for model selection when LDA parameters were estimated using $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$ and $\vec{\alpha} = 1$ & $\vec{\beta} = 1$. Under both sets of hyperparameters, perplexity 5-CV accuracy increased as the average document length increased. For instance, on average, perplexity 5-CV selected the simulated topic model on 73.2% (*sd* = 38.5%) of the replications for corpora with $\lambda = 20$. However, its performance improved to an average of 93.5% (*sd* = 14.5%), 98.4% (*sd* = 4.04%), and 99.8% (*sd* = 0.57%) of the replications for corpora with $\lambda = 50,$ $\lambda = 100, \lambda = 200$. Further, the results showed that the larger the number of topics, the worse the performance was. For example, on average, perplexity 5-CV detected a 3-topic model in 91.9% (*sd* = 20.3%) of the replications but decreased to 73% (*sd* = 30.7%) for a 5-topic model using $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$.

The effect of the set of hyperparameters, the average document length, and the number of topics seemed to decrease considerably for corpora containing 500 and 1,000 documents. On average, perplexity 5-CV increased its performance for detecting the simulated topic model to 94.7% (*sd* = 10.4%) and 87.7% (*sd* = 12.7%) replications for LDA models estimated using $\vec{\alpha} = 1$ & $\vec{\beta} = 1$ and $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$, respectively. This implies that perplexity 5-CV was more useful for model selection for corpora with an average document length of at least 20 words. No apparent differences were found for the shortest average length of documents.

Additionally, the results showed that perplexity 5-CV had the opposite performance for model selection when LDA was estimated using $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$. In this case, perplexity 5-CV detected the generated number of topics in 92.6% (*sd* = 5.48%) of the replications for corpora with $\lambda = 5$. Accuracy of perplexity 5-CV decreased to, however, 33.3% (*sd* = 20%) replications for corpora with $\lambda = 200$. This suggested that perplexity 5-CV was more accurate in detecting the simulated topic structure for corpora containing shorter document lengths when LDA was estimated using $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$.

The utility for model selection of both similarity measures appeared to be good across conditions for parameters estimated using $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$. For instance, $_{CS}$ selected the simulated number of topics, on average, in 89% (*sd* = 4.52%) of the replications for corpora containing 200 documents with an average document length of 5 words (i.e., $\lambda = 5$). Similarly, $\overline{JSD}$ selected the simulated number of topics, on average, in 78.3% (*sd* = 14.4%) of the replications. When LDA was estimated using $\vec{\alpha} = 1$ & $\vec{\beta} = 1$, the accuracy of both $\overline{CS}$ and $\overline{JSD}$decreased for detecting the generated number of topics for corpora with $\lambda = 5$ and $\lambda = 20$. For example, on average, $\overline{CS}$ and $\overline{JSD}$ selected the generated topic model in 59.2% (*sd* = 40.8%) and 40.9% (*sd* = 39.4%) of the replications, respectively. However, as the number of documents increased, the effect of the average answer length seemed to decrease.

For LDA models estimated using $\vec{\alpha} = 1/K$ & $\vec{\beta} = 1$, $\overline{CS}$ tended to perform better as, for example, selected the simulated number of topics for corpora containing at least 500 documents with an average document length of 20 words and whose topic structure was 3 or 4-topics. Additionally, the results of this study indicated that it was possible for $\overline{CS}$ to detect the best topic structure with corpora with $\lambda = 5$ as long as they included 1,000 documents.

Results for VEM estimation

When LDA parameters were estimated using the VEM algorithm, perplexity 5-CV and $\overline{CS}$ tended to be more useful for model selection than $\overline{JSD}$. The performance for model selection of these three indices was also influenced by the set of hyperparameters, the average document length, and in some cases, the number of topics. The performance of each model selection index under the tested conditions is described below.

In general, the performance of perplexity 5-CV tended to be better, when LDA parameters were estimated using the same hyperparameter α used to generate the data. In such a case, the accuracy of perplexity 5-CV for model selection appeared to vary across the average answer length, the number of topics, and to a lesser degree, the set of hyperparameters. For example, on average, perplexity 5-CV recovered a 3-topic model almost perfectly, regardless of the set of hyperparameters, using corpora with $\lambda = 5$. However, the accuracies of perplexity 5-CV to recover a 5-topic model decreased to 71.2% (*sd* = 8.41%), 62.5% (*sd* = 13.6%), and 40.5% (*sd* = 8.41%) when LDA estimated the parameters using $\vec{\alpha} = 1/K$ & $\vec{\beta} = 1$, $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$, and $\vec{\alpha} = 1$ & $\vec{\beta} = 1$, respectively. No apparent differences were found using corpora with various document or vocabulary sizes.

The accuracy of perplexity 5-CV for model selection decreased for LDA models estimated using $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free$. The performance of perplexity 5-CV tended to be better across conditions for data simulated using $\vec{\alpha} = \frac{1}{K}$. For example, on average, perplexity 5-CV detected the simulated topic model in 95.9% (*sd* = 2.67%) and 97.6% (*sd* = 3.01%) of the replications for corpora with $\lambda = 20$ and $\lambda = 200$, respectively. However, it decreased for corpora with $\lambda = 5$. Similar results were obtained for data generated using $\vec{\alpha} = 0.5$ as perplexity 5-CV tended to be more accurate in detecting the simulated number of topics for corpora containing average document lengths of at least 20 words ($\lambda = 20$). Further, its performance seemed to be better for detecting a 3- or 4-topic model than a 5-topic model. For instance, on average, perplexity 5-CV was accurate for selecting 3-topics in 83.2% (*sd* = 8.48%), but it decreased to 64.2% (*sd* = 16.5%) of the replications for detecting a 5-topic models. Additionally, for data generated using $\vec{\alpha} = 1$, perplexity 5-CV performed well for model selection for corpora with average document lengths of 50, 100, or 200 words.

In general, accuracy of similarity indices appeared to decrease for model selection for LDA models estimated using the VEM algorithm. Similar to perplexity 5-CV, $\overline{CS}$ and $\overline{JSD}$ were more useful for model selection when LDA used the same α value to generate and estimate the parameters. When this was the case, $\overline{CS}$ provided substantially better results than $\overline{JSD}$. The performance of $\overline{CS}$ in detecting the simulated topic model varied according to the average answer length, the number of topics, and to a lesser extent, the set of hyperparameters. For instance, on average, $\overline{CS}$ was useful for model selection in 57.1% (*sd* = 22.4 %), 88.3% (*sd* = 8.36%), 92.3% (*sd* = 6.93%), 93.4% (*sd* = 6.63%), and 92.8% (*sd* = 6.91%) of the replications for $\lambda = 5, \lambda = 20, \lambda = 100,$ and $\lambda = 200,$ respectively. These results suggest that $\overline{CS}$ performed better on average for corpora with an average document length of at least 20 words. Although it appeared to be more difficult for $\overline{CS}$ to select the simulated number of topics with corpora containing $\lambda = 5$, it was still possible to obtain good results under this condition for data simulated using $\vec{\alpha} = 0.5$ or $\vec{\alpha} = \frac{1}{K}$ for the 3-topic model.

The performance of $\overline{CS}$ decreased when LDA estimated the parameters using $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free,$ particularly for corpora containing average document lengths of 5 words. On average, $\overline{CS}$ misidentify the simulated topic model in about 85.3% (*sd* = 11.7%) of the replications for corpora with $\lambda = 5$. The effect of $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free$ decreased, however, for corpora with average document lengths of more than 20 words.

Similar to $\overline{CS},$ the accuracy of $\overline{JSD}$ was influenced by the average answer length when LDA was estimated using the same $\vec{\alpha}$ value to generate and estimate the parameters. On average, $\overline{JSD}$ performed poorly for model selection as $\overline{JSD}$ misidentified the simulated topic model in about 70% of the replications for corpora with average document lengths shorter than 200 words. However, when LDA was estimated using $\vec{\alpha} = \frac{1}{K}$ & $\vec{\beta} = 1$, the performance of $\overline{JSD}$ for model selection improved to 59.2% (*sd* = 17.7%) of the replications for corpora with $\lambda = 50$. The utility of $\overline{JSD}$ in selecting the simulated topic model also varied across average document lengths for LDA models estimated using $\vec{\alpha} = \frac{50}{K}$ & $\vec{\beta} = free$. In this case, $\overline{JSD}$ appeared to be accurate for model selection for corpora with $\lambda = 100$ and $\lambda = 200$ as, on average, $\overline{JSD}$ correctly identified the topic structure in 65.2% (*sd* = 32.7%) and 94.5% (*sd* = 5.81%) of the replications.

Discussion

LDA has primarily been used for exploratory analysis of large corpora (e.g., thousands of documents containing a relatively large number of unique words). Efforts to use LDA with classroom measurement data have been promising and have been shown to provide useful information about examinees' thinking as reflected in their responses to CR items (Cardozo-Gaibisso et al., 2019; Wheeler, Raczynski, et al., 2022). It also has been useful in identifying students' misconceptions (Shin et al., 2019). Further, it has enabled researchers to augment the estimates of ability obtained using traditional psychometric models for analyzing students' answers to CR items (Wheeler, Wang, et al., 2022).

Although LDA results are promising, studies of the accuracy of methods for selecting the best-fitting LDA model do not yet appear to have been reported. In this context, data sets from CR tests often consist of smaller numbers of documents, fewer unique words, and smaller vocabularies than those under which this method was developed. In this study, therefore, the focus was on investigating. Results from this study suggested that the recovery of the topics did not seem to rely upon the tested condition or the estimation method. Rather, results were consistent with previous studies indicating that the $\vec{\beta}$ values do not influence the estimates of the topics (Mardones-Segovia et al., 2022; Syed & Spruit, 2018). On the other hand, the recovery of the topic proportions does appear to depend on the average answer length, the set of hyperparameters, and the estimation conditions. Particularly, the results showed that the recovery of the topic proportions was lower for corpora with average document lengths of 5 and 20 words.

These results were more evident when VEM was used to estimate the $\vec{\alpha}$ (i.e., $\frac{50}{K}$) hyperparameter. Although the average answer length did affect recovery of topic proportions, when the set of hyperparameters was $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$, both the Gibbs and VEM algorithms were able to accurately recover the topic proportion even with short answers and small numbers of documents and unique words.

This study also suggested that the accuracy of the model selection indices was influenced to a greater degree by the average answer length, the estimation method, the set of hyperparameters, and the number of topics. Overall, perplexity 5-CV, $\overline{CS}$, $\overline{JSD}$ tended to be less accurate for model selection for corpora with average document lengths of 5 and 20 words. These results were consistent with previous evidence that showed that short average answer lengths might not have sufficient word co-occurrences to detect the topic structure (e.g., Chen et al., 2016; Hu et al., 2009).

Additionally, the results suggested that the selection of the sets of hyperparameters affected the performance of the model selection indices for corpora containing average document lengths of 5 and 20 words. For example, perplexity 5-CV, $\overline{CS}$, $\overline{JSD}$ did provide useful and accurate information for model selection for the shortest average document length for LDA models estimated using the Gibbs sampling with hyperparameters $\vec{\alpha} = 0.5$ & $\vec{\beta} = 0.05$. The accuracy of those model selection indices was lower, however, when VEM was used to estimate the parameters using the default hyperparameter in the R package *topicmodels* (i.e., $\vec{\alpha} = \frac{50}{K}$).

These results may suggest that although the $\vec{\beta}$ values do not impact the estimates for the topics, they appear to influence the performance of the model selection indices, as using a different $\vec{\beta}$ value to estimate the LDA models decreased the overall accuracy of the model selection indices. Similarly, these results may imply that $\vec{\alpha}$ values also appeared to influence the accuracy for model selection as the hyperparameters proposed by Griffiths and Steyvers (2004) were not useful for detecting the simulated topic model for corpora with average document lengths of 5 and 20 words. Therefore, the hyperparameters for the topics and the topic proportions appeared to be relevant for model selection purposes. It is important to mention that the influence of the set of hyperparameters tended to decrease for average document lengths of at least 50 words.

The results of this study provided useful information about the accuracy of model selection indices for detecting the best-fitting topic structure for topic modeling of CR answers in which small numbers of topics, small numbers of unique words, or shorter answer lengths are typical. The performance of model selection indices appeared to depend mainly on the average document length. The longer the average answer length, the better the accuracy of perplexity 5-CV, $\overline{CS}$, $\overline{JSD}$ for detecting the best topic structure. Evidence was also provided regarding the accuracy of Gibbs sampling and VEM as estimation algorithms. Results suggest that both the Gibbs sampling and VEM seemed to be accurate algorithms for estimating the latent topic structure. Gibbs sampling appeared to be better, in general, for corpora containing average document lengths of 5 and 20 words.

Overall, the results suggested that researchers and practitioners applying LDA to corpora of 200 or 300 documents, including short answers (e.g., 5 words or a single line), should carefully analyze the results as the data might not contain sufficient information to estimate the topics and topic proportions accurately. Further, although perplexity 5-CV and $\overline{CS}$ using Gibbs sampling tended to be more useful for model selection under those conditions, their variability was high. An important consideration is that as the corpus of documents increased, the model selection indices appeared to improve their performance for short answers to CR items. Additionally, the LDA estimates and the model selection indices were more accurate for answers to CR items that contained at least 50 words. Therefore, researchers and practitioners should consider the average answer lengths when applying LDA models to answers to CR classroom assessments. In addition to providing evidence for the performance of the three model selection methods presented in this study, R code is provided in the Appendix so that future empirical studies that apply LDA can investigate which model selection method is best for their corpora.

Although the results of this study are limited to the conditions evaluated, this research provides evidence about which model selection indices performed best for model selection in LDA models in conditions that are typical in classroom assessments. Future studies might investigate the accuracy of other model selection indices using similar, or different corpora sizes as the ones presented in this study. Additionally, due to students' answers to CR items being focused on the prompts, topics may be more related than in less constrained kinds of text. Therefore, empirical evidence would be useful to evaluate $\overline{CS}$ and $\overline{JSD}$ in these conditions.

Aknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant No. 1920407. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additionally, this material was funded, in part, by the Doctorado en el Extranjero Becas Chile/2019-7200118.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike,* 199–213.

Anderson, D., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. *Educational Measurement: Issues and Practice, 39*(4), 53–64.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. *Pacific-Asia conference on knowledge discovery and data mining,* 391–402.

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement, 74*(5), 795–808.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics, 1*, 391–402.

Berrar, D. (2019). Cross-validation. In S. Ranganathan, M. Gribskov, K. Nakai, &

C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (pp. 542–545). Academic Press.

Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th international conference on machine learning (icml-12),* 201–208.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association, 112*(518), 859–877.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*.

Buxton, C., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S., Choi, Y.-J., & Cohen, A. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of language of scientific investigation practices. *Double Helix, 2*(1), 1–21.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing, 72*(7-9), 1775–1781.

Cardozo-Gaibisso, L., Kim, S., Buxton, C., & Cohen, A. (2019). Thinking beyond the score: Multidimensional analysis of student performance to inform the next generation of

science assessments. *Journal of Research in Science Teaching*, *1*(57), 856–878.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems, 22*.

Chen, Q., Yao, L., & Yang, J. (2016). Short text classification based on lda topic model. *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, 749–753.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain monte carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83*(2), 278–306.

Cohen, A. S., & Cho, S.-J. (2017). Information criteria. In *Handbook of item response theory volume 2: Statistical tools* (pp. 363–378). CRC Press.

Copur-Gencturk, Y., Choi, H.-J., & Cohen, A. (2022). Investigating teachers' understanding through topic modeling: A promising approach to studying teachers' knowledge. *Journal of Mathematics Teacher Education*, 1–22.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique, 17*(1), 61–84.

Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), 16–34.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences, 101*(1), 5228–5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Hasan, M., Rahman, A., Karim, M., Khan, M., Islam, S., Islam, M., et al. (2021). Normalized approach to find optimal number of topics in latent dirichlet allocation (lda). *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*, 341–354.

Hu, X., Sun, N., Zhang, C., & Chua, T.-S. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. *Proceedings of the 18th ACM conference on Information and knowledge management*, 919–928.

Hübner, R., & Pelzer, T. (2020). Improving parameter recovery for conflict drift-diffusion models. *Behavior Research Methods*, *52*(5), 1848–1866.

Kang, T., & Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358.

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*, 82–102.

Lau, J. H., Baldwin, T., & Newman, D. (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing (TSLP)*, *10*(3), 1–14.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous irt models. *Applied Psychological Measurement*, *33*(5), 353–373.

Lockwood, J. (2014). Handbook of automated essay evaluation current applications and new directions mark d. shermis and jill burstein (eds.)(2013) new york: Routledge. pp. 194 isbn: 9780415810968. *Writing and Pedagogy*, *6*(2), 437–441.

Mardones-Segovia, C., Choi, H.-J., Hong, M., Wheeler, J. M., & Cohen, A. S. (2022). Comparison of estimation algorithms for latent dirichlet allocation. *Quantitative Psychology: The 86th Annual Meeting of the Psychometric Society, Virtual, 2021*, 27–37.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 conference on empirical methods in natural language processing*, 262–272.

Myung, J. I., & Pitt, M. A. (2004). Model comparison methods. In *Methods in enzymology (pp.* 351–366). Elsevier.

Neishabouri, A., & Desmarais, M. C. (2020). Reliability of perplexity to find number of latent topics. *The Thirty-Third International Flairs Conference*.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108.

Nikita, M. (2019). *Ldatuning: Tuning of the latent dirichlet allocation models parameters* [R package version 1.0.0]. https://CRAN.R-project.org/package=ldatuning Nikolenko, S. I. (2016). Topic quality metrics based on distributed word representations. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1029–1032.

Ponweiser, M. (2012). *Latent dirichlet allocation in r* (Doctoral dissertation). https://epub.wu.ac.at/3558/1/main.pdf

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, *5*, 532–538.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Roque, C., Cardoso, J. L., Connell, T., Schermers, G., & Weber, R. (2019). Topic analysis of road safety inspections using latent dirichlet allocation: A case study of roadside safety in irish main roads. *Accident Analysis & Prevention*, *131*, 336–349.

Schröder, N., Falke, A., Hruschka, H., & Reutterer, T. (2017). Analyzing browsing and purchasing across multiple websites based on latent dirichlet allocation. *ALLDATA 2017*, *40*, 44.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343.

Sen, S., & Cohen, A. S. (2019). Applications of mixture irt models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in psychology*, *10*, 825.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Syed, S., & Spruit, M. (2018). Selecting priors for latent dirichlet allocation. *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 194–202.

Taddy, M. (2012). On estimation and selection for topic models. *Artificial Intelligence and Statistics*, 1184–1193.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. *International*

Conference on Machine Learning, 190–198.

Vu, H. Q., Li, G., & Law, R. (2019). Discovering implicit activity preferences in travel itineraries by topic modeling. *Tourism Management*, *75*, 435–446.

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking lda: Why priors matter. *Advances in neural information processing systems*, 1973–1981.

Wang, Y., Luo, J., Niemi, R., Li, Y., & Hu, T. (2016). Catching fire via" likes": Inferring topic preferences of trump followers on twitter. *arXiv preprint arXiv:1603.03099*.

Wesslen, R. (2018). Computer-assisted text analysis for social science: Topic models and beyond. *arXiv preprint arXiv:1803.11045*.

Wheeler, J. M., Cohen, A. S., Xiong, J., Lee, J., & Choi, H.-J. (2021). Sample size for latent dirichlet allocation of constructed-response items. *Quantitative Psychology: The 85th Annual Meeting of the Psychometric Society, Virtual*, 263–273.

Wheeler, J. M., Engelhard, G., & Wang, J. (2022). Exploring rater accuracy using unfolding models combined with topic models: Incorporating supervised latent Dirichlet allocation. *Measurement: Interdisciplinary Research and Perspectives*, *20*(1), 34–46. Wheeler, J. M., Raczynski, K., Cohen, A. S., & Engelhard Jr, G. (2022). Using topic models to understand rater-mediated writing assessments. *The Journal of Experimental Education, 1–20.*

Wheeler, J. M., Wang, S., Tan, Y., & Cohen, A. (2022). Textual data as process data: A new scoring procedure for mixed-format assessments.

Wheeler, J. M., Xiong, J., Mardones-Segovia, C., Choi, H.-J., & Cohen, A. S. (2022). An investigation of prior specification on parameter recovery for latent dirichlet allocation of constructed-response items. *Quantitative Psychology: The 86th Annual Meeting of the Psychometric Society, Virtual, 2021*, 203–215.

Wild, F. (2020). *Lsa: Latent semantic analysis* [R package version 0.73.2]. https://CRAN.R-project.org/package=lsa

Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science, 171*, 109203.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, *48*(2), 379–398.