Journal of Machine Learning Research 25 (2024) 1-30 Submitted 2/23; Published 3/24

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

Victor Bystrov VICTOR.BYSTROV@UNI.LODZ.PL
Faculty of Economics and Sociology
University of Lodz
Rewolucji 1905r. 41, 90-214 Lodz, Poland

Viktoriia Naboka-Krell VIKTORIIA.NABOKA@WIRTSCHAFT.UNI-GIESSEN.DE
Department of Statistics and Econometrics
Justus Liebig University Giessen
Licher Strasse 64, 35394 Giessen, Germany

Anna Staszewska-Bystrova ANNA.BYSTROVA@UNI.LODZ.PL
Faculty of Economics and Sociology
University of Lodz
Rewolucji 1905r. 37/39, 90-214 Lodz, Poland

Peter Winker PETER.WINKER@WIRTSCHAFT.UNI-GIESSEN.DE
Department of Statistics and Econometrics
Justus Liebig University Giessen
Licher Strasse 64, 35394 Giessen, Germany

Editor: Matthew Hoffman

Abstract

Selecting the number of topics in Latent Dirichlet Allocation (LDA) models is considered to be a difficult task, for which various approaches have been proposed. In this paper 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 applied 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 considered data generation processes (DGPs) are revealed. Practical recommendations for LDA model selection in applications are derived.

Keywords: Topic models, text analysis, latent Dirichlet allocation, singular Bayesian information criterion, Monte Carlo simulation, text generation

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 et al. (2003). For some

©2024 Victor Bystrov and Viktoriia Naboka-Krell and Anna Staszewska-Bystrova and Peter Winker.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/23-0188.html.

recent applications making use of this method, see, e.g., Lüdering and Winker (2016), Thorsrud (2020), Ellingsen et al. (2022), and Savin and Teplyakov (2022).

LDA is a generative model 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 determined by a mixture of terms 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, various 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 et al., 2009), maximize the topic coherence (Mimno et al., 2011) or maximize the goodness-of-fit between the estimated and the actual document-term 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 et al. (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.[1] In this paper, comprehensive Monte Carlo (MC) simulations are proposed, which allow 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 under the assumption that the underlying DGPs follow an LDA process (LDA based text generation). Afterwards, to each of the generated corpora in each of the considered DGPs, LDA models with different numbers of topics are fitted. Then, we apply several alternative criteria to select the number of topics and evaluate the performance of criteria over many MC replications. To the best of our knowledge, no such systematic and comprehensive comparison analysis of the criteria used for selecting the number of topics in LDA models 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 proper Monte Carlo (MC) simulations to evaluate the proposed criterion as well as several alternatives commonly used in applications. Third, we evaluate the performance of studied criteria quantitatively and qualitatively, i.e., we consider whether the actual number of topics as well as the content and structure of the estimated topics are approximated well

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.

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

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-term frequency matrix $X = \{x_{ji}\}_{j,i=1}^{J,I}$, where x_{ji} is the frequency of term i encountered in document j and I is the number of different terms 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-term 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:

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

where

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

and β_{ki} is the frequency of term 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 et al., 2021; Tiba et al., 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 et al., 2021).

2.2 Topic Coherence

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

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

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

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

$\mathbf{i}^{(k)} = (i_1^{(k)}, \ldots, i_v^{(k)})$ is the list of the v most frequent terms in topic k, $f(i)$ is the document frequency of term 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 terms 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 smoothing parameter ϵ is included to avoid taking the logarithm of zero and its default value is given by e^{-12}. The number of the most frequent terms, v, is set to the default value of 20.

The average semantic coherence is often used for selecting the number of topics in applied topic mining, e.g., for the analysis of monetary policy speeches (Ferrara et al., 2022), stock market news (Adämmer and Schüssler, 2020), tweets concerning the energy market (Polyzos and Wang, 2022), or survey responses on the consequences of the Covid-19 pandemic (Kleinberg et al., 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-term frequencies. The frequency of term 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-term frequencies is usually sparse, Lewis and Grossetti (2022) suggest collapsing relatively rare terms in a single frequency bin. For document j, they order terms from the smallest to the largest estimated frequency, $(i_1^{(j)}, i_2^{(j)}, \ldots, i_I^{(j)})$ such that $\hat{x}_{ji_1^{(j)}}^{(K)} \leq \hat{x}_{ji_2^{(j)}}^{(K)} \leq \ldots \leq \hat{x}_{ji_I^{(j)}}^{(K)}$, and select a sub-vector of relatively rare terms $(i_1^{(j)}, i_2^{(j)}, \ldots, i_p^{(j)})$. The cumulative frequency of relatively rare terms 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 $\hat{x}_{ji_p^{(j)}}^{(K)}$ is the largest frequency such that $\sum_{i=i_1^{(j)}}^{i_p^{(j)}}\hat{x}_{ji}^{(K)} < x_{\text{cutoff}}$, and x_{cutoff} is a cumulative frequency cut-off value. Following Lewis and Grossetti (2022), we use $x_{\text{cutoff}} = 0.05$ as

a baseline cut-off value. (For a robustness check, we also consider a cut-off value of 0.20). The actual cumulative frequency of relatively rare terms in document j is

$$x_{j,\min} = \sum_{i \in (i_1^{(j)}, \ldots, i_p^{(j)})} x_{ji}.$$

The resulting goodness-of-fit statistic is

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

where $(i_{p+1}^{(j)}, \ldots, i_I^{(j)})$ is a sub-vector of relatively frequent terms 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). Unlike the 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 – the singular Bayesian information criterion – is a generalization of the Bayesian information criterion (BIC) that can be applied to singular statistical models (Drton and Plummer (2017)). The criterion was successfully used by Bystrov et al. (2022) for selecting parsimonious LDA models with coherent topics, however the properties of the criterion as applied to LDA modelling have not been studied in a simulation setup.

The standard BIC for an LDA model with K topics is of the form

$$\text{BIC}(K) = \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K) - \frac{d_K}{2} \log(N), \quad (2.2)$$

where $P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K)$ is the value of the likelihood function for corpus $\mathcal{D}$ given the estimated matrices of document-topic and topic-term probabilities ($\hat{\theta}$ and $\hat{\beta}$), $d_K = J(K-1)+(I-1)K$ is the number of estimated parameters (model dimension), and N is the total number of words in the corpus. The model dimension, d_K, is a linear function of the number of topics, K, and the term $\frac{d_K}{2} \log(N)$ in equation (2.2) is a penalty for increasing the number of parameters.

The general formula of the BIC was derived by Schwartz (1978) as a quadratic approximation for the logarithm of the marginal likelihood under the assumption of a regular (non-singular) model for which the Fisher information matrix is positive definite. For Latent Dirichlet Allocation (LDA) models the Fisher matrix is singular and the quadratic approximation of the log-marginal likelihood, which is used in the derivation of the standard BIC (2.2), is not possible.

The singular Bayesian information criterion (sBIC) can be derived using an approximation of the log-marginal likelihood described by Watanabe (2009). For an LDA model, this approximation can be written as

$$\log L(\mathcal{D}|K) \simeq \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K) - [\lambda_{Kr} \log(N) - (m_{Kr} - 1) \log\log(N)], \tag{2.3}$$

where λ_{Kr} is a rational number in the interval $[0, d_K/2]$, m_{Kr} is a natural number in the range $\{1, 2, \ldots, d_K\}$, and r is an intrinsic value of the true distribution, $r = \text{rank}(\theta \times \beta)$, that depends on the true number of topics (see Hayashi (2021)). The term $[\lambda_{Kr} \log(N) - (m_{Kr} - 1) \log\log(N)]$ is an approximation of the model complexity in LDA, which is determined by the number of non-redundant parameters in matrices of document-topic and topic-term probabilities. It is smaller than the penalty in the standard BIC and, moreover, as a sub-linear function of the number of topics, K, the term $[\lambda_{Kr} \log(N) - (m_{Kr} - 1) \log\log(N)]$ grows slower than the penalty in the standard BIC (see Watanabe (2009) and Hayashi (2021)). Therefore, a criterion based on the approximation (2.3) selects an LDA model with more topics than the standard BIC (2.2).

The coefficients λ_{Kr} and m_{Kr} cannot be computed directly, because they depend on the true number of topics. This problem can be overcome by applying model averaging as proposed by Drton and Plummer (2017). In this approach, a feasible singular Bayesian information criterion for an LDA model with K topics is defined as an approximation of the log-marginal likelihood obtained by the averaging of sub-models (models with smaller or equal number of topics). The feasible sBIC satisfies the equation

$$\text{sBIC}(K) = \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K) - \log \left[\sum_{k \leq K} \omega_{Kk} N^{\lambda_{Kk}} (\log N)^{-(m_{Kk}-)} \right],$$

where the penalty term is the logarithm of the weighted average of $\left[N^{\lambda_{Kk}} (\log N)^{-(m_{Kk}-1)}\right]$ with coefficients λ_{Kk} and m_{Kk} depending on the number of topics in sub-models, $k \leq K$, and weights ω_{Kk} depending on the data. The computation of the feasible sBIC involves calculating λ_{Kk} and m_{Kk} for all $k \leq K$ as well as solving a system of quadratic equations. Therefore, full details of computing the feasible sBIC for an LDA model are provided in Appendix A.

3. Monte Carlo Simulations

Despite the broad usage of some metrics described in the previous section, there is not yet consensus on which metric performs best, when it comes to selecting the number of topics. Given that the data generating process (DGP) is unknown in applications, the relative performance of the metrics can only be assessed on the basis of a subjective analysis of the estimated topics. To account for this issue, we carry out a Monte Carlo (MC) simulation study, for which the data are generated by well-defined DGPs with known numbers of topics.[2] This allows us to compare the performance of alternative metrics with regard to

2. The idea of using Monte Carlo simulations for obtaining well-defined text corpora has been applied recently by Wang et al. (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.

the number of topics identified as well as to evaluate whether certain characteristics of the corpora such as number or length of documents might affect the relative performance. Furthermore, in the MC simulation study, we not only compare the selected number of topics to the number of topics in the DGP, but we also evaluate whether the estimated topics closely match the topics in the DGP.

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.

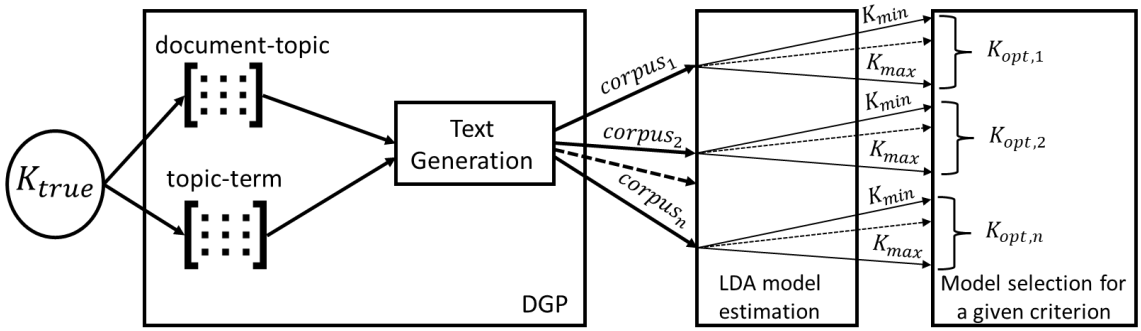


Figure 1: Generic procedure for Monte Carlo simulations with a given selection criterion

As described in Section 2, redLatent Dirichlet Allocation (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 can be described by two matrices, the first containing the probabilities of occurrence of each term in each topic (topic-term distribution), and the second providing the probabilities of each topic occurring in a single document (document-topic distribution). These matrices are used to generate text corpora in the MC simulations.

In the first step of the procedure, an LDA model is estimated using a real document corpus with a number of topics that was used in previous analysis of the selected corpus. In order to make certain that only distinct topics will be used for the MC generation of text corpora in the following step, topics exhibiting a cosine similarity with other topics larger than a selected cut-off value (95% or 99% percentile) are dropped and the document-topic matrix is re-scaled to ensure that topic weights add up to one (see Appendix B for more details). The data generating process based on distinct topics is intended to approximate a feature of the generative LDA model described by Blei et al. (2003) where

topics are independently drawn from a Dirichlet distribution. Dealing with well-separated topics stabilizes the performance of the estimation methods developed for LDA.

In the second step, text corpora are generated using the estimated document-topic and topic-term distributions. 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, a topic is randomly selected based on the known document-topic distribution and then a term is drawn from the vocabulary using the known topic-term distribution.

As mentioned above, Algorithm 1 does not exactly reproduce the generative procedure described in Blei et al. (2003) where rows of document-topic and topic-term frequency matrices are drawn from Dirichlet distributions. In applications, hyper-parameters of these distributions are not often estimated, and using exchangeable Dirichlet distributions in the generating process could result in document-topic and topic-term frequency matrices that structurally differ from frequency matrices estimated for actual text corpora. Therefore, we use a synthetic approach that, on the one hand, approximates the essential features of the generative LDA model and, on the other hand, replicates properties of frequency matrices estimated for real data.

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 term from the topic-term distribution
6: Append the selected term to the current document
7: **end for**
8: Append the generated document to the corpus.
9: **end for**

Algorithm 1 is implemented in each DGP with 1 000 Monte Carlo replications, i.e., 1 000 corpora containing the same number of documents of same length as the original corpus.

In the third step of the procedure, we estimate LDA models with the number of topics ranging from $\max\{2; K_{\text{true}} - 20\}$ to $K_{\text{true}} + 20$, where K_{true} is the number of topics in the data generating process. The maximum length of the range of admitted values for the number of topics is equal to 41 with the true number of topics, K_{true}, located in the center of the range if $K_{\text{true}} > 20$. 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 the estimated models.

In the final step, we compare the number of topics selected by different criteria using descriptive statistics such as standard deviation, mean, median, and skewness (see Subsection 4.1). 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 provide information about the extent to which the content of topics used for generating

texts is revealed in the estimated LDA with the number of topics selected by the different criteria.

3.2 Data Generating Processes

The three data generating processes (DGPs) used for the Monte Carlo (MC) simulations are related to three actual text 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 terms. 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 ($K_{\text{true}} = 38$).

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 vocabulary is composed of 1,796 terms. The focused nature of conference abstracts suggests a limited number of topics. The initial number of topics selected for this corpus was equal to 20. This number was reduced to 12 ($K_{\text{true}} = 12$) after removing the topics that were too similar.

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 terms 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. After removing the most similar topics, the final number of topics used in DGP 3 was equal to 70 ($K_{\text{true}} = 70$). In the analysis we used only the most recent 50,000 documents from this corpus because using the whole dataset would increase the computational costs of MC simulations beyond the available capacities.

At this point, we would like to emphasize that in the described experiments, texts are generated using an LDA model with distinct topics. It means that each generated text is a "bag-of-words", where semantic and syntactic relationships between words, observed in actual texts, are neglected. However, it allows us to create a controlled setting for text generation as well as for evaluating model selection criteria. The results of applying the considered criteria to actual corpora, which do not emerge from the generative LDA model, may therefore differ from those presented in this study. Nevertheless, the results of the described experiments provide insights into the usability of the model selection criteria in settings when LDA constitutes a reasonable approximation to the actual DGP.

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/`). The number of iterations was set to a relatively small value of 1 000 due to computational constraints. Most other parameters of the package were used at the default values. The point estimates of document-topic and topic-term frequency matrices are computed as in Griffiths and Steyvers (2004).

For DGP 1, the numbers of topics in the estimated models ranged from 18 to 58; for DGP, the number of topics ranged from 2 to 32; and for DGP 3 - from 50 to 90.

The average topic similarity (Cao_Juan) and the average semantic coherence (Mimno) criteria were computed using the Python package "tmtoolkit" (`https://pypi.org/project/tmtoolkit/`). The Python implementations of the singular Bayesian information criterion (sBIC) and the goodness-of-fit statistic (OpTop) model selection criteria were written by the authors.

For high-precision computations in the implementation of sBIC we used the Python module "decimal" and wrote a procedure that augments precision if it is necessary. The outer limits allowable for exponents of floating-point numbers have to be sufficiently large in order to avoid exponent underflow and overflow in the computation of sBIC which is a solution of a recursive system of quadratic equations parameterized by likelihood values in sub-models (see Appendix A). Compared to the estimation time of LDA models, the additional time needed for high-precision computations in the sBIC algorithm is not substantial.

Computations were performed using the high-performance-computing-cluster at Justus Liebig University Giessen (justHPC) (`https://www.hkhlr.de/de/cluster/justhpc-giessen`).[3]

4. Results

This section summarizes the results of the Monte Carlo simulations. It is divided into two subsections. Subsection 4.1 presents and discusses the results of estimating the optimal number of topics and subsection 4.2 evaluates the structure and contents of the estimated topics.

4.1 Number of Topics

The first set of results concerns the estimation of the number of topics, K. Figures 2-4 present histograms of the numbers of topics selected by different criteria for the three considered data generating processes (DGPs). In each of the histograms, the red vertical line depicts the true number of topics, K_{true}, used for generating the corpora. The shape and location of histograms shown in Figures 2-4 suggest that the sBIC is clearly the best

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

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.

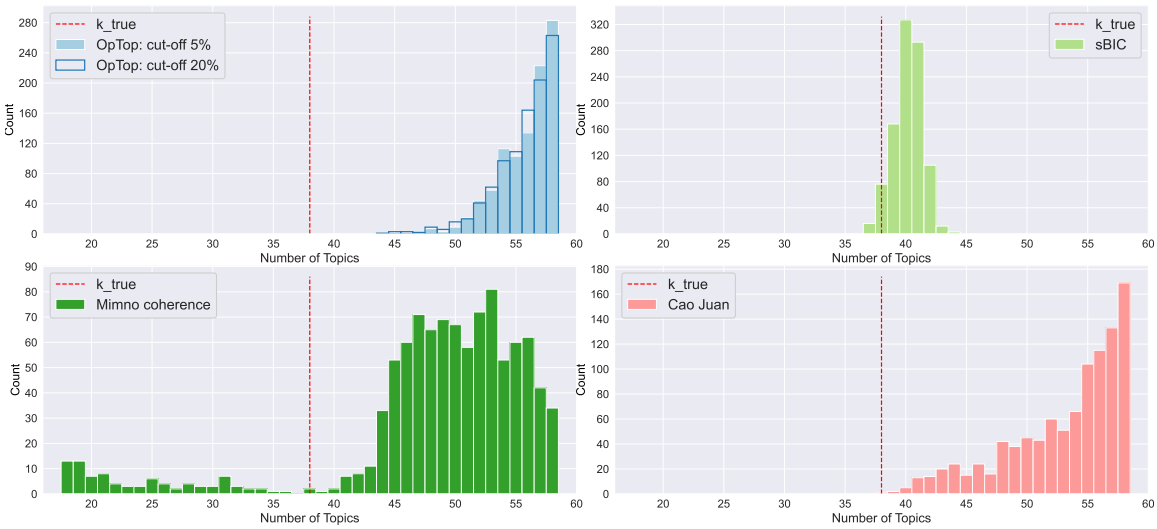


Figure 2: Comparison of evaluation metrics for DGP1

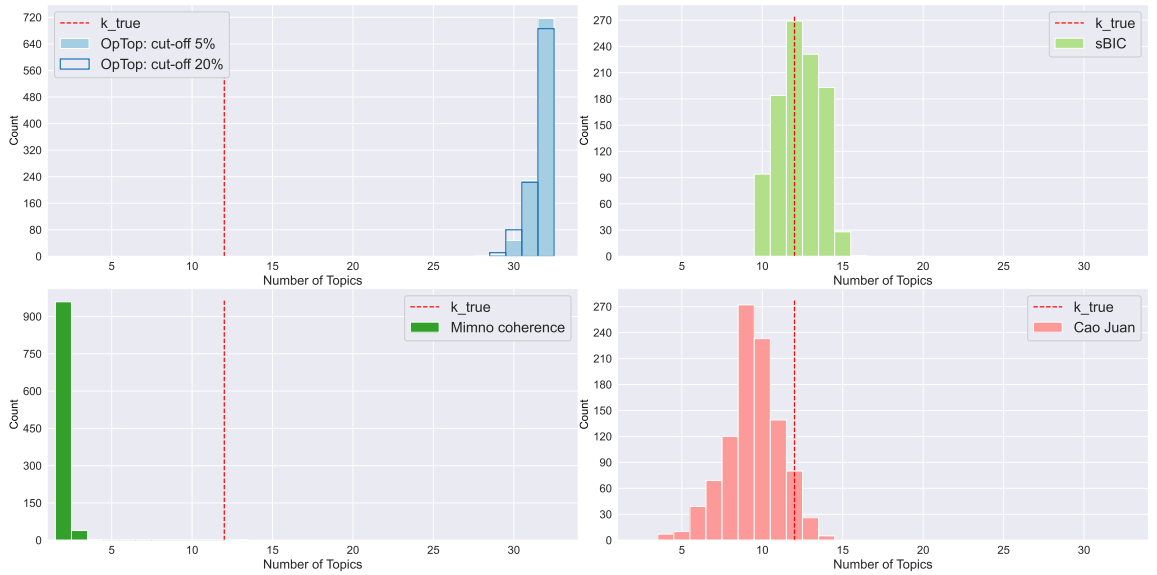


Figure 3: Comparison of evaluation metrics for DGP2

Table 1 provides descriptive statistics computed for the number of topics selected by each criterion. The results in Table 1 demonstrate that the mean and the median of the number of topics estimated by the sBIC is the closest to the true value for all DGPs. For DGP 2 the median of the estimates provided by the sBIC is the actual number of topics. The performance of the criterion differs for DGP 1 and DGP 3. In the first case, the

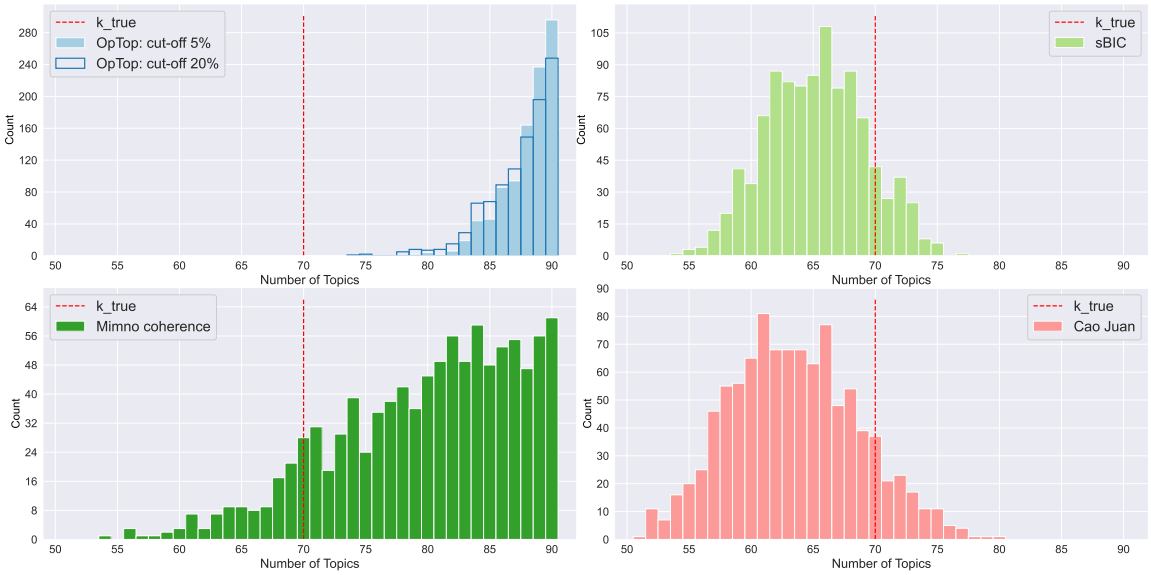


Figure 4: Comparison of evaluation metrics for DGP3

sBIC tends to select too many topics and in the second case, on average, it selects too few topics. The differences between the true and the estimated values are relatively small, but both types of estimation errors have their consequences. Overestimation of the number of topics means that some spurious topics will be generated, while underestimation implies that relevant topics 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 (goodness-of-fit) is rather poor as it has a strong tendency to select too many topics for each DGP. This result is robust with respect to the choice of the cut-off value for low-frequency terms (5% or 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 irrelevant topics would be estimated. Since, as noted by Mimno et al. (2011), there is a trade-off between obtaining many refined and meaningful topics, the quality of these additional topics found by the OpTop method might be expected to be rather low.

The performance of the average cosine similarity (Cao_Juan) varies depending on the DGP. The mean/median number of topics selected for DGP 1 is too large as compared to 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 the sBIC and smaller than in case of the OpTop criterion.

The unsystematic behaviour in terms of the tendency to over- or underestimate can be also seen for the average semantic coherence (Mimno). The mean/median number of topics selected for DGP 1 and DGP 3 is too large as compared to 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 the estimates have the largest variance compared to the remaining methods for DGP 1 and DGP 3.

		DGP1 ($K_{\text{true}} = 38$)	DGP2 ($K_{\text{true}} = 12$)	DGP3 ($K_{\text{true}} = 70$)
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

Table 2 reports the percentages of the number of topics, estimated by each criterion, falling within symmetric intervals centered at the true number of topics, $[K_{\text{true}}-k, K_{\text{true}}+k]$, $k \in \{0, 1, 2, 3, 4, 5\}$. The sBIC clearly outperforms other criteria for all DGPs. This result holds for all considered intervals. For example, in 88% of the cases sBIC delivers a topic number between 35 and 41 for DGP 1 ($K_{\text{true}} = 38$). In nearly 60% of cases sBIc proposes a number of topics between 65 and 75 for DGP 3 ($K_{\text{true}} = 70$) as opposed to 40% by Cao_Juan and 23% by Mimno.

4.2 Structure and Content 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

		$K_{\text{true}} - k \leq K_{\text{metric}} \leq K_{\text{true}} + k$					
DGP	Metric	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
DGP 1	Cao_Juan	0.1	0.3	0.8	2.1	3.5	5.5
	Mimno	0.2	0.3	0.6	1.4	2.4	3.7
	OpTop 20%	0.0	0.0	0.0	0.0	0.0	0.0
	OpTop 5%	0.0	0.0	0.0	0.0	0.0	0.0
	sBIC	7.6	26.0	58.7	88.0	98.5	99.7
DGP 2	Cao_Juan	8.0	24.5	48.3	75.5	87.5	94.4
	Mimno	0.0	0.1	0.1	0.1	0.1	0.1
	OpTop 20%	0.0	0.0	0.0	0.0	0.0	0.0
	OpTop 5%	0.0	0.0	0.0	0.0	0.0	0.0
	sBIC	26.9	68.4	97.1	99.9	100.0	100.0
DGP 3	Cao_Juan	3.7	9.7	17.4	23.9	32.7	40.1
	Mimno	2.8	8.0	11.6	15.4	20.1	23.4
	OpTop 20%	0.0	0.0	0.0	0.0	0.1	0.3
	OpTop 5%	0.0	0.0	0.0	0.0	0.0	0.0
	sBIC	4.2	13.4	25.8	36.2	47.8	56.9

Table 2: Percentages of the estimated number of topics, K_{metric}, falling within intervals around the true number of topics, K_{true}

topics should be also evaluated.[4] To this end, we propose to consider the problem as a classification task. This allows us to compare the results obtained using all the different selection criteria quantitatively making the use of well established performance measures, precision and recall. 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 criteria 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:

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

4. 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.

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

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

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 data generating process (DGP). This matching can be done using the topic matching technique proposed by Bystrov et al. (2022), the so-called *best matching*. Thereby, each topic is represented as a probability vector over the vocabulary. For each "true" topic, a match in the set of estimated topics is identified using the cosine similarity measure. Cosine similarity is often used in natural language processing to measure similarity between high-dimensional text representations. For each topic of the "true" topic set, cosine similarities to all the topics from the estimated topic set are calculated. Initially, a topic pair with the highest cosine similarity value is considered a "best match". 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 B for further details). If one of the "true" topics finds several sensible matches, we only consider the matches with the highest cosine similarities. The number of identified matches corresponds to "true positives", i.e. the number of correctly identified topics. This number of reproduced topics is then divided either by the true number of topic K_{true} (recall) or the estimated number of topics K_{metric} (precision).

Table 3 describes the distribution of precision and recall for each DGP and each evaluation metric.[5] 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.

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 example, in case of DGP 2, the average precision of topics selected by the Mimno criterion (average semantic coherence) is equal to 1, while the average recall is equal to 0.17. In the previous subsection, it was shown

5. 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 D. 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 3: Descriptive statistics of recall, precision, and F1 scores

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. On the other hand, for the sBIC, which performs very well in case of DGP 2, there are relatively high values of both recall and precision (0.93 and 0.92, respectively) indicating that mostly true topics and most of the true topics are recovered.

A recall value of 1 means that all of the true topics are uncovered by the estimated topics. However, it does not imply that $K_{\text{metric}} = K_{\text{true}}$. 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 (average cosine similarity) 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 combine precision and recall measures. This is done by using the F1 score which is defined in equation 4.3 a the harmonic mean of precision and recall. The interpretation of F1 is straightforward: the higher the values the better the joint score of precision and recall. The results indicate that the sBIC outperforms the other evaluation metrics for DGP 1 and DGP 2. For DGP 3, according to the F1 score the 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 Latent Dirichlet Allocation (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 performance 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 (DGPs) based on empirical text corpora which differed with respect to the number and length of documents and the number of topics. This text generation process was based on the assumption that the considered DGPs actually follow an LDA process (or could be approximated by an LDA process). The generalizability of the results is therefore limited. 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 (sBIC) 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 performance of the 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) (the average cosine similarity) and Mimno et al. (2011) (the average semantic coherence) 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 the 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 (the goodness-of-fit statistic). 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 direction of future research.

Acknowledgments

Financial support from the German Research Foundation (DFG) (WI 2024/8-1) and the National Science Centre (NCN) (Beethoven Classic 3: UMO-2018/31/G/HS4/00869) for the project TEXTMOD is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

Appendix A. Computation of the singular Bayesian Information Criterion (sBIC)

The marginal likelihood of a corpus $\mathcal{D}$ composed of J documents with a vocabulary including I terms given an LDA model with K topics is defined as

$$L(\mathcal{D}|K) = \int_{\theta,\beta} P(\mathcal{D}|\theta,\beta,K)dP(\theta,\beta|K), \tag{A.1}$$

where $P(\mathcal{D}|\theta,\beta,K)$ is the value of the likelihood function for the corpus $\mathcal{D}$ given $(J \times K)$ matrix of document-topic probabilities θ and $(K \times I)$ matrix of topic-term probabilities β, and $P(\theta,\beta|K)$ is a prior distribution of matrices θ and β in the LDA model with K topics.

An approximation of the marginal likelihood (A.1), based on the averaging of sub-models with number of topics $k = K_{\min},\ldots,K$, is defined as (see Drton and Plummer (2017))

$$L'(\mathcal{D}|K) = \frac{\sum_{k=K_{\min}}^{K} L'_{Kk} L(\mathcal{D}|k)P(k)}{\sum_{k=K_{\min}}^{K} L(\mathcal{D}|k)P(k)}, \tag{A.2}$$

where $P(k)$ is a prior for a model with k topics (assumed to be a known positive constant). The term L'_{Kk} is

$$L'_{Kk} = P(\mathcal{D}|\hat{\theta},\hat{\beta},K)N^{-\lambda_{Kk}}(\log N)^{m_{Kk}-1}, \tag{A.3}$$

where $P(\mathcal{D}|\hat{\theta},\hat{\beta},K)$ is the value of the likelihood function given the estimated parameter matrices $\hat{\theta}$ and $\hat{\beta}$ in the LDA model with K topics, and coefficients λ_{Kk}, m_{Kk} are computed for $k = K_{\min},\ldots,K$ using formulas from Hayashi (2021) where the rank of the matrix product $\theta \times \beta$ in true distribution, r, is replaced by the number of topics in a sub-model, k:

1. If $J + k \leq I + K$, $I + k \leq J + K$, $K + k \leq I + J$ and
   (a) if $I + J + K + k - 1$ is odd, then
   $\lambda_{Kk} = \frac{1}{8}\left\{2(K+k)(I+J) - (I-J)^2 - (K+k)^2\right\} - \frac{1}{2}J$ and $m_{Kk} = 1$
   (b) if $I + J + K + k - 1$ is even, then
   $\lambda_{Kk} = \frac{1}{8}\left\{2(K+k)(I+J) - (I-J)^2 - (K+k)^2 + 1\right\} - \frac{1}{2}J$ and $m_{Kk} = 2$
2. Else if $I + K < J + k$, then $\lambda_{Kk} = \frac{1}{2}\{IK + Jk - Kk - J\}$, $m_{Kk} = 1$
3. Else if $J + K < I + k$, then $\lambda_{Kk} = \frac{1}{2}\{JK + Ik - Kk - J\}$, $m_{Kk} = 1$

4. Else (i.e. $I + J < K + k$), then $\lambda_{Kk} = \frac{1}{2}(IJ - J)$, $m_{Kk} = 1$.

In equation (A.2) the approximation of the marginal likelihood, $L'(\mathcal{D}|K)$, is expressed as a function of the actual marginal likelihoods $L(\mathcal{D}|k)$, $k = K_{\min}, \ldots, K$. Drton and Plummer (2017) resolve this problem by replacing the unknown marginal likelihoods $L(\mathcal{D}|k)$ on the right-hand side of (A.2) by their approximations, $L'(\mathcal{D}|k)$, and considering a system of equations

$$L'(\mathcal{D}|K) = \sum_{k=K_{\min}}^{K} \frac{L'(\mathcal{D}|k)P(k)}{\sum_{k=K_{\min}}^{K} L'(\mathcal{D}|k)P(k)} L'_{Kk}, \qquad K = K_{\min}, \ldots, K_{\max}, \tag{A.4}$$

where L'_{Kk} and $P(k)$ are known constants and $L'(\mathcal{D}|K)$ are unknowns to be found. Then the singular Bayesian information criterion for a model with K topics is defined as

$$\text{sBIC}(K) = \log L'(\mathcal{D}|K), \tag{A.5}$$

where $L'(\mathcal{D}|K)$ is the unique solution of the transformed equation system (assuming that $P(K) > 0$ for $K = K_{\min}, \ldots, K_{\max}$)

$$\sum_{k=K_{\min}}^{K} [L'(\mathcal{D}|K) - L'_{Kk}]L'(\mathcal{D}|k) = 0,\ K = K_{\min}, \ldots, K_{\max} \tag{A.6}$$

that can be found inductively with $L'(\mathcal{D}|K_{\min}) = L'_{K_{\min}K_{\min}} > 0$ for the minimal model. Proceeding by induction, if $L'(\mathcal{D}|k)$ have been computed for all $k = K_{\min}, \ldots, (K-1)$, then $L'(\mathcal{D}|K)$ is the unique positive solution of quadratic equation

$$L'(\mathcal{D}|K)^2 + b_K L'(\mathcal{D}|K) - c_K = 0, \tag{A.7}$$

with

$$b_K = -L'_{KK} + \sum_{k=K_{\min}}^{K-1} L'(\mathcal{D}|k)\frac{P(k)}{P(K)} \text{ and } c_K = \sum_{k=K_{\min}}^{K-1} L'_{Kk}L'(\mathcal{D}|k)\frac{P(k)}{P(K)}.$$

Since $c_K > 0$ by induction, the quadratic equation (A.7) has the unique positive solution

$$L'(\mathcal{D}|K) = \frac{1}{2}\left(-b_K + \sqrt{b_K^2 + 4c_K}\right)$$

for $K = K_{\min} + 1, \ldots, K_{\max}$. Given formulas (A.3), (A.4) and (A.5), sBIC should satisfy

$$\text{sBIC}(K) = \log P(\mathcal{D}|\hat{\theta}, \hat{\beta}, K) - \log\left[\sum_{k=K_{\min}}^{K} \omega_{Kk} N^{\lambda_{Kk}} (\log N)^{-(m_{Kk}-1)}\right]$$

where weights ω_{Kk} are defined as

$$\omega_{Kk} = \frac{L'(\mathcal{D}|k)P(k)}{\sum_{k=K_{\min}}^{K} L'(\mathcal{D}|k)P(k)}, \; k = K_{\min}, \ldots, K.$$

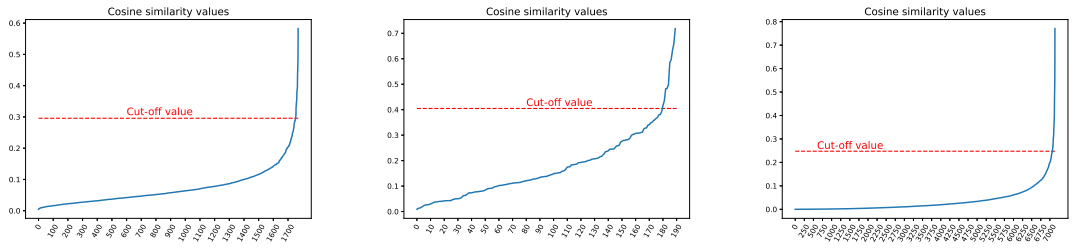
Because the coefficient λ_{Kk} is less than half the model dimension, $\frac{1}{2}[J(K-1)+(I-1)K]$, for every $k = K_{\min}, \ldots, K$, the penalty in the singular BIC is less than in the standard BIC for an LDA model with the same number of topics. Moreover, the penalty for increasing the number of topics in the singular BIC grows slower as compared to the standard BIC.

Appendix B. Topic Number Reduction

The goal of the topic number reduction step in preparing our DGPs for the Monte Carlo simulations 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.

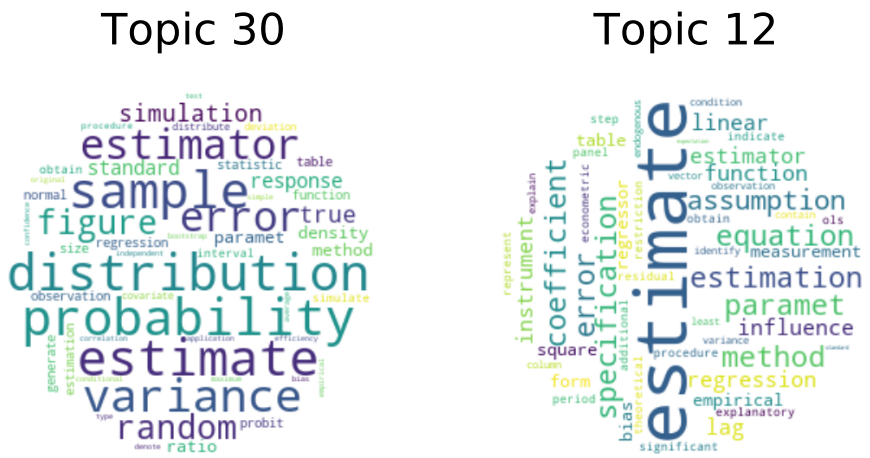


Figure 6: Similar topics in DGP 1

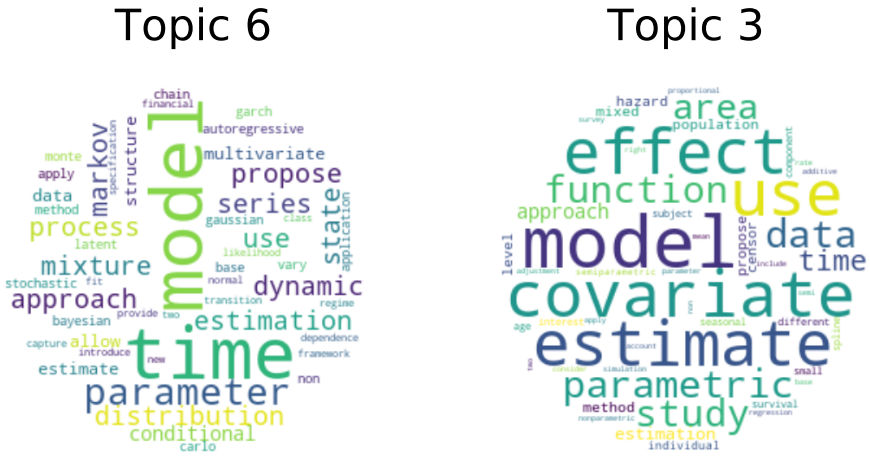


Figure 7: Similar topics in DGP 2

Appendix C. 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.

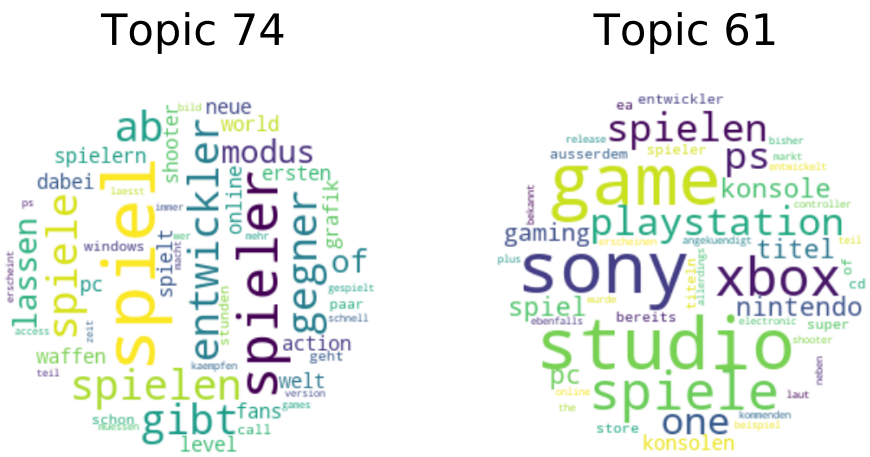


Figure 8: Similar topics in DGP 3

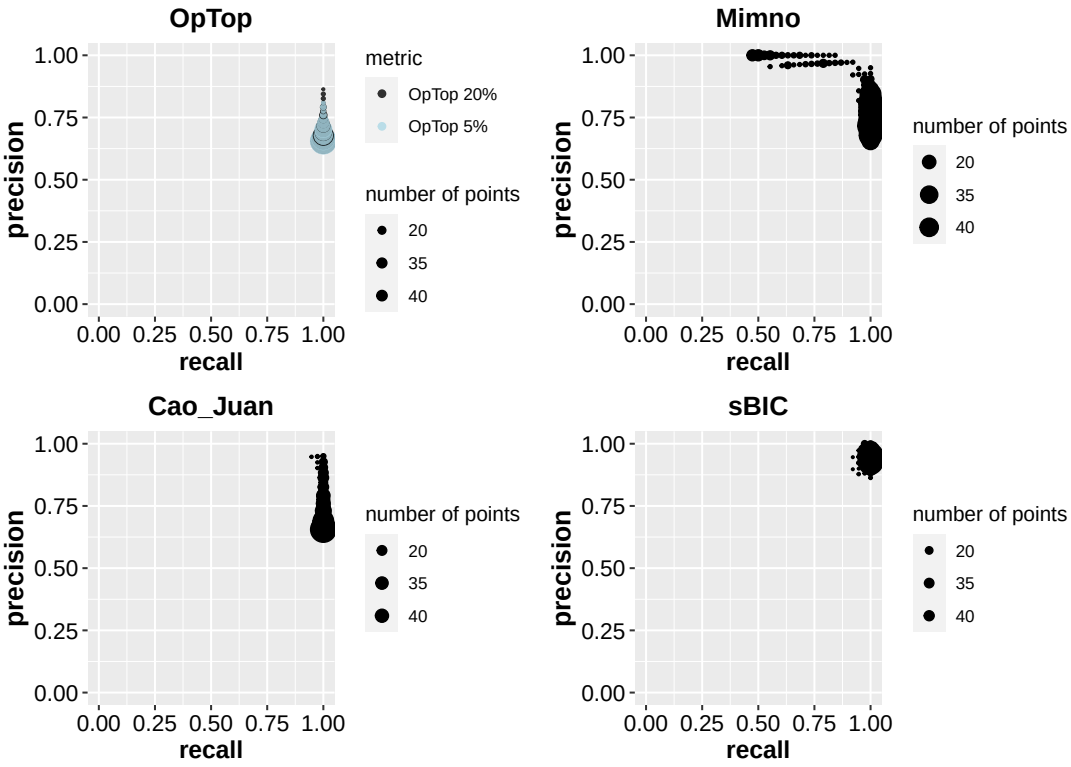


Figure 9: Precision and recall for DGP1

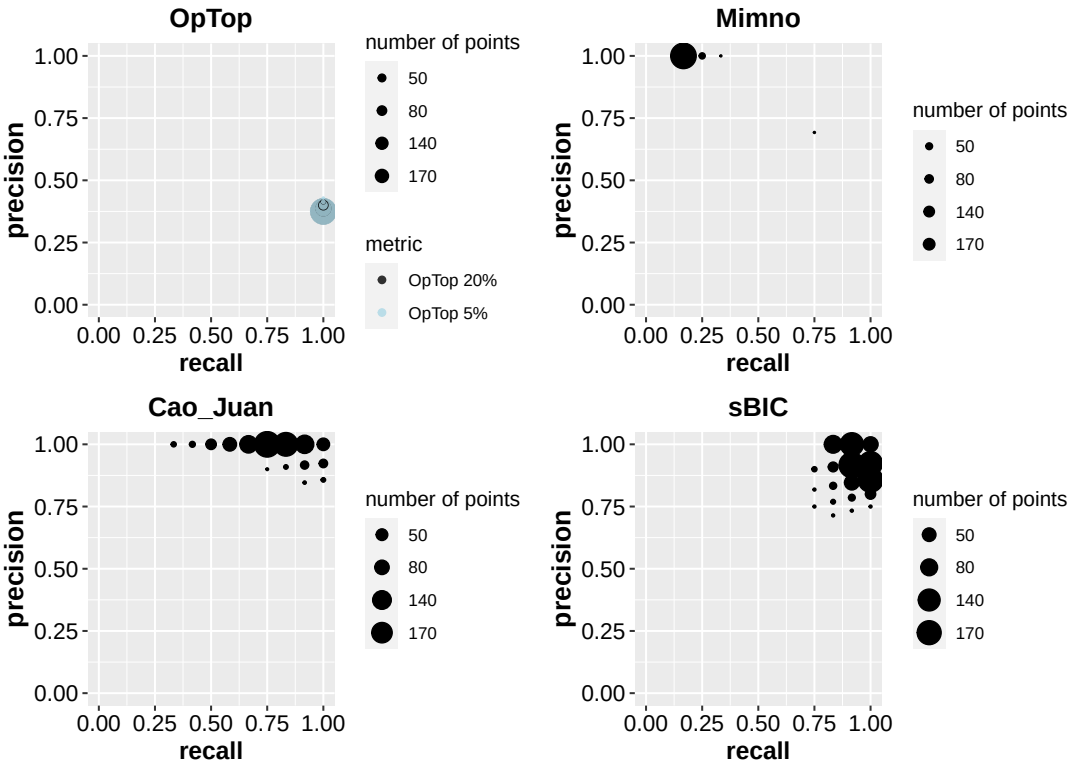


Figure 10: Precision and recall for DGP2

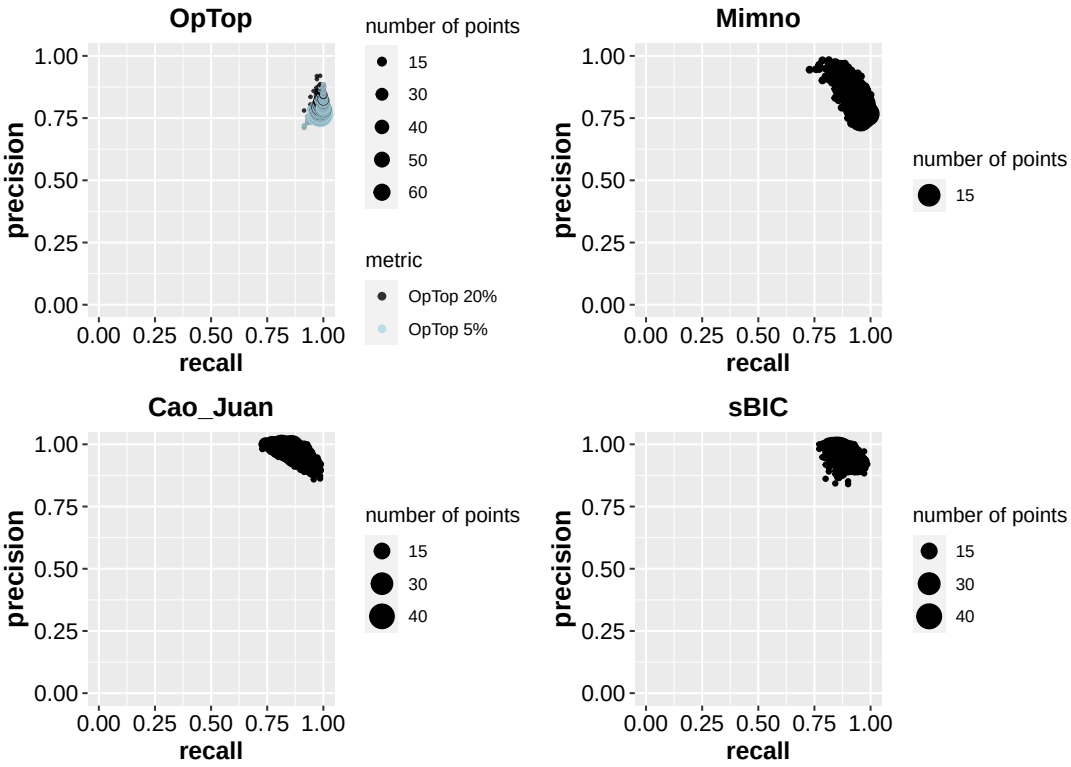


Figure 11: Precision and recall for DGP3

Appendix D. 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:

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

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

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 4 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 3 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.

While recall and precision values of our standard implementation are discrete leading to clustering of points in the scatter plots shown in Appendix C, 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 C taking into account the clustering, but standard scatter plots in Figures 12, 13, and 14.

		Recall		Precision		F1	
data	metric	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 4: Descriptive statistics of recall, precision, and F1 scores based on cosine similarity

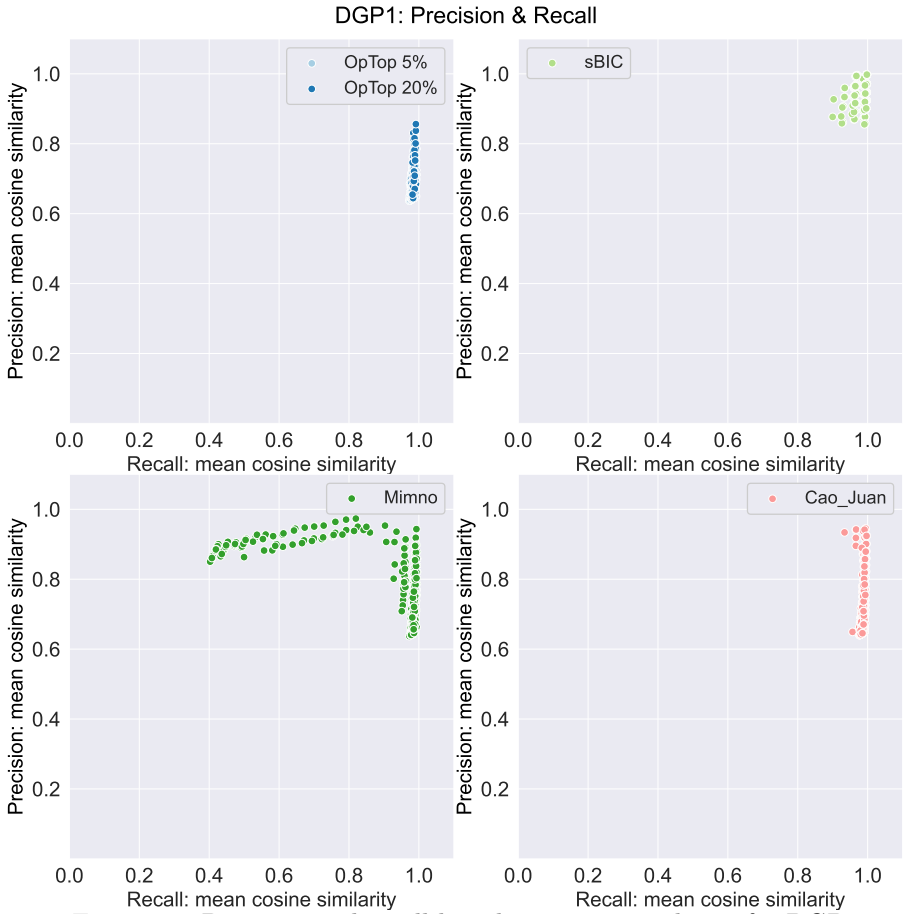


Figure 12: Precision and recall based on cosine similarity for DGP1

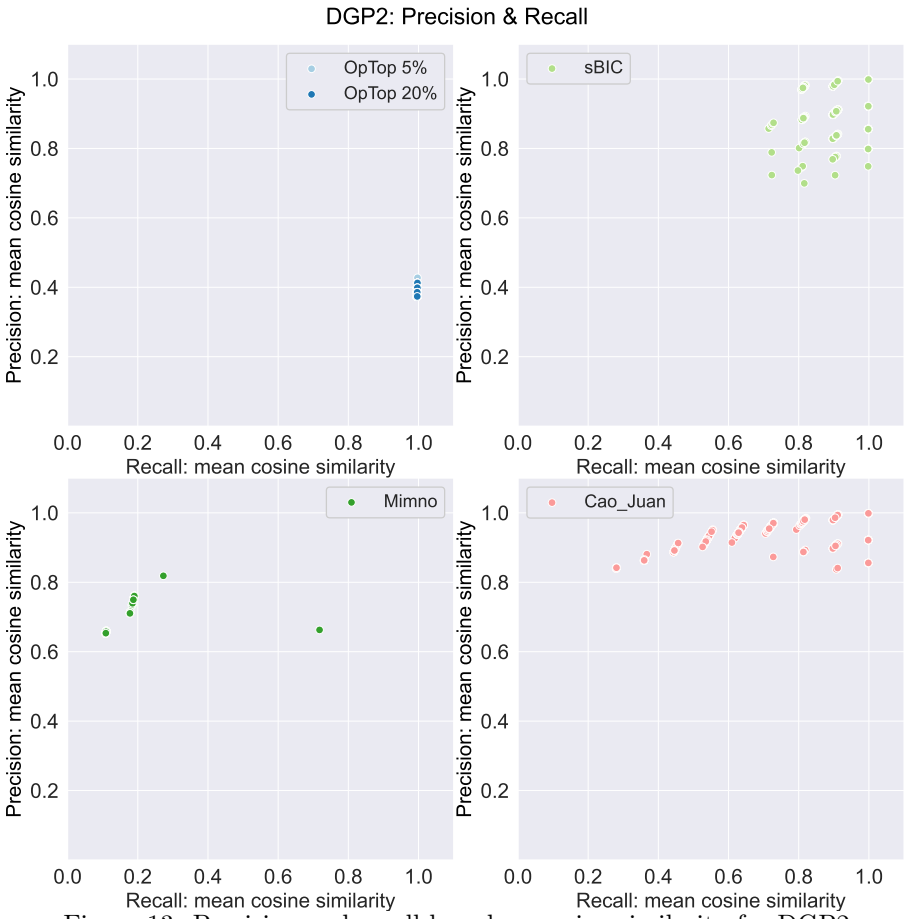


Figure 13: Precision and recall based on cosine similarity for DGP2

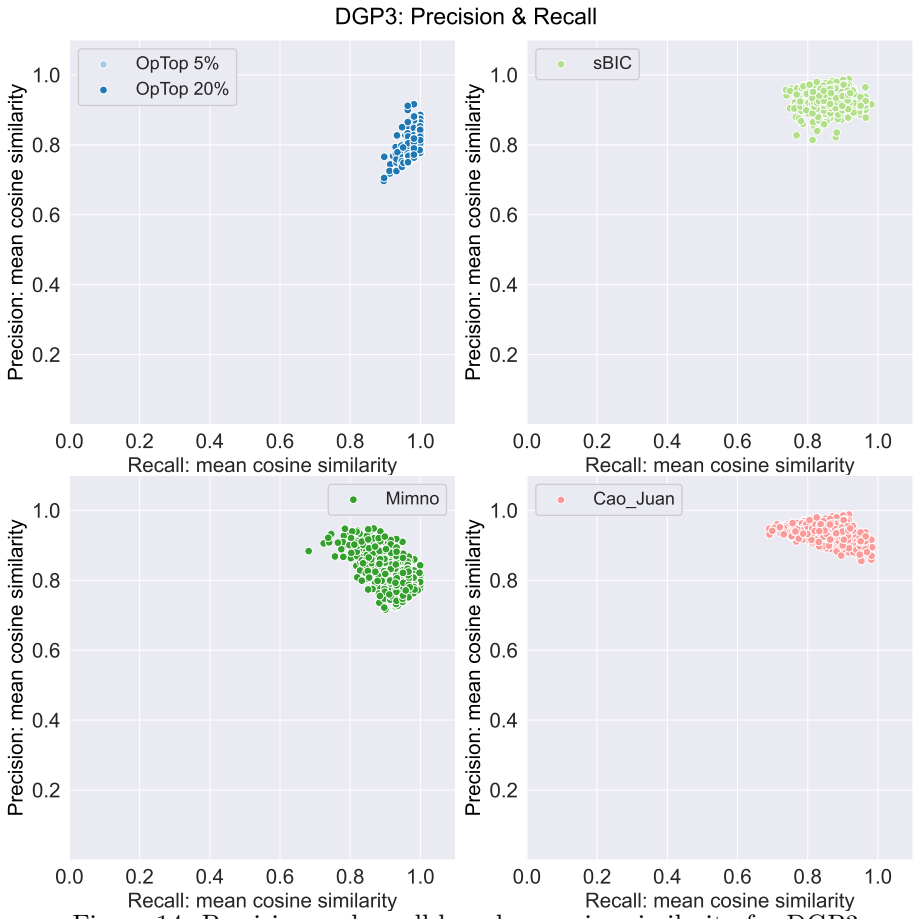


Figure 14: Precision and recall based on cosine similarity for DGP3

References

Philipp Adämmer and Rainer A. Schüssler. Forecasting the equity premium: Mind the news! *Review of Finance*, 24(6):1313–1355, 2020. doi: https://doi.org/10.1093/rof/rfaa007. URL `https://academic.oup.com/rof/article/24/6/1313/5788550`.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

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

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7):1775 – 1781, 2009.

Johannes Dahlke, Kristina Bogner, Maike Becker, Michael P. Schlaile, Andreas Pyka, and Bernd Ebersberger. Crisis-driven innovation and fundamental human needs: A typologi-

cal framework of rapid-response covid-19 innovations. *Technological Forecasting & Social Change*, 169, 2021. ISSN 0040-1625. doi: https://doi.org/10.1016/j.techfore.2021.120799. URL https://www.sciencedirect.com/science/article/pii/S0040162521002316.

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

Jon Ellingsen, Vegard H. Larsen, and Leif Anders Thorsrud. News media versus fred-md for macroeconomic forecasting. *Journal of Applied Econometrics*, 37(1):63–81, 2022. doi: https://doi.org/10.1002/jae.2859.

Federico M. Ferrara, Donato Masciandaro, Manuela Moschella, and Davide Romelli. Political voice on monetary policy: Evidence from the parliamentary hearings of the european central bank. *European Journal of Political Economy*, 74:102143, 2022. doi: https://doi.org/10.1016/j.ejpoleco.2021.102143. URL https://www.sciencedirect.com/science/article/pii/S0176268021001178.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Philipp Hartmann and Frank Smets. The european central bank's monetary policy during its first 20 years. *Brooking Papers on Economic Activity*, Fall 2018:1–146, 2018. ISSN 1533-4465. doi: https://www.brookings.edu/wp-content/uploads/2018/09/Hartmann-Smets_final-draft.pdf.

Naoki Hayashi. The exact asymptotic form of Bayesian generalization error in latent Dirichlet allocation. *Neural Networks*, 137:127–137, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.01.024. URL https://www.sciencedirect.com/science/article/pii/S0893608021000320.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. Measuring emotions in the COVID-19 real world worry dataset. *arXiv:2004.04225*, 2020. doi: https://doi.org/10.48550/arXiv.2004.04225.

Craig Lewis and Francesco Grossetti. A statistical approach for optimal topic model identification. *Journal of Machine Learning Research*, 23:1–20, 05 2022.

Alex Yu-Ting Lin and Saori N. Katada. Striving for greatness: status aspirations, rhetorical entrapment, and domestic reforms. *Review of International Political Economy*, 29, NO. 1:175–201, 2022. ISSN 1466-4526. doi: https://doi.org/10.1080/09692290.2020.1801486. URL https://www.tandfonline.com/doi/full/10.1080/09692290.2020.1801486.

Sandra Maria Correia Loureiro, Jao Guerreiro, and Iis Tussyadiah. Artificial intelligence in business: State of the art and future research agenda. *Journal of Business Research*, 129:911–926, 2021. ISSN 1535-3966. doi: https://doi.org/10.1016/j.jbusres.2020.11.001. URL https://www.sciencedirect.com/science/article/pii/S0148296320307451.

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

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1024.

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

Efstathios Polyzos and Fang Wang. Twitter and market efficiency in energy markets: Evidence using lda clustered topic extraction. *Energy Economics*, 114:106264, 2022. doi: https://doi.org/10.1016/j.eneco.2022.106264. URL https://www.sciencedirect.com/science/article/pii/S0140988322004017.

Ivan Savin and Nikita Teplyakov. Topics of the nationwide phone-ins with Vladimir Putin and their role for public support and Russian economy. *Information Processing & Management*, 59(5):103043, 2022. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.103043. URL https://www.sciencedirect.com/science/article/pii/S0306457322001480.

Gideon Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

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

Sarah Tiba, Frank J. van Rijnsoever, and Marko P. Hekkert. Firms with benefits: A systematic review of responsible entrepreneurship and corporate social responsibility literature. *Corporate Social Responsibility and Environmental Management*, 26(2):265–284, 2018. ISSN 1535-3966. doi: https://doi.org/10.1002/csr.1682. URL https://onlinelibrary.wiley.com/doi/full/10.1002/csr.1682.

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

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