Choosing the Number of Topics in LDA Models – A Monte Carlo Comparison of Selection Criteria[1]

Victor Bystrov
University of Lodz
Rewolucji 1905r. 41, 90-214 Lodz, Poland
email: victor.bystrov@uni.lodz.pl

Viktoriia Naboka-Krell
Justus Liebig University Giessen
Licher Strasse 64, 35394 Giessen, Germany
email: viktoriia.naboka@wirtschaft.uni-giessen.de

Anna Staszewska-Bystrova
University of Lodz
Rewolucji 1905r. 37/39, 90-214 Lodz, Poland
email: anna.bystrova@uni.lodz.pl

Peter Winker
Justus Liebig University Giessen
Licher Strasse 64, 35394 Giessen, Germany
email: peter.winker@wirtschaft.uni-giessen.de

Abstract

Selecting the number of topics in LDA models is considered to be a difficult task, for which alternative approaches have been proposed. The performance of the recently developed singular Bayesian information criterion (sBIC) is evaluated and compared to the performance of alternative model selection criteria. The sBIC is a generalization of the standard BIC that can be implemented to singular statistical models. The comparison is based on Monte Carlo simulations and carried out for several alternative settings, varying with respect to the number of topics, the number of documents and the size of documents in the corpora. Performance is measured using different criteria which take into account the correct number of topics, but also whether the relevant topics from the DGPs are identified. Practical recommendations for LDA model selection in applications are derived.

Key Words: Topic models, text analysis, latent Dirichlet allocation, singular Bayesian information criterion, Monte Carlo simulation, text generation
JEL classification: C49

---

1 Introduction

Text data have been increasingly used in different applications lately. One of the main challenges in working with text data is to structure and to quantify these data. To this end, probabilistic topic modelling approaches are often applied, as they allow to uncover hidden structures behind text data. One of the best-known and widely used topic modelling approaches is Latent Dirichlet Allocation (LDA) introduced by Blei, Ng and Jordan (2003). For some recent applications making use of this method, see, e.g., Lüdering and Winker (2016), Thorsrud (2020), Ellingsen, Larsen and Thorsrud (2022), and Savin and Teplyakov (2022).

LDA is an unsupervised method that builds on two basic assumptions. First, it is assumed that each document in a corpus represents a mixture of topics. The second assumption is that each topic is given by a mixture of words from the vocabulary. The number of these topics, or themes, is a parameter to be set by the researcher. Often this decision is based on human/expert judgment and is, therefore, rather subjective. In order to account for possible subjectivity and to allow for a more standardised estimation procedure, different evaluation metrics have been developed for identifying an optimal number of topics in LDA models. Some of them aim to minimize the similarity of different topics (Cao, Xia, Li, Zhang and Tang, 2009), maximize the topic coherence (Mimno, Wallach, Talley, Leenders and Mc-Callum, 2011) or maximize the goodness-of-fit between the estimated and the actual document-word frequencies (Lewis and Grossetti, 2022). These criteria, however, often result in (substantially) different numbers when applied to the same corpus. Their performance might also differ across corpora depending on the underlying data set (see examples in Section 2). Bystrov, Naboka, Staszewska-Bystrova and Winker (2022) propose to use a new measure for selecting an optimal number of topics, namely the singular Bayesian information criterion (sBIC). This information criterion reflects the trade-off between goodness-of-fit and model complexity and showed promising results in a first application.

There have been some attempts to compare selected criteria based on individual real datasets.[2] In this paper, a comprehensive Monte Carlo (MC) simulation is proposed, which allows a systematic evaluation going beyond individual case reports by using a large number of datasets coming from well defined data generating processes (DGP) with known properties. Thereby, we consider three different data generating processes to reflect different types of text data commonly used in applications. In a first step, we generate corpora with a known (true) number of topics, to which LDA models with different numbers of topics are fitted. Then, we apply the metrics to select the number of topics and evaluate their performance over many MC replications. To the best of our knowledge, no such systematic and comprehensive comparison analysis of the metrics used for choosing the number of topics in LDA models

---

[2]A notable exception including also a small scale Monte Carlo simulation is Lewis and Grossetti (2022).

has been performed yet.

The contribution of this paper is threefold. First, with the sBIC we implement a new measure for identifying the true number of topics in LDA models. Second, we perform a proper MC simulation study to evaluate the proposed criterion as well as other evaluation criteria commonly used in applications.[3] Third, we evaluate the considered metrics quantitatively and qualitatively, i.e., we consider whether the actual number of topics is approximated well, and also the content and structure of the estimated topics.

The remainder of this paper is structured as follows. The considered model selection criteria are described in Section 2. Section 3 presents the design and the implementation details of the MC simulations. The results of the MC simulations for three different DGPs are presented in Section 4, which is divided in two subsections to address the main trade-off between *number of topics* and *coherence/structure* of the uncovered topics. The final section summarises the findings and provides recommendations for applications.

2 Model Selection Criteria for LDA

The selection of the optimal number of topics for LDA models can be based either on measures of topic quality (similarity or coherence) or on measures of goodness-of-fit and model complexity.

Let us consider an LDA model under a standard "bag-of-words" assumption. For a document corpus $\mathcal{D}$ that consists of J documents, each document j ($j = 1, 2, \ldots, J$) is a set of N_j words, where the ordering of words is ignored. The total number of words in the corpus is equal to $N = \sum_{j=1}^{J} N_j$. The document corpus $\mathcal{D}$ can be characterized by a $J \times I$ document-word frequency matrix $X = \{x_{ji}\}_{j,i=1}^{J,I}$, where x_{ji} is the frequency of word i encountered in document j and I is the number of different words in the vocabulary.

Under the "bag-of-words" assumption, an LDA model can be summarized by a $J \times K$ matrix θ of document-topic frequencies and a $K \times I$ matrix β of topic-word frequencies with the dimensions of these matrices depending on the number of topics K. The estimated document-term matrix is a product of estimates $\hat{\theta}$ and $\hat{\beta}$: $\hat{X} = \hat{\theta} \times \hat{\beta}$. A set of candidate LDA models is determined by the numbers of topics in candidate models: $K \in \{K_{min}, \ldots, K_{max}\}$.

In the following, we describe two popular semantic measures of topic quality, which are often used in applications, and two recently developed goodness-of-fit measures.

2.1 Topic Similarity

Following Cao et al. (2009), the optimal number of topics is often selected by minimizing the average cosine similarity across topics:

---

[3]This version of the paper is preliminary regarding the limited number of replications conducted for the Monte Carlo simulations. Due to constraints in available computational resources, current work focuses on extending the number of replications to 1 000 for all setups.

$$Cao\_Juan(K) = \frac{\sum_{k=1}^{K} \sum_{l=k+1}^{K} corr(k,l)}{K \times (K-1)/2},$$

where

$$corr(k,l) = \frac{\sum_{i=1}^{I} \beta_{ki}\beta_{li}}{\sqrt{\sum_{i=1}^{I} \beta_{ki}^2}\sqrt{\sum_{i=1}^{I} \beta_{li}^2}},$$

and β_{ki} is the frequency of word type i in topic k.

The average cosine similarity is extensively used for selecting the number of topics in different text-as-data applications, e.g. analyzing scientific articles to examine the evolution of research over time and identify future fields of research (Loureiro, Guerreiro and Tussyadiah, 2021; Tiba, Rijnsoever and Hekkert, 2018), analyzing the speeches by Executive Board members of the European Central Bank (Hartmann and Smets, 2018), investigating news data in the context of economic reforms (Lin and Katada, 2022), analyzing and categorizing innovation projects (Dahlke, Bogner, Becker, Schlaile, Pyka and Ebersberger, 2021).

2.2 Topic Coherence

Mimno et al. (2011) proposed a model selection procedure that maximizes the average semantic coherence of topics:

$$Mimno(K) = \frac{1}{K} \sum_{k=1}^{K} coh(k, \mathbf{i}^{(k)}),$$

where $coh(k, \mathbf{i}^{(k)})$ is the coherence metric for topic k,

$$coh(k, \mathbf{i}^{(k)}) = \frac{2}{v \times (v-1)} \sum_{m=2}^{v} \sum_{n=1}^{m-1} \log \frac{f(i_m^{(k)}, i_n^{(k)}) + \epsilon}{f(i_n^{(k)})},$$

$\mathbf{i}^{(k)} = (i_1^{(k)}, \ldots, i_v^{(k)})$ is the list of the v most frequent word types in topic k, $f(i)$ is the document frequency of word i (i.e., the number of documents with at least one token of type i), and $f(i, i')$ is the co-document frequency of words i and i' (i.e., the number of documents containing one or more tokens of type i and at least one token of type i'). The default number of the most probable words used is equal to 20. The smoothing parameter ϵ is included to avoid taking the logarithm of zero and its default value is equal to e^{-12}.

The average semantic coherence is also often used for selecting the number of topics in applied topic mining analyzing, e.g., monetary policy speeches (Ferrara, Masciandaro, Moschella and Romelli, 2022), news data to forecast aggregated stock returns (Adämmer and Schüssler, 2020), energy market tweets with regard to their impact on market movements (Polyzos and Wang, 2022), or survey responses on the consequences of Covid-19 pandemic (Kleinberg, van der Vegt and Mozes, 2020).

2.3 OpTop Criterion

Lewis and Grossetti (2022) proposed to use a goodness-of-fit statistic based on the comparison of actual and estimated document-word frequencies. The frequency of word type i in document j estimated in an LDA model with K topics is

$$\hat{x}_{ji}^{(K)} = \sum_{k=1}^{K} \hat{\theta}_{jk}^{(K)} \hat{\beta}_{ki}^{(K)}.$$

Because the matrix of document-word frequencies is usually sparse, Lewis and Grossetti (2022) suggest collapsing relatively unimportant words in a single frequency bin. For document j, they order word types from the smallest to the largest estimated frequency, $(i_1^{(j)}, i_2^{(j)}, \ldots, i_I^{(j)})$ such that $\hat{x}_{ji_1}^{(K)} \leq \hat{x}_{ji_2}^{(K)} \leq \ldots \leq \hat{x}_{ji_I}^{(K)}$, and select a sub-vector of relatively unimportant word types $(i_1^{(j)}, i_2^{(j)}, \ldots, i_p^{(j)})$. The cumulative frequency of relatively unimportant word types in document j estimated in an LDA model with K topics is

$$\hat{x}_{j,min}^{(K)} = \sum_{i \in (i_1^{(j)}, \ldots, i_p^{(j)})} \hat{x}_{ji}^{(K)},$$

where $\hat{x}_{ji_p}^{(K)}$ is the largest frequency such that $\sum_{i=i_1^{(j)}}^{i_p^{(j)}} \hat{x}_{ji}^{(K)} < x_{cutoff}$, and x_{cutoff} is a cumulative frequency cut-off value. Following Lewis and Grossetti (2022), we use $x_{cutoff} = 0.05$ as a baseline cut-off value. (For a robustness check, we also consider a cut-off value of 0.20) The resulting goodness-of-fit statistic is

$$OpTop(K) = \sum_{j=1}^{J} \left[(P_j + 1) \left(\sum_{i \in (i_{p+1}^{(j)}, \ldots, i_I^{(j)})} \frac{(\hat{x}_{ji}^{(K)} - x_{ji})^2}{\hat{x}_{ji}^{(K)}} + \frac{(\hat{x}_{j,min}^{(K)} - x_{j,min})^2}{\hat{x}_{j,min}^{(K)}} \right) \right],$$
$$(2.1)$$

where $(i_{p+1}^{(j)}, \ldots, i_I^{(j)})$ is a sub-vector of relatively important word types in the jth document and P_j is the length of this sub-vector. Lewis and Grossetti (2022) propose to select an optimal number of topics by minimizing the OpTop statistic (2.1) over a range of numbers of topics. Unlike criteria proposed by Cao et al. (2009) and Mimno et al. (2011), the OpTop statistic is not a semantic measure of topic quality, but a goodness-of-fit measure that can be easily computed.

2.4 Singular Bayesian Information Criterion

The last model selection criterion – singular Bayesian information criterion is a version of the Bayesian information criterion (BIC) which can be applied

to singular statistical models (Hayashi and Watanabe, 2020), for which it is more suitable for model selection than BIC. The method was successfully applied by Bystrov et al. (2022) for selecting parsimonous LDA models with coherent topics, however the properties of the criterion as applied to LDA modelling have not been studied in a simulation setup.

Computation of sBIC is based on several results from the literature. The first one, is the decomposition of log-marginal likelihood of a text corpus $\mathcal{D}$ with K topics, described by Watanabe (2009). In the context of an LDA model, this representation can be written as

$$\log L(\mathcal{D}|K) = \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K) - \lambda(K) \log(N) + (m(K)-1) \log \log(N) + O_p(1),$$

where $\hat{\theta}$ and $\hat{\beta}$ are consistent estimators of the document-topic and topic-word probability matrices, respectively, $\lambda(K)$ is a learning coefficient, measuring stochastic complexity of a model with K topics, $m(K)$ is its multiplicity and N stands for the total number of words in the corpus.

In the next step, to deal with the problem of unknown values of $\lambda(K)$ and $m(K)$ which depend on the true value of K, model averaging described in Drton and Plummer (2017) is applied. In this approach, the singular Bayesian information criterion for an LDA model with K topics can be defined as an approximation of the log-marginal likelihood obtained by averaging of models with smaller number of topics (see Drton and Plummer (2017)):

$$sBIC(K) = \log L'(\mathcal{D}|K), \tag{2.2}$$

where $L'(\mathcal{D}|K)$ denotes the approximation following from the Drton and Plummer (2017) procedure. To compute the marginal likelihood for every sub-model with the number of topics k where $k \leq K$, the formulas for the learning coefficient $\lambda(k)$ and its multiplicity $m(k)$ derived for LDA by Hayashi (2021) are used. Model selection is then based on maximizing the sBIC value.

As described by Bystrov et al. (2022) evaluation of (2.2) for some datasets may be associated with numerical problems resulting from very small values of the likelihood function. In order to avoid these problems, in the experiments we use high-precision computations.

3 Monte Carlo Simulation

Despite the broad usage of the evaluation metrics described in the previous section, there is no consensus yet on which metric performs best, when it comes to choosing the number of topics. Given that the ground truth, i.e., the actual data generating process is unknown in real applications, the performance of the metrics can only be assessed based on a subjective analysis of the uncovered themes. To account for this problem, a Monte Carlo simulation study is required, for which the data are generated by a well defined

DGP with a known number of different topics.[4] This allows not only to compare the performance of alternative metrics with regard to the number of topics identified, but also to evaluate whether certain characteristics of the corpora such as number or length of documents might affect the relative performance. Furthermore, it enables to evaluate not only the number of topics identified, but also whether these topics correspond closely to the topics underlying the DGP, i.e., focusing also on the content in an objective approach.

This section provides the details of the Monte Carlo simulation setup used for the comparison of the methods described in Section 2. First, in Subsection 3.1 we present the general framework that is applied for each of three different DGPs. Second, in Subsection 3.2 we describe the DGPs, which are derived from actual corpora with typical characteristics of textual data used in applications. Finally, Subsection 3.3 provides some technical implementation details.

3.1 Procedure

The three DGPs used in the Monte Carlo simulations are designed to replicate the characteristics of a given real document corpus. Figure 1 presents the generic procedure which is applied to each of these DGPs.
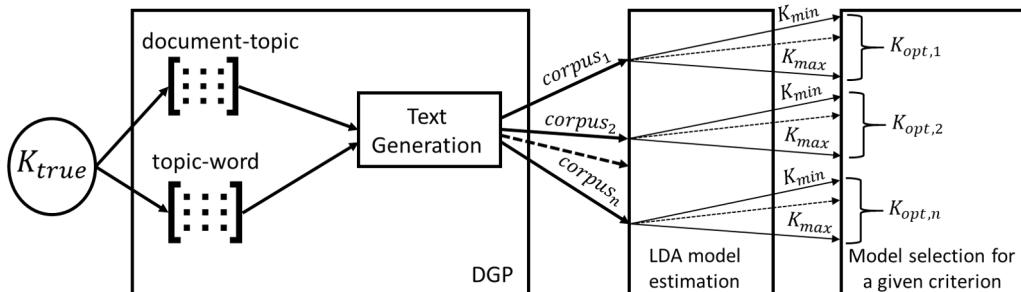

Figure 1: Generic procedure for Monte Carlo simulations with a given selection criterion

As described in Section 2, LDA is based on the assumption that each document in a corpus is a distribution over a given number K_{true} of latent topics and each topic is a distribution over a fixed corpus vocabulary (Blei et al., 2003). Thus, an LDA model is described by two matrices, the first containing the probabilities of occurrence of each word in each topic (topic-word distribution), and the second providing the probabilities of each topic occurring in a single document (document-topic distribution). The approach for generating Monte Carlo text corpora is based on these two matrices.

---

[4]The idea of using Monte Carlo simulations for obtaining well-defined text corpora has been applied recently by Wang, Zhang, Li, Deng and Liu (2021) in the context of model selection for text classification tasks. The authors use the generated data to evaluate the classification performance of different topic models.

Therefore, in the first step of the procedure, an LDA model is estimated using a given real document corpus with a number of topics that was used in previous analysis of the selected corpus. In order to make sure that only distinct topics will be used for the generation of Monte Carlo text corpora in the following step, topics exhibiting a cosine similarity measure with other topics larger than a selected cut-off value (95% or 99% percentile) are dropped (see Appendix A for more details). This correction allows to reduce a potential bias that can be induced by high topic similarity observed in the generated data and ensures a more stable performance of all model selection criteria.

For the remaining K_{true} topics, the document-topic matrix is re-scaled to ensure that topic weights add up to one and passed on to the second step of text generation. The text generation process based on LDA is presented in Algorithm 1. For each document in the original corpus, a new Monte Carlo document is created with the same number of words and document-topic distribution. For each word in this document, first, a topic is randomly selected based on the known document-topic distribution. Then, the word is randomly selected from the known vocabulary using the known topic-word distribution. It should be noted that the algorithm as presented does not exactly reproduce the generative procedure described in Blei et al. (2003), where document-topic and topic-word frequency matrices are obtained using hyper parameters. In applications, these hyper parameters are not often estimated, and using flat priors would result in distributions of document-topic and topic-word frequencies different from the ones actually observed. Therefore, we follow a data-driven approach and try to replicate closely the properties of the observed datasets serving as the benchmarks for our Monte Carlo corpora and, consequently, use the estimated document-topic and topic-word frequency matrices.

Algorithm 1 Text generation

1: **for** $document = 1, 2, \ldots, J$ **do**
2:      $document\_length = original\_document\_length$
3:      **for** $word = 1, 2, \ldots, document\_length$ **do**
4:          Randomly select a topic from the document-topic distribution
             of the current document
5:          Randomly select a word from the topic-word distribution
6:          Append the selected word to the current document
7:      **end for**
8:      Append the generated document to the corpus.
9: **end for**

Algorithm 1 is applied to each DGP with 500 Monte Carlo replications, i.e., 500 corpora containing the same number of documents of same length as the original corpus.

In the third step, for each criterion that we use, we estimate the LDA model with number of topics within the interval $[\max\{2; K_{true} - 20\}, K_{true} + 20]$, where K_{true} is the number of topics used when generating text corpora

from the DGPs. The maximum length of the range of admitted values for the number of topics is equal to 40 with the true number of topics K_{true} of the DGP in the center of the interval if larger than 20. Otherwise, the lower bound is set to 2, the lowest sensible number of topics. This limited range of admitted values for the number of topics is due to the high computational costs of model estimation. The optimal number of topics is determined for each of the selected criteria based on these estimated models.

For the final step, the comparison of the outcomes of different model selection criteria in Subsection 4.1, we use descriptive statistics such as standard deviation, mean, median, and skewness. For the visualisation of the distributions over the number of topics determined according to the considered criteria, we use histograms. Furthermore, in Subsection 4.2, we will also provide information about the extent to which the topics used for generating the texts are found when applying LDA with the number of topics selected by the different criteria.

3.2 Data Generation Processes

The three DPGs used for the Monte Carlo simulations are related to three real world corpora:

- DGP 1 replicates the characteristics of a corpus consisting of scientific papers published in the Journal of Economics and Statistics (JES).

- DGP 2 reproduces features of the corpus consisting of abstracts submitted to European Research Consortium for Informatics and Mathematics (ERCIM) and Computational and Financial Econometrics (CFE) conferences.

- DGP 3 reproduces the properties of a corpus containing Newsticker items from heise online.

The data from JES used for DGP 1 cover the period from 1984 to 2020 and consists of 704 documents with an average text length of about 3,000 words. The size of the vocabulary for this corpus is equal to 3,911 words. The collection focuses on scientific publications in empirical economics and applied statistics. The initial number of topics selected was equal to 60 as in Bystrov et al. (2022). After removing topics which were too similar, the final number of topics used in DGP 1 is equal to 38 ($K_{true} = 38$).

The conference abstract data used for DGP 2 cover the period from 2007 to 2019 and consists of 11,387 documents with an average text length of about 80 words. For this corpus the dictionary is composed of 1,796 words. The focused nature of conference abstracts allows to expect a limited number of topics. The initial number of topics selected for these data was equal to 20. This number is reduced to 12 ($K_{true} = 12$) after removing the topics that were too close to each other.

The heise data used for DGP 3 cover the period from 1996 to 2021 and include 181,402 documents with an average length of about 120 words. The

number of words in the vocabulary for this corpus is equal to 4,675. The news platform discusses a significant number of topics concerning technological advances. The initial number of topics selected was equal to 120. Keeping the most distinct topics, the final number of topics used in DGP 3 was equal to 70 ($K_{true} = 70$). In the analysis we use only the most recent 50,000 documents from this corpus because using the whole dataset would increase the computational costs for the Monte Carlo simulation beyond the available capacities.

3.3   Details of Implementation

All Monte Carlo simulations were implemented using Python. To generate random sequences used in the text generation stage (Algorithm 1), the random number generator from Pythons' numpy package was used (`https://numpy.org/doc/stable/reference/random/generator.html`). LDA models were estimated using the Gibbs sampler as implemented in the Python package "lda" (`https://pypi.org/project/lda/`). For each corpus generated based on DGP 1, models with topic numbers in the interval $[18; 58]$ were estimated, for corpora generated based on DGP 2 - in the interval $[2; 32]$, and for corpora generated based on DGP 3 - in the interval $[50; 90]$. Most other parameters of the package were used at the default values. The number of iterations was set to a relatively small value of $1\,000$ due to computational constraints.

The Cao_Juan and Mimno criteria were computed using the Python package "tmtoolkit" (`https://pypi.org/project/tmtoolkit/`). The Python implementations of the sBIC and OpTop model selection criteria were written by the authors. Computations were performed using the high-performance-computing-cluster at Justus Liebig University Giessen (justHPC) (`https://www.hkhlr.de/de/cluster/justhpc-giessen`).[5]

4   Results

This section summarizes the results of the Monte Carlo simulations. It is divided into two subsections. The first (4.1) presents and discusses the results on finding the optimal (true) number of topics. The second one (4.2) focuses on topics' contents and structure, and proposes a procedure to analyse/evaluate both.

4.1   Number of Topics

The first set of results concerns the estimation of the number of topics K. Figures 2, 3 and 4 present histograms for the numbers of themes selected by the evaluation metrics described in Section 2 for all three considered DGPs. In each of the histograms, the red vertical line depicts the true number of

---

[5]Code details can be found in the Github repository for this paper at `https://github.com/VikaNa/sBIC` .

topics (K_{true}) used for generating the corpora. Table 1 provides selected descriptive statistics computed for these estimates. The shape and location of histograms shown in Figures 2-4 suggest that sBIC is clearly the best method for selecting the number of topics for DGP 1 and DGP 2, while it performs similarly to the method of Cao et al. (2009) for DGP 3.
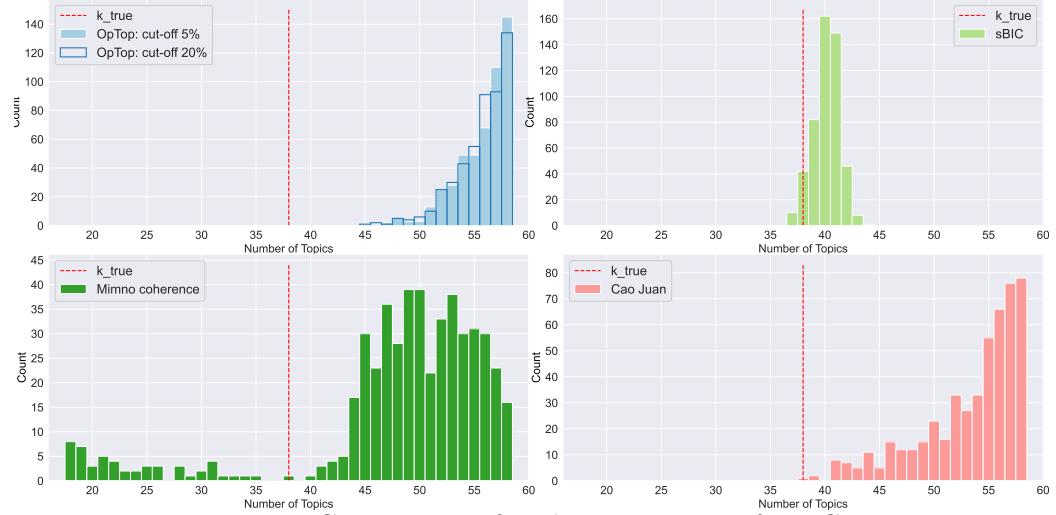

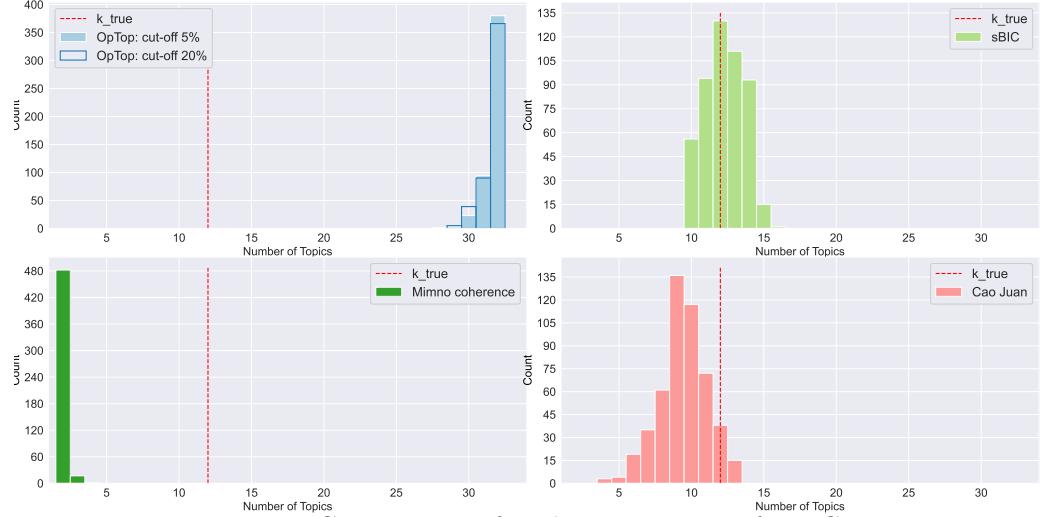Figure 2: Comparison of evaluation metrics for DGP1

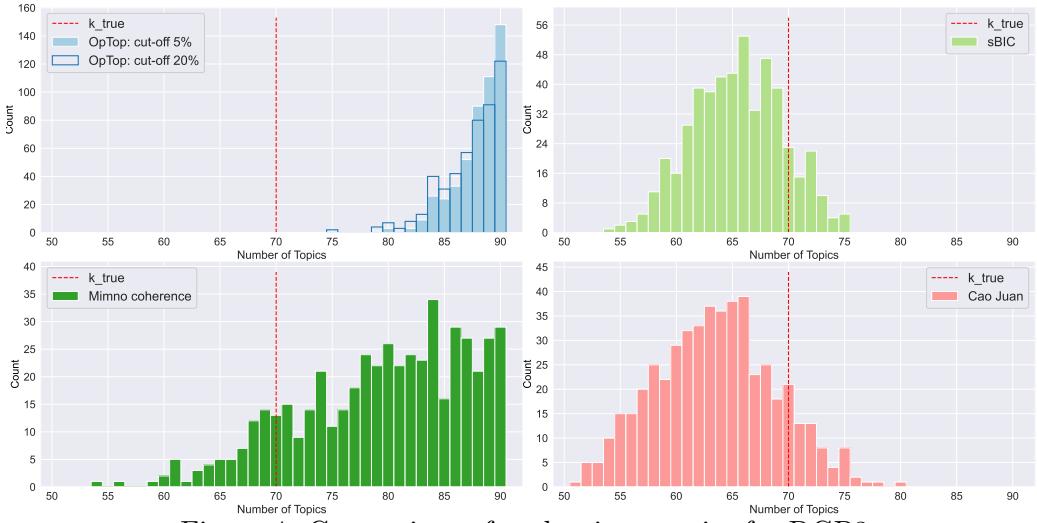Figure 3: Comparison of evaluation metrics for DGP2

Figure 4: Comparison of evaluation metrics for DGP3

Statistics from Table 1 further indicate that the mean and the median of the estimated number of topics for sBIC is in fact the closest to the true value for all DGPs. For DGP 2 the median of the estimates provided by sBIC is the actual number of themes. The performance of the metric differs for DGP 1 and DGP 3. In the first case, sBIC tends to select too many topics and in the second it chooses too few topics on average. The differences between the true and the estimated values are relatively small, however both types of estimation errors have their consequences. Overestimation means that some artificial topics will be generated, while underestimation implies that a number of relevant themes will be omitted. These issues are further discussed in Section 4.2 where the structure and content of the estimated topics is evaluated.

The performance of the OpTop statistic is rather poor as the procedure has a strong tendency to select too many topics for each DGP for both cut-off values of low-frequency words (5% and 20%). In each case, the mean and median values of the estimates are very close to the maximum of the range of candidates for the optimal number of topics. Such large overestimation errors mean that a substantial number of topics that do not belong would be estimated. Since, as noted by Mimno et al. (2011), there is a trade-off between obtaining many refined topics and meaningful themes, the quality of these additional topics found by the OpTop method might be expected to be rather low.

The working of the average cosine similarity (Cao_Juan) depends on the DGP. The mean/median number of topics selected for DGP 1 is too large as compared with the true number of topics, while the mean/median number of topics selected for DGPs 2 and 3 is too low as compared to the true number of topics. This outcome might depend on particular features of the DGPs (e.g. DGP 1 including a relatively small number of longer documents) which could be subject to further analyses. On the whole, the estimation errors are larger than for sBIC and smaller than in case of the OpTop procedure.

The unsystematic behaviour in terms of the tendency to over- or under-

11

estimate can be also seen for the average semantic coherence (Mimno). The mean/median number of topics selected for DGP 1 and DGP 3 are too large as compared with the true number of topics, while there is severe underestimation problem for DGP 2. The performance of this procedure seems to be quite unstable as in the case of DGPs 1 and 3 the estimates have the largest variance as compared to the remaining methods.

		DGP1 $(K_{true} = 38)$	DGP2 $(K_{true} = 12)$	DGP3 $(K_{true} = 70)$
sBIC	std	1.23	1.35	4.09
	mean	40.15	12.28	65.43
	median	40.00	12.00	66.00
	skewness	-0.24	0.00	-0.04
Cao_Juan	std	4.54	1.68	5.36
	mean	53.40	9.43	63.53
	median	55.00	9.00	64.00
	skewness	-1.16	-0.28	0.12
Mimno	std	9.23	0.20	7.55
	mean	47.88	2.04	79.50
	median	50.00	2.00	81.00
	skewness	-1.89	5.55	-0.61
OpTop 5%	std	2.30	0.59	2.06
	mean	55.81	31.70	88.03
	median	57.00	32.00	89.00
	skewness	-1.22	-2.17	-1.28
OpTop 20%	std	2.39	0.67	2.61
	mean	55.67	31.63	87.38
	median	56.00	32.00	88.00
	skewness	-1.37	-1.78	-1.30

Table 1: Evaluation of different criteria

4.2 Content and Structure of Topics

While the selected *number* of topics delivers first general insights on the performance of different criteria, this indicator does not contain information on the correspondence between the topics used to generate the text corpora and the topics obtained using the selected number of topics in the estimation procedure. Therefore, the structure and the *content* of topics should be also considered.[6] To this end, we propose to consider the problem as a classifica-

---

[6]In applications, sometimes the quality of topics is analyzed based on human judgment. For example, Morstatter and Liu (2018) present an approach based on existing measures of topic coherence and extending them by a measure of topic consensus by humans. Although this approach delivers some measure of interpretability by humans, the authors point out the need for automated and reproducible measures of topic quality.

tion task. This allows to compare the results obtained using all the different selection criteria quantitatively making use of well established performance metrics. We use precision and recall as commonly used performance metrics. In standard applications, these are defined as follows:

- **Recall** describes how many relevant items are retrieved.

- **Precision** indicates how many retrieved items are relevant.

In standard classification tasks, the length of predicted and actual labels is the same. In our case it might be different, as the number of topics selected by each of the considered evaluation metrics can deviate from the true number of topics as described in the previous subsection. Thus, we define the True Positive (TP) class as those topics that were correctly identified, i.e., true topics which find their match in the set of estimated topics for the number of topics indicated by the given selection criterion. Using this definition, precision and recall can be defined and calculated as follows:

$$\text{Recall} = \frac{|\text{TP}|}{K_{true}}, \tag{4.1}$$

where $|\text{TP}|$ denotes the cardinality of the set TP and K_{true} is the true number of topics in a particular DGP.

$$\text{Precision} = \frac{|\text{TP}|}{K_{metric}}, \tag{4.2}$$

where K_{metric} is the proposed number of topics according to the selection criterion considered.

As there might be a trade-off between recall and precision, the F1 measure is often used as a combined measure. F1 is calculated as follows:

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.3}$$

For computing these measures, estimated topics have to be matched with true topics from the DGP. This matching can be done using topic matching technique proposed by Bystrov et al. (2022), the so-called *best matching*. For each "true" topic, a match in the set of estimated topics is identified using the cosine similarity measure. If one of the "true" topics finds several matches, we only consider the matches with the highest cosine similarities. Obviously, a "best match" does not have to be a sensible match, i.e., close to the true topic. Therefore, we apply a threshold for the cosine similarity which has to be surpassed in order to consider a match as being a sensible match. This threshold is the same as used for the topic number reduction step for each DGP described in Subsection 3.2 (see Appendix A for further details).

Table 2 describes the distribution of precision and recall for each DGP and each evaluation metric.[7] As mentioned before, our application differs from standard classification problems as the number of true and estimated topics might differ. Hence, the interpretation of the results is slightly different. Here, a precision value of 1 means that all of the estimated topics are sensible matches to some of the true topics. However, it does not imply that all of the true topics are uncovered. Consequently, this measure might overestimate the performance of a metric if it tends to underestimate the true number of topics. For DGP 2, for example, the Mimno metric is described by an average precision value of 1, while the average recall value is 0.17. In the previous subsection, it was shown that the Mimno metric tends to underestimate the true number of topics for DGP 2. Thus, the high precision value only indicates that these few estimated topics are related to the true topics. sBIC, on the other hand, shows relatively high values for both recall and precision, 0.93 and 0.92 respectively indicating that mostly true topics and most of the true topics are found.

As for recall, a value of 1 means that all of the true topics are uncovered by the estimated topics. However, it does not imply that $K_{metric} = K_{true}$. Consequently, this measure might lead to overestimation of the performance of a metric if it tends to select too many topics. For DGP 1, for example, the Cao_Juan metric reveals an average recall value of 1, while the average precision value of 0.72 is substantially lower. Also in this example, sBIC performs well with average recall and precision values of 0.99 and 0.94, respectively.

To take account of the trade-off described above, it seems appropriate to consider both evaluation metrics simultaneously. This is done making use of the F1 score defined in equation 4.3 a the harmonic mean of recall and precision. The interpretation of F1 is straightforward: the higher the values the better the joint score for both recall and precision. The results indicate that sBIC outperforms the other evaluation metrics for DGP 1 and DGP 2. For DGP 3, according to the F1 score sBIC is found to perform similarly to the Cao_Juan criterion, while still exhibiting some advantages compared to the other criteria.

5   Conclusions and Outlook

Estimating LDA models requires making a number of decisions regarding parameter settings. This paper considered the problem of selecting the value of one of those essential parameters, viz. the number of topics discussed in the text corpus. The main aim was to analyze the properties of various model selection criteria with special focus on the recently proposed singular Bayesian information criterion. The performance of the methods was examined via Monte Carlo experiments using synthetic data generating processes based on empirical text corpora which differed with respect to the number and length

---

[7]As a robustness check we also calculate the described performance metrics using cosine similarities instead of the binary indicator match/no match. The procedure is described in Appendix C. The results do not differ qualitatively.

		Recall		Precision		F1	
data	metric	mean	std	mean	std	mean	std
DGP1	Cao_Juan	1.00	0.00	0.72	0.07	0.83	0.04
	Mimno	0.96	0.12	0.78	0.09	0.85	0.06
	OpTop 20%	1.00	0.00	0.68	0.03	0.81	0.02
	OpTop 5%	1.00	0.00	0.68	0.03	0.81	0.02
	sBIC	0.99	0.01	0.94	0.03	0.97	0.01
DGP2	Cao_Juan	0.78	0.13	1.00	0.02	0.87	0.09
	Mimno	0.17	0.02	1.00	0.00	0.29	0.02
	OpTop 20%	1.00	0.00	0.38	0.01	0.55	0.01
	OpTop 5%	1.00	0.00	0.38	0.01	0.55	0.01
	sBIC	0.93	0.06	0.92	0.07	0.92	0.04
DGP3	Cao_Juan	0.87	0.05	0.96	0.03	0.91	0.02
	Mimno	0.93	0.04	0.83	0.05	0.87	0.02
	OpTop 20%	0.98	0.01	0.79	0.03	0.87	0.02
	OpTop 5%	0.98	0.01	0.78	0.02	0.87	0.02
	sBIC	0.88	0.04	0.95	0.03	0.91	0.02

Table 2: Descriptive statistics of recall, precision, and F1 scores

of documents and the number of topics. The performance of different model selection procedures was evaluated by not only examining the accuracy of estimating the actual number of topics but also by analyzing the structure and contents of the estimated topics.

Simulation results showed that the singular Bayesian information criterion performed relatively well for all data generating processes considered in the experiments. It was the best method for estimating the number of topics as it was associated with the smallest estimation errors as compared to the competitors. In addition, it resulted in topics with good content and structure and performed in a relatively stable fashion for all data generating processes. Across the DGPs, the working of the method based on sBIC was worst for DGP 3 corresponding to a text corpus with a large number of short documents and a substantial number of topics. In this setting, sBIC exhibited a certain downward bias in the selected number of topics which might be taken into account in applied work. The reasons for this finding and possible adjustments to the method might be subject to further analyses.

The performance of the methods proposed by Cao et al. (2009) and Mimno et al. (2011) depended on the DGP. For each of these methods, the experiments revealed cases of systematic under- or overestimation of the true number of topics. The estimation errors were larger than those found for sBIC and had some negative consequences for the structure and content of the estimated topics. Dependence on the DGP implies that reliability and stability of these methods cannot be guaranteed in applied work unless further analyses will explain the relation between features of a DGP and the model selection results. Despite these drawbacks, the method of Cao et al.

(2009) was still overall the second best approach to LDA model selection in the experiments reported in this paper. It was found that the method could be particularly useful for modelling collections of many short texts related to a large range of topics.

The final set of conclusions relates to the OpTop criterion. It was shown that the method tends to select models with an excessively large number of topics. The estimation errors were very substantial and led to small precision and F1 metric values used for examining the content and structure of estimated topics. These results imply that using this criterion in applied work can result in obtaining some spurious topics, which do not correspond to the data generating process. It seems that poor estimation properties of the OpTop procedure could be improved by the introduction of an appropriate penalty for model complexity (which increases with the number of topics) into the test statistic formula. This adjustment constitutes a further direction of future research.

References

Adämmer, P. and Schüssler, R. A. (2020). Forecasting the equity premium: Mind the news!, *Review of Finance* **24**(6): 1313–1355.
URL: *https://academic.oup.com/rof/article/24/6/1313/5788550*

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research* **3**: 993–1022.

Bystrov, V., Naboka, V., Staszewska-Bystrova, A. and Winker, P. (2022). Cross-corpora comparisons of topics and topic trends, *Journal of Economics and Statistics* **242**(4): 433–469.
URL: *https://doi.org/10.1515/jbnst-2022-0024*

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

Dahlke, J., Bogner, K., Becker, M., Schlaile, M. P., Pyka, A. and Ebersberger, B. (2021). Crisis-driven innovation and fundamental human needs: A typological framework of rapid-response covid-19 innovations, *Technological Forecasting & Social Change* **169**.
URL: *https://www.sciencedirect.com/science/article/pii/S0040162521002316*

Drton, M. and Plummer, M. (2017). A Bayesian information criterion for singular models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(2): 323–380.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12187*

Ellingsen, J., Larsen, V. H. and Thorsrud, L. A. (2022). News media versus fred-md for macroeconomic forecasting, *Journal of Applied Econometrics* **37**(1): 63–81.

Ferrara, F. M., Masciandaro, D., Moschella, M. and Romelli, D. (2022). Political voice on monetary policy: Evidence from the parliamentary hearings of the european central bank, *European Journal of Political Economy* **74**: 102143.
URL: *https://www.sciencedirect.com/science/article/pii/S0176268021001178*

Hartmann, P. and Smets, F. (2018). The european central bank's monetary policy during its first 20 years, *Brooking Papers on Economic Activity* **Fall 2018**: 1–146.

Hayashi, N. (2021). The exact asymptotic form of Bayesian generalization error in latent Dirichlet allocation, *Neural Networks* **137**: 127–137.
URL: *https://www.sciencedirect.com/science/article/pii/S0893608021000320*

Hayashi, N. and Watanabe, S. (2020). Asymptotic Bayesian generalization error in latent Dirichlet allocation and stochastic matrix factorization, *SN Computer Science* **1**(69).

Kleinberg, B., van der Vegt, I. and Mozes, M. (2020). Measuring emotions in the COVID-19 real world worry dataset, *arXiv:2004.04225* .

Lewis, C. and Grossetti, F. (2022). A statistical approach for optimal topic model identification, *Journal of Machine Learning Research* **23**: 1–20.

Lin, A. Y.-T. and Katada, S. N. (2022). Striving for greatness: status aspirations, rhetorical entrapment, and domestic reforms, *Review of International Political Economy* **29, NO. 1**: 175–201.
URL: *https://www.tandfonline.com/doi/full/10.1080/09692290.2020.1801486*

Loureiro, S. M. C., Guerreiro, J. and Tussyadiah, I. (2021). Artificial intelligence in business: State of the art and future research agenda, *Journal of Business Research* **129**: 911–926.
URL: *https://www.sciencedirect.com/science/article/pii/S0148296320307451*

Lüdering, J. and Winker, P. (2016). Forward or backward looking? The economic discourse and the observed reality, *Journal of Economics and Statistics* **236**(4): 483–515.

Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. (2011). Optimizing semantic coherence in topic models, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 262–272.
URL: *https://aclanthology.org/D11-1024*

Morstatter, F. and Liu, H. (2018). In search of coherence and consensus: Measuring the interpretability of statistical topics, *Journal of Machine Learning Research* **18**(169): 1–32.
URL: *http://jmlr.org/papers/v18/17-069.html*

Polyzos, E. and Wang, F. (2022). Twitter and market efficiency in energy markets: Evidence using lda clustered topic extraction, *Energy Economics* **114**: 106264.
URL: *https://www.sciencedirect.com/science/article/pii/S0140988322004017*

Savin, I. and Teplyakov, N. (2022). Topics of the nationwide phone-ins with Vladimir Putin and their role for public support and Russian economy, *Information Processing & Management* **59**(5): 103043.
URL: *https://www.sciencedirect.com/science/article/pii/S0306457322001480*

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle, *Journal of Business & Economic Statistics* **38**: 393–409.

Tiba, S., Rijnsoever, F. J. v. and Hekkert, M. P. (2018). Firms with benefits: A systematic review of responsible entrepreneurship and corporate social responsibility literature, *Corporate Social Responsibility and Environmental Management* **26**(2): 265–284.
URL: *https://onlinelibrary.wiley.com/doi/full/10.1002/csr.1682*

Wang, F., Zhang, J. L., Li, Y., Deng, K. and Liu, J. S. (2021). Bayesian text classification and summarization via a class-specified topic model, *Journal of Machine Learning Research* **22**(89): 1–48.
  URL: *http://jmlr.org/papers/v22/18-332.html*

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press.

Appendices

Appendix A Topic Number Reduction

The goal of the topic number reduction step in preparing our DGPs for the Monte Carlo simulation was to use well separated topics allowing for a robust comparison of the topics estimated with the underlying DGPs. The process of topic number reduction comprises the following three steps:

1. Starting with the estimated LDA for a given corpus, for each topic the most similar other topic is identified using the standard matching proposed by Bystrov et al. (2022).

2. For deciding whether a pair of topics is "too similar", i.e., will be excluded before generating synthetic data within the Monte Carlo simulation, a threshold value has to be defined. This value is also obtained by a data driven approach. We calculate all pairwise cosine similarity scores for each DGP providing $\frac{K^2-K}{2}$ typical values. Sorting them in increasing order provides the distributions shown in Figure 5. Following the approach of the "elbow" criterion, we set percentile values defining the cut-off value for each DGP. These values are shown in the figure by the red horizontal line and correspond to the 95% percentile for DGP2 and to the 99% percentile for DGPs 1 and 3, respectively.

(a) DGP1 (b) DGP2 (c) DGP3

Figure 5: Distribution of the pairwise cosine similarity values.

3. All topics belonging to matched topic pairs above the cut-off value are considered as being too similar and, consequently, are removed from the model before starting the data generation within the Monte Carlo simulation. Figures 6, 7, and 8 show examples of pairs including redundant topics in each DGP, which are eliminated by this method.
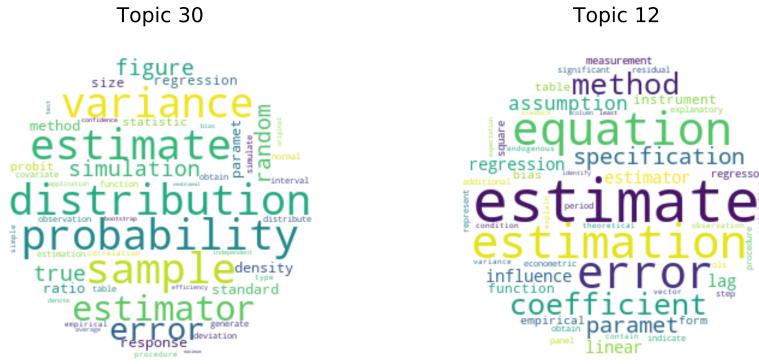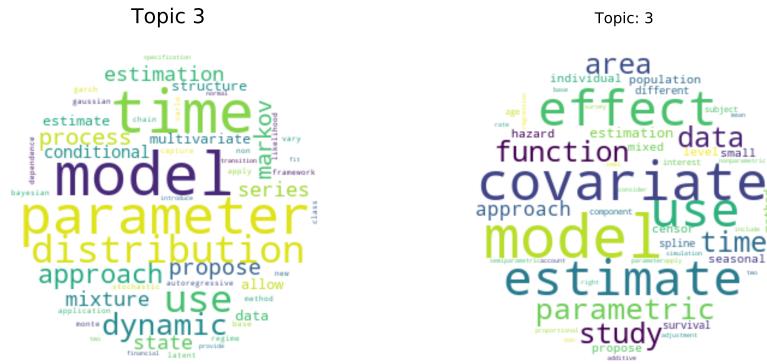
Figure 6: Similar topics in DGP 1
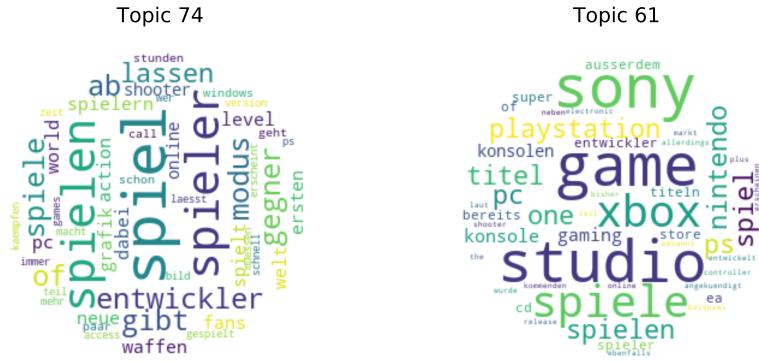

Figure 7: Similar topics in DGP 2

Figure 8: Similar topics in DGP 3

Appendix B   Recall and Precision

Figures 9, 10, and 11 exhibit the scatter plots of recall and precision values for each DGP separately. Thereby, each point corresponds to one of the simulated corpora. Consequently, there is a total of 300 points in each plot. However, the evaluation metrics considered may result in the same recall and precision scores for multiple corpora. Thus, some points may overlap.

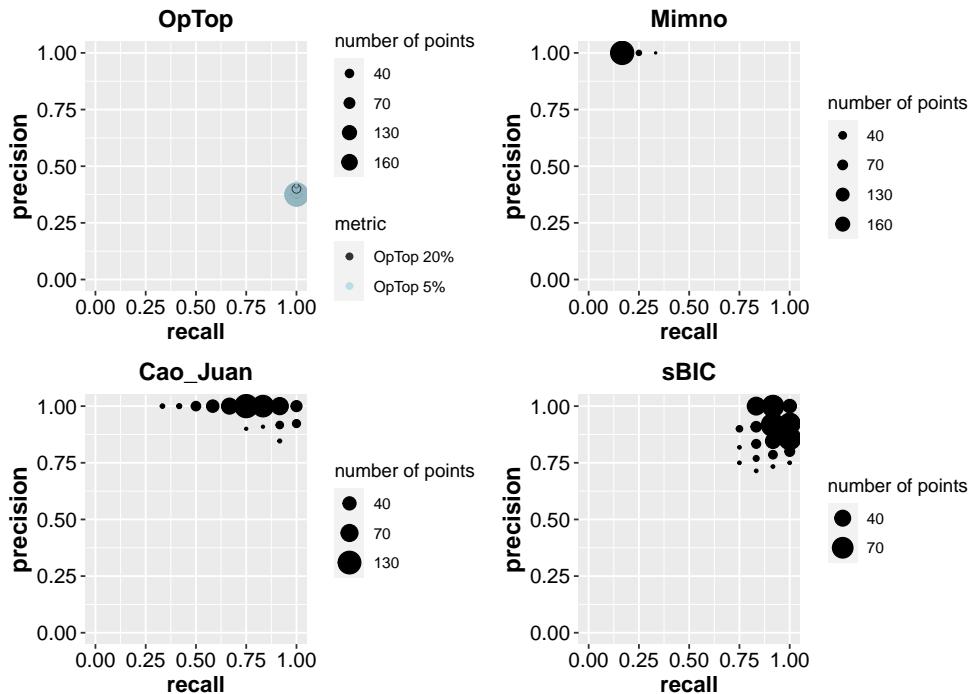Figure 9: Precision and recall for DGP1
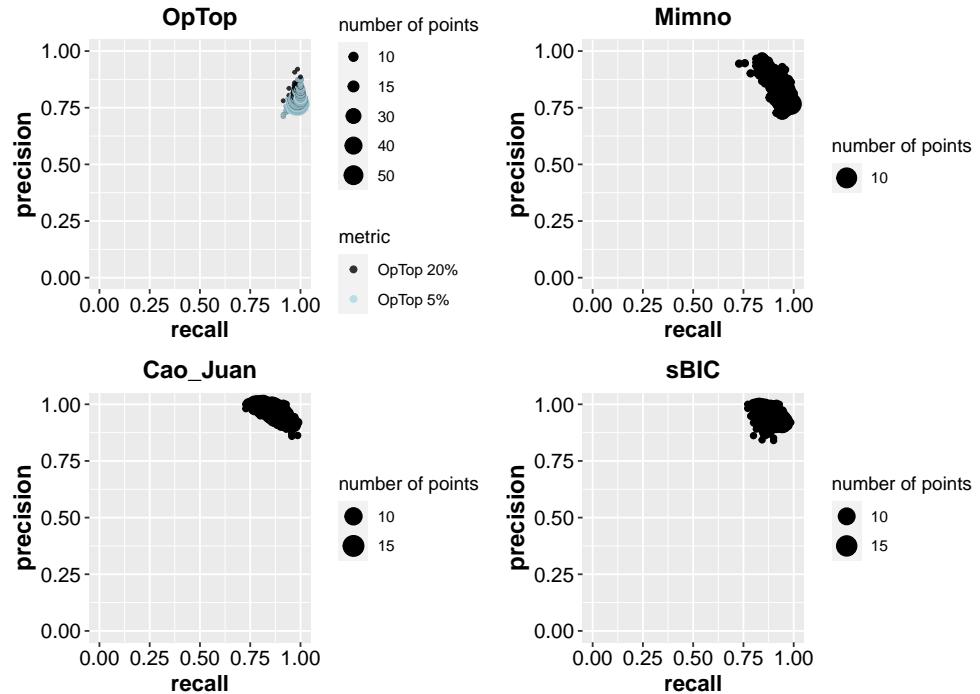

Figure 10: Precision and recall for DGP2

22

Figure 11: Precision and recall for DGP3

Appendix C   Weighted Recall & Precision

As a robustness analysis, we report results for alternative definitions of recall and precision. We identified a true positive (TP) for the measures in Section 4.2, when the similarity of matched topics was above a predefined threshold. Here, we use the actual cosine similarity scores instead, which would be close to 1 for good matches. Hence, recall and precision values are calculated as follows:

$$\text{Recall} = \frac{\sum_{i=1}^{n} \text{cosine\_similarity\_score}_i}{K_{true}}, \tag{C.1}$$

$$\text{Precision} = \frac{\sum_{i=1}^{n} \text{cosine\_similarity\_score}_i}{K_{metric}}, \tag{C.2}$$

where K_{true} is the true number of topics in a particular DGP. K_{metric} is the proposed number of topics for the evaluation metric considered. The numerator contains the sum of cosine similarity values of all the n identified matches. Therefore, recall presents the average cosine similarity value among the matches relative to the true number of topics. Precision presents the average cosine similarity value between the matches relative to the estimated number of topics.

Table 3 summarizes the recall, precision, and F1 score values for this alternative definitions of recall and precision. As expected, the values are smaller than the values shown in Table 2 for the original definitions, but the qualitative findings about the relative performance of the different criteria remain unchanged. According to the F1 scores, sBIC performs best for DGP 1 and DGP 2, while the average F1 scores are quite similar for all the considered metrics in DGP 3, still with a minor advantage for Cao_Juan and sBIC.

data	metric	Recall		Precision		F1	
		mean	std	mean	std	mean	std
DGP1	Cao_Juan	0.99	0.00	0.71	0.07	0.82	0.04
	Mimno	0.95	0.13	0.76	0.08	0.83	0.07
	OpTop 20%	0.98	0.00	0.67	0.03	0.80	0.02
	OpTop 5%	0.98	0.00	0.67	0.03	0.80	0.02
	sBIC	0.99	0.02	0.93	0.03	0.96	0.02
DGP2	Cao_Juan	0.76	0.14	0.96	0.03	0.84	0.10
	Mimno	0.11	0.02	0.66	0.02	0.19	0.02
	OpTop 20%	1.00	0.00	0.38	0.01	0.55	0.01
	OpTop 5%	1.00	0.00	0.38	0.01	0.55	0.01
	sBIC	0.92	0.07	0.91	0.06	0.91	0.05
DGP3	Cao_Juan	0.85	0.06	0.94	0.02	0.89	0.03
	Mimno	0.92	0.04	0.81	0.05	0.86	0.02
	OpTop 20%	0.98	0.02	0.78	0.03	0.87	0.02
	OpTop 5%	0.98	0.02	0.78	0.02	0.87	0.02
	sBIC	0.86	0.05	0.92	0.03	0.89	0.03

Table 3: Descriptive statistics of recall, precision, and F1 scores based on cosine similarity

While recall and precision values of our standard implementation are discrete leading to clustering of points in the scatter plots shown in Appendix B, the weighted recall and precision values reported in this section are continuous and each point is actually unique due to the differences in the cosine values, although these might be minor. Therefore, we do not use the type of plots from Appendix B taking into account the clustering, but standard scatter plots in Figures 12, 13, and 14.
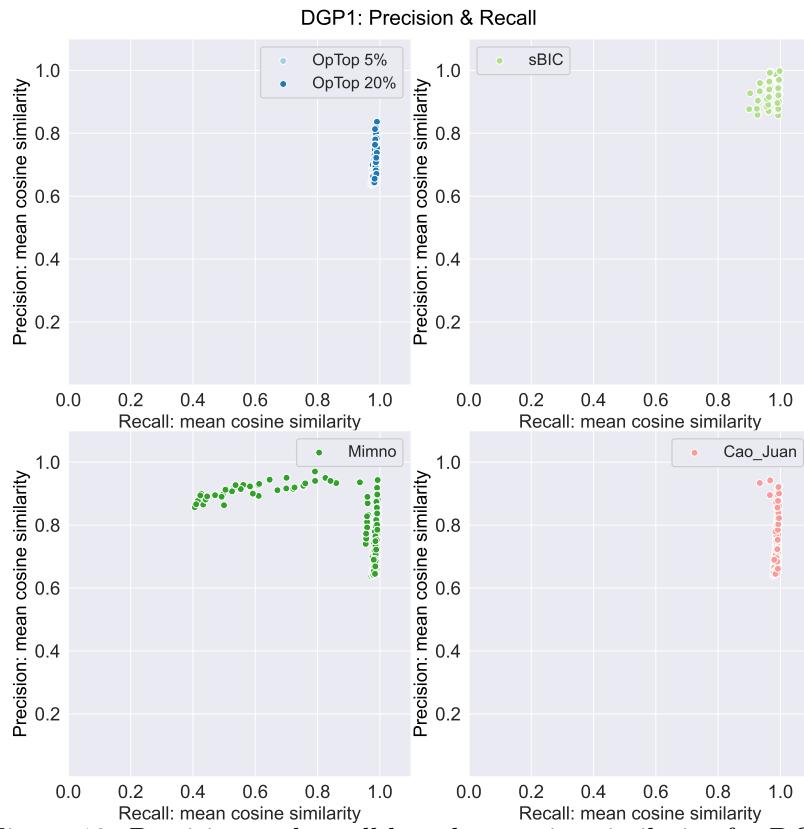
Figure 12: Precision and recall based on cosine similarity for DGP1
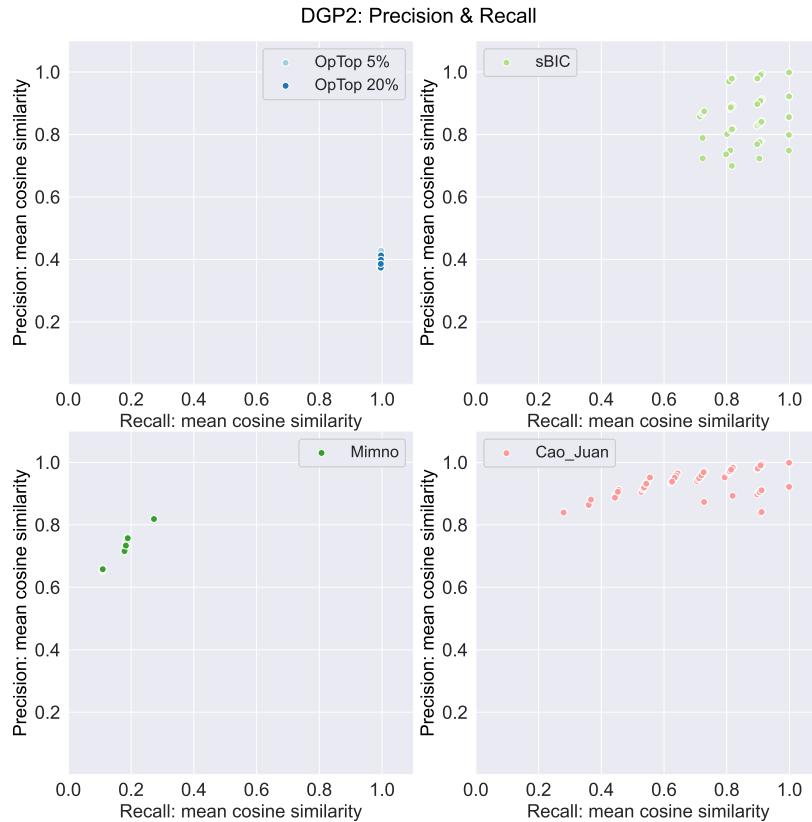

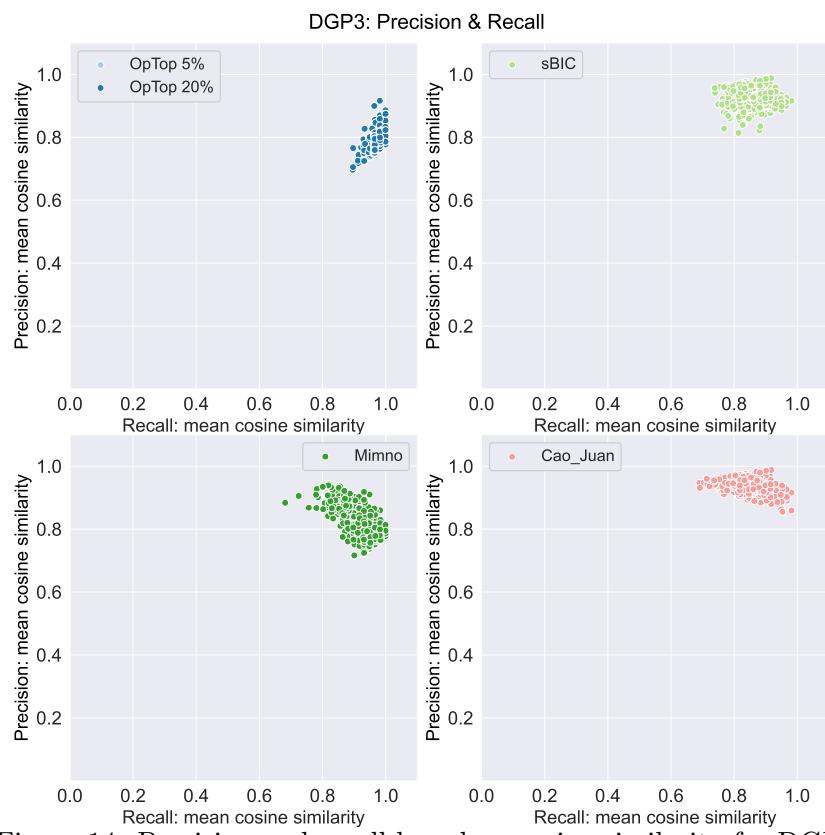
Figure 13: Precision and recall based on cosine similarity for DGP2

Figure 14: Precision and recall based on cosine similarity for DGP3