Erasmus University Rotterdam
Erasmus School of Economics

Master Thesis Econometrics

Analyzing Text Data in Economics:
The Stability and Performance of LDA

Written by E.J. van de Winkel (student number 531022)

First assessor dr. C. Cavicchia (Erasmus School of Economics)
Second assessor dr. M. van de Velden (Erasmus School of Economics)

External supervision from:
dr. A.A. Dubovik (CPB Netherlands Bureau for Economic Policy Analysis)
dr. M.A.C. Kattenberg (CPB Netherlands Bureau for Economic Policy Analysis)
prof. dr. C.N. Teulings (Utrecht School of Economics)
dr. M.P. Schraagen (Utrecht University Department of Information and Computing Sciences)

February 15, 2022

Abstract

One machine learning method that has become increasingly popular to study text data within economics is Latent Dirichlet Allocation (LDA). LDA can be applied to a corpus of text documents to determine which latent topics underlie these documents. The found topics can then be used as a classification of the documents in the corpus. LDA is a stochastic algorithm and each run of the LDA algorithm thus produces a different topic classification. Previous research has shown that when LDA is applied to non-structured text data, the results of the algorithm are not stable. To account for this instability, new forms of LDA have been developed that can be applied to non-structured text data. Researchers have not yet determined the stability of LDA on structured text data, which is the research objective of this Thesis. Making use of the data and framework as used by Bandiera, Prat, Hansen, and Sadun (2020) in their paper, we show that when LDA is applied to structured text data there is instability in the LDA results. However, this instability is relatively small and does not affect the main conclusion of the paper of Bandiera et al. (2020) when statistical significance is taken into account. Besides, we show that the perplexity, which is a statistical measure of model fit, can be used as a selection criterion for each run of the LDA algorithm to decrease the instability: if only the runs are being selected with the lowest perplexity, the LDA results are stable. These results are robust to changes in both the vocabulary size and the choice of the prior hyperparameters. Furthermore, we show that the LDA algorithm produces results that are only moderately correlated to results produced by Lasso regression, indicating that both methods serve a different purpose and cannot be used interchangeably.

Table of Contents

1 Introduction

In the last couple of years there has been a rapid increase in the use of machine learning algorithms in economics. These algorithms have come to rise as they can help answer new research questions that could not have been answered using standard econometric techniques. Where econometric methods focus on establishing the relationship between variables, machine learning methods try to cluster, classify and predict data sets (Athey, 2019; Mullainathan & Spiess, 2017). With machine learning it is possible to uncover unobserved structures in a data set without any pre-specification, which is different from running regressions in which all variables need to be pre-specified: in machine learning the functional form of the model is determined based on the data. This feature of machine learning methods makes it possible to study new data types. One of the new types of data that can be studied is text data (Athey & Imbens, 2019; Gentzkow, Kelly, & Taddy, 2019). Recent examples are Gissler, Oldfather, and Ruffino (2016) who analyse bank comments to set up a new measure of banks' perception of uncertainty and Cohen, Malloy, and Nguyen (2020) who show that changes in the language of financial reports, can predict a firm's future returns.

One of the machine learning methods that economists use to study text data is Latent Dirichlet Allocation (LDA). LDA was developed by Blei, Ng, and Jordan (2003) to discover latent topics in a text data set. In their model, a data set consists of documents which in turn are formed by words. The LDA algorithm determines a posterior distribution over topics for each document based on the words it consists of. For example, if you have a data set with documents about movie descriptions, the algorithm could determine if a movie description belongs to the topic "horror" or "comedy". Different from other clustering methods like k-means (MacQueen et al., 1967), LDA does not provide the researcher with a strict clustering: where k-means would indicate if a movie description is either horror or comedy, LDA would classify the description as a mixture of both topics (Mazarura & De Waal, 2016). So applying the LDA algorithm to a specific movie description will result in a vector of probabilities corresponding to each topic, e.g. a movie description corresponds to the topic of horror with a probability of 0.95 and to the topic of comedy with a probability of 0.05. The labels of the topic are not provided by the algorithm itself, and need to be interpreted based on the output of the LDA algorithm.

LDA is very useful in economic research as it has been proven to classify text data the same as human interpreters do at a much lower cost (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). Besides, it has a high representational power, because of the assumption that a document is not generated from a single topic but from a mix of topics (Anandkumar, Foster, Hsu, Kakade, & Liu, 2015). This assumption is specifically relevant for large text documents, as it is rather unlikely that such documents have only one specific topic at its core. Within economics also dictionary methods are used to classify text data. In these methods researchers choose keywords and/or phrases themselves to classify documents into certain topics (Bao & Datta, 2014). Hansen, McMahon, and Prat (2018) show that LDA is preferred to these methods as it is

more objective and flexible to different data sets as it requires less pre-specification.

Within economics, LDA has three main usages. In the first case, LDA is used to construct an index based on text data (Azqueta-Gavaldón, 2017; Bandiera et al., 2020; Larsen & Thorsrud, 2019). Using this new developed index, researchers are able to conduct empirical analyses based on text data. Bandiera et al. (2020), for example, determine a CEO behavior index based on CEO diary data to determine how CEO behavior affects firm performance. In a second approach, LDA is used to determine trends in text data over time (Dyer, Lang, & Stice-Lawrence, 2017; Hansen et al., 2018; Larsen, Thorsrud, & Zhulanova, 2021). E.g., Ambrosino et al. (2018) study all economic articles in the JSTOR database from 1845 to 2013 to see how the focus of economic research shifts over time. Similar to the second approach, LDA is also used in economics to determine topics in text data at one point in time. Stantcheva (2021) use LDA to classify US citizens in two profiles based on a socio-economic survey, whereas Weigel (2020) use it to determine the main topics that emerged during town hall meetings in Congo.

Although the number of times that LDA has been cited within economics has grown steadily over the preceding years, there are still some methodological concerns regarding this method that have not been addressed in the literature before. Hereby, a distinction is made between non-structured and structured text data: the first refers to pure text documents, whereas the second is used for data sets that have text data in an ordered way, e.g. survey data. As this method has a big scientific relevance to economics (it can study new forms of data and thus explores new research directions), it is of importance to elevate these concerns. This Thesis will seek to address two of the more important open research problems concerning the use of LDA within economics.

First, an issue with many machine learning algorithms, including LDA, is that different results are produced after each run of the algorithm. This is because the LDA algorithm is stochastic and thus has a certain degree of randomness to it. For LDA it has been proven that, when applied on non-structured text data, the algorithm in itself is not very stable (Agrawal, Fu, & Menzies, 2018; Mantyla, Claes, & Farooq, 2018). Here stability is defined as the ability of the LDA model to replicate its solutions (De Waal & Barnard, 2008). In other words, it has been shown that when LDA is applied to a non-structured text data set only once, the results are not reliable and they cannot be used to provide interpretation. A high model stability thus indicates that a model is appropriate to analyse the data (Greene, O'Callaghan, & Cunningham, 2014). New forms of the LDA model on non-structured text data have been developed that can increase its stability (Agrawal et al., 2018). The current literature all focus on the stability of LDA on non-structured text data. As Bandiera et al. (2020) show that LDA can also be applied to structured text data it is of importance to study the stability of LDA on this type of data to make a case that LDA produces or does not produce valid results on structured text data sets.

Second, within the current literature the output of LDA has been compared to the output of Lasso regression on non-structured text data sets (Blei & McAuliffe, 2010). Lasso regression is used by researchers when a regression equation contains a large number of independent variables. By shrinking multiple regressors at the same time to zero, Lasso regression serves as a technique to reduce the dimensions for a regression (Adragni & Cook, 2009). Gentzkow et al. (2019) explain that Lasso regression can be applied to text data, where Lasso results indicate which words in a data set are rare and not informative, and which words are informative. As LDA also serves as a technique to reduce the dimensionality in a data set (all information in a text data set is reduced to several topics), comparing LDA to Lasso makes an interesting case. Blei and McAuliffe (2010) show how an extended LDA algorithm outperforms Lasso regression on a non-structured text data set. No research yet has compared the output of LDA to the output of Lasso regression on the same structured text data set. As both methods serve a same goal, it is of importance to see if both methods give similar classification results on a single data set.

In this Thesis we will study both the stability of LDA on structured text data and we will compare the performance of LDA to Lasso on this type of data. To do this we will study the data used by Bandiera et al. (2020) in their paper on the influence between CEO behavior and firm performance. The LDA model used by Bandiera et al. (2020) will be referred to as the baseline model. We use the data by Bandiera et al. (2020) as it is publicly available and is the first structured text data set within economics to which LDA has been applied. Our analysis consists of three steps.

In the first step of the analysis, we run the baseline model for a total of 250 times. Each of the 250 runs of the algorithm yields a different parameter estimate. For these estimates we will determine the instability, which is defined as the standard deviation over all the runs. Our results suggest that the instability in these 250 runs is substantial (the standard deviation covers roughly 25% of the total range of the coefficients). This instability indicates that a single run of the LDA algorithm is not sufficient for finding reproduceable results that can be interpreted. This instability is largely removed (it is roughly a third of the instability over all 250 runs) by only considering those observations that are statistically significant. However, only selecting those runs of the algorithm that are statistically significant is a form of "cherry picking" (beforehand it is unknown if there is a significant effect), thus making a selection of the runs based on the significance is bad practice. However, we show that the perplexity, which is a measure of fit of the model, can serve as a rejection criterion to the LDA algorithm: runs of the LDA algorithm with a perplexity lower than a post-determined value can be interpreted as stable. Running the algorithm 250 times gives us a way to determine a critical perplexity, that can be used to indicate which of the 250 runs of the LDA algorithm are stable and can be interpreted, and which runs cannot be interpreted.

In the second step of the analysis we perform robustness checks of the baseline model. We are interested in two things: do the baseline results change if we use alternative model settings, and is the stability of the LDA

results influenced by the alternative model settings? We study two alternative model settings: a different vocabulary size and different hyperparameter values for the priors. The results indicate that choosing alternative model settings result in slightly different regression coefficient estimates, but the interpretation of the results is not different compared to the interpretation of the baseline results. Besides, we show that the instability of the LDA results for the alternative model settings is comparable to the instability of the baseline model results: the results are unstable, but the perplexity can serve as a rejection criteria to produce a set of stable results. However, there is one model setting that does provide us with very unstable results that cannot be accounted for. If we trim down the model too much, by reducing the total number of words in our data set to roughly 10% of the initial baseline model, the results are very unstable and this instability cannot be accounted for by considering the perplexity.

In the third step of our analysis we compare the output of LDA on our data set to the output of a Lasso regression. We do this in two different ways. Firstly, we check if the LDA topics can be interpreted in a similar way to the Lasso coefficient estimates. The LDA topics suggests that there are two types of CEO behavior: leader CEOs have a higher firm performance than manager CEOs. Leader CEOs are hereby characterized as having high-level meetings and taking part in many video meeting and conference calls, whereas manager CEOs mostly do site visits and have many meetings with people from production and suppliers. This clear distinction between leader and manager CEOs cannot be observed by interpreting the Lasso regression estimates. Although the Lasso estimates also suggest that high-level meetings are associated with a positive effect and site visits with a negative effect on firm performance, the Lasso estimates make no clear distinction between different forms of CEO behavior. Our second way of comparing both methods is considering the correlation between a CEO behavior index formed by the Lasso regression results and a CEO behavior index formed by using the LDA analysis. We observe that both indices are moderately correlated, suggesting that the results of the two methods are not clearly in accordance to each other. Both ways of comparing suggest that LDA does not produce results that are in agreement with the Lasso results. As we have already shown that LDA produces robust and interpretable on this data set, we conclude that LDA is more appropriate than Lasso on this data set. This suggests that LDA and Lasso cannot be used interchangeably, although they are both serving as a dimensionality reduction technique.

This Thesis is related to two strands of literature. The first is the literature related to the stability of LDA. Several papers, including Koltcov, Koltsova, and Nikolenko (2014), Agrawal et al. (2018) and Mantyla et al. (2018), find that when LDA is applied to a non-structured text data set, there is an instability of topics. This means that each run of the LDA algorithm provides the researcher with a different topic description, thus the results of a single run of the algorithm are not suitable for interpretation. To account for the topic instability, two different solutions are presented in the literature. A researcher can either use an updated version of the LDA algorithm that directly accounts for the instability (this is done by e.g. Agrawal et al., 2018; Greene et al., 2014; and De Waal & Barnard, 2008) or he can run the LDA algorithm multiple times with the same

settings and then select some of the runs based on an appropriate measure or clustering technique (this is done by e.g. Koltcov et al., 2014; Mantyla et al., 2018; Rieger, Rahnenführer, & Jentsch, 2020; Rieger, Koppers, Jentsch, & Rahnenführer, 2020). What all the papers in this literature have in common is that they study non-structured text data. This Thesis will contribute to the literature by looking at the case of structured text data, which has not been done before in the literature.

Secondly, this Thesis is related to papers that compare the performance of LDA to the performance of other machine learning algorithms. One of the algorithms that is closest to LDA is Latent Semantic Analysis (LSA). LSA is a topic model that retrieves a semantic connection between words (Iaria, Schwarz, & Waldinger, 2018). Analysing articles in *The New York Times*, Stevens, Kegelmeyer, Andrzejewski, and Buttler (2012) show that LDA and LSA both have their own strengths. Whereas LDA best learns descriptive topics, LDA is the best in creating an accurate semantic representation of a document. The first of these conclusions is shared by S. Lee, Song, and Kim (2010), as their analysis shows that LDA outperforms LSA when a document consists of multiple topics. Whether to use LDA or LSA thus depends on the objective of the study. Studying three different text data sets, Péladeau and Davoodi (2018) finds that Factor Analysis provides more coherent topics than LDA. A comparison of LDA and a Hierarchical Dirichlet Process on bug data reports by Limsettho, Hata, and Matsumoto (2014), shows that the performance of LDA is better. Multiple articles study how k-means compares to a combination of k-means and LDA on text data, but there is no uniform conclusion drawn from this analysis (Alhawarat & Hegazi, 2018; Bui, Sayadi, Amor, & Bui, 2017). This Thesis adds to this literature by studying how a Lasso regression compares to LDA on structured text data. This comparison has not been made before in the literature.

This Thesis is organized as follows. We study the LDA model in more detail in Section 2. Special attention is given to how parameter inference takes place within the LDA model. Section 3 takes a closer look at the work of Bandiera et al. (2020) as their work is the basis for this Thesis. Section 4 describes the methodology used within this Thesis. Section 5 provides our results, after which these are concluded and discussed in Section 6.

2 Latent Dirichlet Allocation

2.1 The Model

The LDA model has been developed by Blei et al. (2003) to study large text data sets, as well as other discrete data sets. This model has become popular as it is a completely unsupervised machine learning technique for determining topics in a large data set (Alghamdi & Alfalqi, 2015). The model requires little

pre-specification and a researcher does not need to specify the topics he is searching for beforehand. Following the explanation of the model given by Blei et al. (2003), we will explain the model using language related to text data. Although we use this type of language, the LDA model is not limited to applications in text data sets. E.g., Tang et al. (2012) apply LDA to satellite images, whereas Nakano, Yoshii, and Goto (2014) use it to analyse audio frames.

In the LDA description, **corpus** is used to refer to the text data set. A corpus is denoted by C and its size is equal to M. The elements in the corpus are called **documents**, which are denoted by D_i with $i \in \{1, 2, ..., M\}$. As the size of the corpus C is equal to M, there are in total M documents in the corpus, thus $C = \{D_1, D_2, ..., D_M\}$. Each document is a sequence of N_i **words** w_i, so $D_i = \{w_{i1}, w_{i2}, ..., w_{iN_i}\}$, where N_i is not necessarily the same across two documents. The words are the basic units of the corpus and they are defined as items from a **vocabulary** $\{1, 2, ..., V\}$ with a size V.

A concrete example of the practical use of this terminology is the following: the book *The Handmaid's Tale* (i.e., the corpus) by Margaret Atwood consists of 46 chapters (i.e., the documents), which are filled using words from the Oxford English Dictionary (i.e., the vocabulary). Although this example is very superficial, it gives an easy interpretation of the core terminology of the LDA algorithm. However, two notes have to be made about this example. Firstly, when LDA is applied to a corpus usually the documents itself are independent. Within the example, the documents are not independent as it are all book chapters of the same book. Secondly, the vocabulary of *The Handmaid's Tale* is smaller than the Oxford English Dictionary as not all the words in the dictionary are used in the book. The vocabulary of *The Handmaid's Tale* thus does solely exist of all the unique words in the Oxford English Dictionary that are being used by Margaret Atwood.

With the LDA algorithm, a researcher can search for latent topics in a corpus. We define a topic to be a distribution over the vocabulary (Blei, 2012) and denote the number of topics by K. A topic is characterised by words in the vocabulary that have a higher probability of occurrence. An important assumption in the LDA model is that the number of topics is known beforehand, and is thus given as input to the algorithm (in Section 2.3.1 we give more insights on choosing the number of topics for a LDA model). The LDA algorithm is generative, which means that the model contains a distribution of the corpus, after which it gives an indication of how likely a given word is. In the context of the LDA model, we thus have that the model tries to generate a document given the latent topics (Alghamdi & Alfalqi, 2015). The premise of the LDA model is that documents in a corpus can be represented as a random mixture over latent topics, where each word in a document is attributable to the presence of the topics. The model assumes that a document is a so-called "bag of words", thus there is no structure in a document and it is just a random mixture of words. As Blei (2012) states, this assumption is not realistic in general for documents, but it is reasonable when the goal of the analysis is to uncover the semantic structure of a corpus. Following Blei et al. (2003),

LDA assumes the following process for each document D_i in a corpus C:

1. Choose the number of words N_i in the document following the Poisson distribution, thus $N_i \sim \text{Poisson}(\xi)$.
2. Choose a topic distribution θ_i following a Dirichlet distribution, thus $\theta_i \sim \text{Dir}(\alpha)$. The parameter α can be interpreted as the per-document topic distribution.
3. For each of the N_i words w_{in} in the document D_i:
   (a) Choose a topic z_{in} following a multinomial distribution, thus $z_{in} \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_{in} from a multinomial probability distribution conditional on the topic z_{in} and β, thus $w_{in} \sim p(w_{in}|z_{in}, \beta)$. The parameter β can be interpreted as the per-topic word distribution.

In Figure 1 the LDA model is graphically represented. As we see in the figure, the LDA model consists of two "layers": the outer layer is the corpus and consists of M documents, whereas the inner layer shows the document which consist in itself of N words. The parameters α and β are pre-specified by a researcher and are given as input for the corpus layer. The parameter θ is sampled for each document, whereas z and w are sampled for each word in a document. The fact that w is shaded in the figure indicates that this variable is the only variable observed by a researcher.

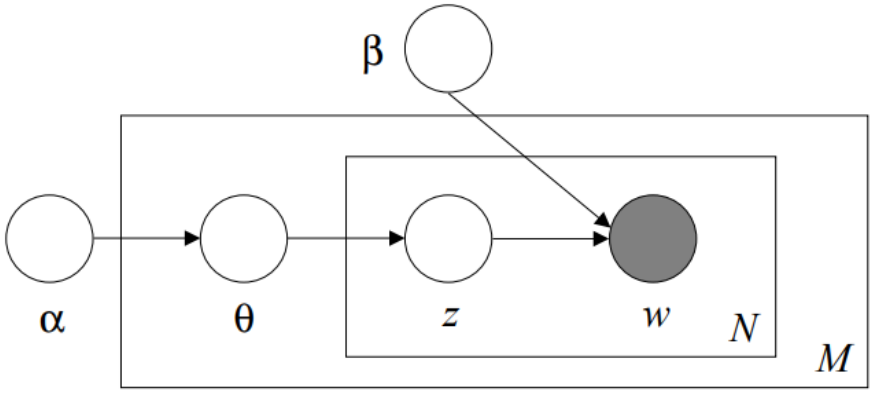


Figure 1: Graphic representation of the LDA model. This representation is taken from Blei et al. (2003).

2.2 Parameter Inference

To estimate the parameters of the LDA model, we set up the likelihood-function of the model. Following Blei et al. (2003), the likelihood-function is given by

$$\mathcal{L} = \prod_{d=1}^{M} \prod_{n=1}^{N_d} p(w_{dn}|z_{dn}, \beta) p(z_{dn}|\theta_d) p(\theta_d|\alpha), \tag{1}$$

where d is an index over the M documents in the corpus and n is an index over the N_d words in each document. A researcher who aims to find an estimate for θ can find this estimate by maximizing the likelihood-function (Equation 1). However, direct optimization of $\mathcal{L}$ is not possible as we do not directly observe the topics z_{dn} (Crain, Zhou, Yang, & Zha, 2012).

Within the literature, several methods have been proposed to solve the estimation problem of LDA. In this Thesis we will follow the collapsed Gibbs sampling algorithm for LDA inference developed by Griffiths and Steyvers (2004). We choose to use this algorithm as it is most used in the literature (Jelodar et al., 2019), and it is also the algorithm used for inference by Bandiera et al. (2020), which will be the framework of our analysis. Other well-known methods in the literature are expectation-maximization (Minka & Lafferty, 2012), variational approaches (Blei et al., 2003), matrix factorization (Hofmann, 2013; D. D. Lee & Seung, 1999) and the method of moments (Anandkumar et al., 2015).

2.2.1 Gibbs Sampling

Gibbs sampling is an algorithm from the class of Monte Carlo Markov Chain methods, and it is thus an algorithm used to sample from a probability distribution. In the case of Gibbs sampling, we want to sample from a joint distribution. However, for this model the joint distribution cannot be calculated, thus instead we focus on the marginal conditional distribution. Let us denote by $\{\theta_1, \theta_2, ..., \theta_d\}$ the variables in the model, by y the data, and by $p(\theta_1, \theta_2, ..., \theta_d|y)$ the joint distribution we wish to sample from. A standard Gibbs sampling algorithm is then given by (Crain et al., 2012; Greenberg, 2012):

1. Set initial values $\theta^0 = \{\theta_1^0, \theta_2^0, ..., \theta_d\}^0$ for the variables in the model and set a timer $m = 0$.
2. Simulate:
   - a value θ_1^{m+1} from the distribution $p(\theta_1|\theta_2^m, \theta_3^m, ..., \theta_d^m, y)$,
   - a value θ_2^{m+1} from the distribution $p(\theta_2|\theta_1^{m+1}, \theta_3^m, ..., \theta_d^m, y)$,
   - ...,
   - a value θ_d^{m+1} from the distribution $p(\theta_d|\theta_1^{m+1}, \theta_2^{m+1}, ..., \theta_{d-1}^{m+1}, y)$.
3. Set the timer $m \leftarrow m + 1$, and go back to step 2.

In the limit of the number of iterations m of the algorithm, the simulated values of θ^m can be used as a sample from the joint distribution $p(\theta_1, \theta_2, ..., \theta_d|y)$ (Greenberg, 2012). The first m^* iterations of a Gibbs sampler are used as a burn-id period: this period is used by the algorithm to converge to the joint distribution. All the iterations $m > m^*$ are averaged to form the final sample from the joint distribution (Crain et al., 2012).

The Gibbs sampling algorithm has been applied to the case of LDA parameter inference by Griffiths and Steyvers (2004). They developed a so-called collapsed Gibbs sampling algorithm, in which some of the model variables are marginalized out of the model. In the case of the LDA model, the only variable that is sampled is z_{dn} and this sampling is done conditional on the parameters α and β and the topic assignment of the other words in the model $z_{\bar{dn}}$ (Crain et al., 2012). After obtaining samples of z_{dn}, these samples can be used to compute an estimate for the topic distribution θ (Newman, Porteous, & Triglia, 2011).

Within the collapsed Gibbs sampling algorithm of LDA, the joint distribution may have a non-convex shape. Due to this shape, there are multiple local maxima (Griffiths & Steyvers, 2004; Koltcov et al., 2014). Because of these multiple local maxima, parameter estimates may be different in each run of the LDA algorithm. Furthermore, the choose of initial values in step 1 of the Gibbs algorithm may influence the outcomes of the Gibbs sampler, as is shown by Rieger, Koppers, et al. (2020). These two observations indicate that the collapsed Gibbs sampler does not produce similar results each time it is applied for inference on the same model and data set. There is thus some kind of instability in the results across multiple runs.

2.3 Model Pre-Specification

In the preceding sections, we looked at the LDA model and how parameters can be estimated for this model. As we explained, the LDA model is an unsupervised machine learning algorithm and it thus does not require a lot of model pre-specification. However, there are still some parameters that need to be initialised before the model can be applied to a data set: the number of topics K and the hyperparameters α (the per-document topic distribution) and β (the per-topic word distribution). Besides these parameters, a researcher using LDA has also some influence on the vocabulary size of the corpus he is using. We will now look into these three different ways to pre-specify the LDA model in more detail.

2.3.1 Number of Topics

A big assumption in the LDA model is that the number of topics K in the corpus is known beforehand. This assumption may be valid for small data sets for which the topics can be derived and checked manually, but this fails for a large corpus. It is, however, very important to have an accurate guess of the number of topics beforehand, as the accuracy of the LDA model is very sensitive to the number of topics that is pre-specified (Arun, Suresh, Madhavan, & Murthy, 2010). As we read in Zhao, Zou, and Chen (2014), if we choose K too small compared to the actual number of topics, the LDA model is too rough to accurately estimate the parameters. Choosing K too large compared to the actual number of topics, will result in a model that is too complex, and thereby cannot be interpreted.

Within the literature, multiple approaches have been designed to determine the number of topics. The classic approach consists of researchers trying different values of K and than comparing the models to each other in terms of a statistical measure, the so-called measure-based method (Zhao et al., 2015). This approach is updated by Zhao et al. (2015), who do look at the rate of measure change opposed to the measure itself. More recent methods for determining the value of K are based on putting a prior distribution on K and then estimating the value using a Metropolis-Hastings algorithm (Chen & Doss, 2019).

2.3.2 Prior Parameters

As described in Section 2.1, the LDA model has two parameters that are given as input for the corpus layer: the per-document topic distribution α and the per-topic word distribution β. The value of these hyperparameters determines the shape of the documents (α) and the shape of the topic (β). As we read in Syed and Spruit (2018), a large value of α gives rise to a document consisting of many topics, whereas a small value of α makes for a sparse model in which a document consists of a small number of topics. The same pattern can be observed for β: a large value makes for topics with many words, whereas a small value ensures that a topic consists of a small number of words. Following Wallach, Mimno, and McCallum (2009), we can distinguish between the following four hyperparameter settings for the pair (α, β): {AA, AS, SA, SS}, where A stands for an asymmetrical prior and S for a symmetric prior. In the case of AA, it is assumed that each topic has the same probability of being assigned to a document, thus $\alpha = 1/K$, and each word in a document has the same probability of being assigned to a topic, thus $\beta = 1/V$.

Wallach et al. (2009) studied how the symmetry or asymmetry of the hyperparameters influences the LDA mechanism. Compared to a symmetrical prior on α, an asymmetrical prior increases the robustness of the LDA model and leads to less sensitive selection of the number of topics. Choosing a symmetric or asymmetric prior on β does not influence the topic coherence. Newman et al. (2011) showed that hyperparameters can best be determined using a grid search over multiple LDA algorithms with different hyperparameter settings.

2.3.3 Vocabulary Size

Although a corpus has a fixed vocabulary size, a researcher can tweak the vocabulary size by removing certain words from the vocabulary. Applying LDA to a corpus results in a list of the most occurring words in each topic. As Schofield, Magnusson, Thompson, and Mimno (2017) makes clear, this view on LDA shows that a smaller vocabulary size is preferred in terms of representation of a topic, as words with a small frequency of occurrence do not yield much information to the topic distribution. However, Syed and Spruit (2017) show empirically that a larger vocabulary size makes a model more robust to "noise" words, which are words that yield little information to a topic distribution like *using*, *used* and *use*. Under these two views, both a too

small and a too large vocabulary are not preferred. Removing too much words from the vocabulary results in less information, whereas having a too large vocabulary size makes the model influenced by little-occurring words.

3 An Application of LDA within Economics

3.1 Introduction

Within this Thesis, we study the stability of LDA on structured text data. To do this, we make use of the data and framework of Bandiera et al. (2020). We use their data as it is the only application of LDA on structured text data within economics. Besides, the data and code of Bandiera et al. (2020) is publicly accessible. Using their data does not only save us the time of gathering data ourselves, but it does also ensure us that we study a relevant case of LDA within economics, as the work of Bandiera et al. (2020) has been published in the top journal *Journal of Political Economy*.

In their paper, Bandiera et al. (2020) determine how CEO behavior affects firm performance. Their analysis consists of two steps. In the first step a CEO behavior index is constructed using LDA assuming there to be two pure CEO behaviors. In the second step this index is used to regress CEO behavior on firm performance. From their analysis, Bandiera et al. (2020) find that CEOs can be characterized as either "leaders" or "managers". The activities of the leader CEOs are mostly related to high-end meetings with other executives, whereas manager CEOs spent most of their time on the working floor with firm employees and suppliers. It is shown that firms that have a leader CEO perform better than firms with a manager CEO on top.

In this section we will give a short summary of the paper of Bandiera et al. (2020) to ensure that the framework in which we are working is clear. We will look at their data and both steps of the analysis. Furthermore, we shortly reflect on how the authors account for the instability of LDA within their paper.

3.2 Data

Bandiera et al. (2020) want to construct an index for CEO behavior. Within the business literature, there are two main streams of literature related to characterising CEO behavior. In the first research stream (e.g. Mintzberg, 1973) researchers try to obtain information about CEO behavior by personal observations of CEOs. Bandiera et al. (2020) denote that this way of observing CEO behavior produces rich and effective data. However, this method is not suitable for obtaining a big sample. The statistical analysis of data in

this stream of research is thus often difficult due to small samples sizes. In a second stream of research (e.g. Hermalin, 1998), researchers develop categorizations of leadership styles based on theoretical models. As Bandiera et al. (2020) make clear, this way of characterising CEO behavior is difficult to turn into an empirical model.

To overcome the issues that are currently present in the literature, Bandiera et al. (2020) develop a new methodology that provides them with a large data set that is suitable for statistical analysis. Instead of shadowing CEOs in real-life themselves, Bandiera et al. (2020) opt for shadowing CEOs' diaries instead. This way of data collecting provides them with a data set consisting of $1,114$ CEOs in six countries. For each CEO in the data set, they have information about all the meetings the CEOs had in one specific week. All of the meetings have been divided into 15-minute time blocks, which they call activities. For each activity the following features are known: (1) the type of meeting, (2) the total duration of the meeting, (3) whether the meeting was planned or not, (4) the number of participants of the meeting, and (5) the function of the participants of the meeting. The final and cleaned data set of Bandiera et al. (2020) consists of $42,233$ activities. For each activity the information about the five features is stored in a string. The activity that is mentioned the most times in the data set (a total of $3,260$ times) is *interacting__meeting__1hrplus__planned__two_plus_pplzzzzproduction*, which corresponds to an interacting meeting (1) of more than one hour (2), that was planned (3), visited by more than two persons (4) working for production (5). Of the $42,233$ activities in the data set, $4,253$ activities are unique.

The CEO diary data can be seen as a structured text data set. As we have seen each activity in the data set is denoted as a string in which the features are written down. The features can be interpreted semantically and there is a connection between the features in a string, thus the data is presented in a structured textual format. We can also define this model using the terminology of the LDA model as defined by Blei et al. (2003). The CEO diary data set that we have is our corpus. This CEO diary data set consists of data about $1,114$ CEOs. The set of activities of a specific CEO are the documents in the model, and there thus are $1,114$ documents. Each document is filled with activities of a specific CEO. The activities are the words in this corpus. We have seen that there are in total $4,253$ unique activities mentioned in our corpus, thus this is the vocabulary size.

3.3 Step 1: LDA Analysis

The CEO diary data is analysed using LDA. We will follow Bandiera et al. (2020) closely in the explanation of their LDA procedure. As we have seen in the previous section, the data set consists of information regarding the activities a CEO undertakes in a single week. In total there are $4,253$ unique activities in the data set. We define by $X = \{x_1, x_2, ..., x_{4,253}\}$ the set of all the different activities in the LDA model. We define a pure

CEO behavior m to be a probability distribution β^m over the set of all activities X, and this behavior is similar for all the CEOs in the data set. Based on the work by Kotter (2001), Bandiera et al. (2020) assume that there exist only two pure CEO behaviors: β^0 and β^1. In the terminology of the LDA model, the two pure CEO behaviors are the latent topics in our data set and each activity a CEO undertakes can be seen as a mixture between β^0 and β^1. The weight a CEO i gives to pure behavior β^0 is denoted by $\theta_i \in [0, 1]$. The weight θ_i is called the CEO behavior index of CEO i and it is interpreted as the probability that the activity of the CEO is the result of β^0.

For each activity x_a^i of CEO i in the CEO diary data set, the LDA procedure can now be described as follows:

1. Given the CEO behavior index θ_i, one of the two pure behavior β^0 and β^1 is drawn.
2. Given the drawn pure behavior, an activity is drawn according to the corresponding probability distribution β^0 or β^1.

A graphical representation of the LDA procedure is shown in Figure 2. Using this figure, we observe that for each activity, a CEO i first chooses one of the two pure behaviors given the CEO behavior index θ_i. Then, given this pure behavior, the CEO chooses one of the activities from the set X of all activities. The probability that a CEO i gives to the activity x_a^i is now given by $(1-\theta_i)\beta_a^0 + \theta_i\beta_a^1$. If we denote by $n_{i,a}$ the number of times the activity a appears in the time of CEO i we can write the complete likelihood function as

$$\mathcal{L} = \prod_{i=1}^{1{,}114} \prod_{a=1}^{4{,}253} \left((1-\theta_i)\beta_a^0 + \theta_i\beta_a^1\right)^{n_{i,a}},$$

where we use the independence of i and a.

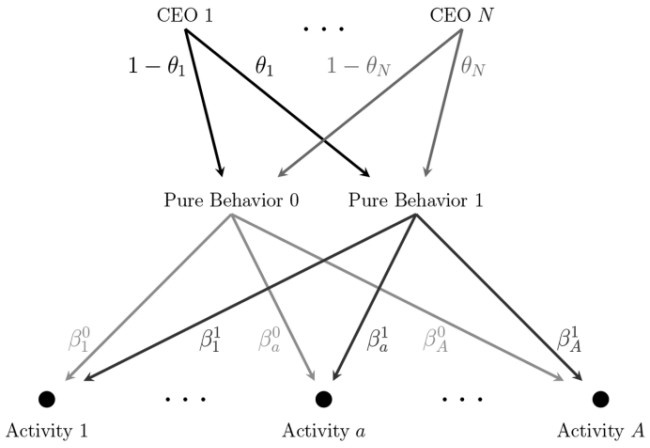


Figure 2: Graphic representation of the LDA model as applied by Bandiera et al. (2020). This figure is taken from Bandiera et al. (2020).

For the likelihood function just discussed, we want parameter estimates of θ and β using Bayesian techniques. In line with the LDA model, Bandiera et al. (2020) place a Dirichlet prior on both of the parameters. The parameters in the model are estimated using the Monte Carlo Markov Chain approach of Griffiths and Steyvers (2004) (see Section 2.2.1 for more information about the collapsed Gibbs sampling procedure).

Bandiera et al. (2020) use the following specifications in their LDA model. From now on in this Thesis we refer to the LDA model with these settings as the baseline model:

- the number of topics $K = 2$,
- the prior of θ follows a symmetric Dirichlet distribution with hyperparameter 1.0,
- the prior of β follows a symmetric Dirichlet distribution with hyperparameter 0.1,
- a total of 10,000 burn-in iterations is used,
- there are 160 samples collected after the burn-in period with a thinning interval of 50,
- the activities that are mentioned in less than 30 CEOs' diaries are removed.

Applying the LDA procedure as explained above on the CEO diary data set gives estimates of θ and β. Using these estimates we can calculate which activities in the data set are less likely to occur under pure CEO behavior β^1 compared to behavior β^0 and which activities are more likely to occur under this behavior. In Table 1 we observe that the activities that are less likely to occur under behavior β^1 compared to behavior β^0 are plants visits, meetings with only outsiders, meetings with people from production and meetings with suppliers. These activities are related to manager CEOs, and as these activities are less likely to occur in behavior β^1, we can denote behavior β^0 as corresponding to a manager CEO. The activities that are more

Table 1: The activities that are most and least likely to occur under CEO behavior β^1 are shown. A value below 1 indicates that the activity is related to CEO behavior β^0, whereas a value above 1 indicates that the activity is related to CEO behavior β^1. This table is taken from Bandiera et al. (2020).

Feature	Times Less/More Likely
Less likely in behavior 1:	
Plant visits	.11
Just outsiders	.58
Production	.46
Suppliers	.32
More likely in behavior 1:	
Communications	1.90
Outsiders and insiders	1.90
C-suite	33.90
Multifunction	1.49

likely to occur under CEO behavior β^1 compared to CEO behavior β^0 are meetings with people from communications, meetings with both outsiders and insiders, meetings with C-suite executives and meetings with people who have multiple functions. These activities are related to leader CEOs, thus we can denote behavior β^1 as corresponding to a leader CEO.

3.4 Step 2: Regression Analysis

Step 1 of the analysis of Bandiera et al. (2020) provides them with an estimate of the CEO behavior index: $\hat{\theta}$. In a second step of the analysis, the estimate is used to determine how CEO behavior affects firm performance. The following simple regression is run:

$$y_{ifts} = \alpha\hat{\theta}_i + CONTROLS + \varepsilon_{ifts}, \tag{2}$$

where y_{ifts} is the log of sales of a firm f operating in sector s led by CEO i at time t and ε is an idiosyncratic error term. We thus have a regression of the estimated CEO behavior index $\hat{\theta}$ on the firm performance y, which is measured in terms of the log of sales of a firm. The parameter of interest is α, and this parameter can be interpreted as the correlation between firm performance and the CEO behavior index. In their paper, Bandiera et al. (2020) find that $\hat{\alpha} = 0.354(0.108)$ which is significant at the 1% level. As θ is related to β^1 this shows that CEOs who are of type β^1 (this type of behavior is related to leaders) have higher firm profits than CEOs who are of type β^0 (this type of behavior is related to managers).

3.5 Multiple Runs of the LDA Algorithm

The main objective of this Thesis is to determine the stability of LDA. As each run of the LDA algorithm provides us with a different estimate of $\hat{\theta}$, each run of the LDA algorithm in return yields a different parameter estimate $\hat{\alpha}$, which is our parameter of interest. Although Bandiera et al. (2020) do not denote it clearly in their paper, their code shows that a small stability check on LDA has been run.

In their code, we see that Bandiera et al. (2020) run the LDA algorithm five times on the CEO diary data set. For each of these five runs of the algorithm (Bandiera et al. (2020) call it five different chains), Figure 3 shows how the perplexity changes over the number of burn-in iterations. Without going into much detail here (we will do this in Section 4.5), the perplexity is a statistical measure of fit. In general, we can state that a model with a lower perplexity, is a better fit to the data. In the figure we observe that four of the five chains have a perplexity that is very similar to that of the others. However, one of the five chains has a perplexity which is way higher and this model should by disregarded from the analysis based on the perplexity. From the remaining four chains, Bandiera et al. (2020) choose to use chain 3 in their subsequent

analysis. Looking at Figure 3 this may be puzzling as chain 4 has a lower perplexity value than chain 3 at the end of the burn-in period. Although Bandiera et al. (2020) do not state in their code why they choose chain 3 instead of chain 4 it may be due to a learning effect. We observe that the perplexity of chain 3 is declining over the number of iterations, whereas the perplexity of chain 4 stays more or less constant over the entire burn-in period. The model that is formed by chain 3 thus has learned more than the model formed by chain 4 and may thus be interpreted as better mimicking the data set.

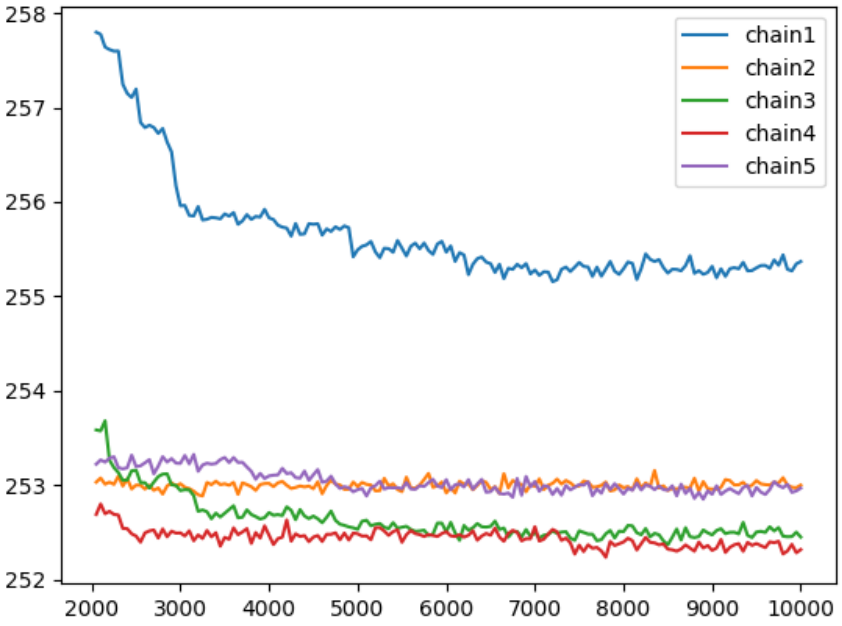


Figure 3: For five different runs of the LDA algorithm (called chains), the perplexity is shown on the vertical axis as a function of the number of burn-in iterations on the horizontal axis. This figure is taken from the code of Bandiera et al. (2020).

An extra step to the analysis of Bandiera et al. (2020) is to calculate the regression coefficient estimate $\hat{\alpha}$ for each of the five chains of the LDA algorithm as shown in Figure 3. In Table 2 we present the results of this analysis. We observe two interesting features. The first thing is that not all of the regression estimates are positive. Chain 1 finds an effect of CEO behavior on firm performance in the opposite direction of the effect found in the paper by Bandiera et al. (2020). This chain would thus invalidate their results. There are two ways to possibly deal with this result. The first way is to disregard this found estimate based on a too high value of the perplexity of the chain as we see in Figure 3. A second way is to check if the labeling of the CEO behaviors is the same in this chain as compared to chain 3 (which is the chain used in the paper by Bandiera et al. (2020)). The LDA model assigns labels to behavior β^0 and β^1 randomly thus it may be the case that chain 1 uses a different labeling than chain 3 (we will go into more detail on this issue in Section

Table 2: For the five different chains of the LDA algorithm that are shown in Figure 3, the estimate $\hat{\alpha}$ is determined. Chain 3 is the value of the parameter estimate as reported in the paper by Bandiera et al. (2020). The number of stars is an indication of the statistical significance. An estimate with $**$ is statistically significant at a 5% level of significance, whereas an estimate with $***$ is statistically significant at a 1% level of significance.

	Chain 1	Chain 2	Chain 3	Chain 4	Chain 5
$\hat{\alpha}$	-0.212^{**}	0.362^{***}	0.343^{***}	0.324^{***}	0.351^{***}

4.3). If this would be the case, the negative value of chain 1 needs to be interpreted as a possible value. What would remain is that the coefficient estimate is than still substantially smaller compared to the other four chains. The second interesting thing from this analysis, is that chains 2, 3, 4 and 5 all have coefficient estimates that are very similar to each other. This may be an indication that the LDA algorithm provides stable results on the data set for different chains of the LDA model who have a similar perplexity (see Figure 3).

4 Methodology

4.1 Introduction

In Section 2.2.1 we have seen that when the collapsed Gibbs sampling procedure is applied to estimate the parameters in the LDA model there are multiple local maxima. In the context of the paper of Bandiera et al. (2020), this means that each new run of the LDA algorithm provides us with a new estimate of the CEO behavior index $\hat{\theta}$ and thus also with a new estimate of the regression coefficient $\hat{\alpha}$. As the regression coefficient is changing in each run of the algorithm, there will be instability in the regression coefficient over all the runs of the algorithm. To determine the instability of the LDA algorithm on structured text data (which is the type of data as used by Bandiera et al. (2020)), we will run the LDA analysis N times. These N estimates provide us with N different estimators for the regression coefficient $\hat{\alpha}$. We define the instability of the regression coefficient α as the standard error of its N estimates, thus:

$$\text{instability} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\alpha}_i - \bar{\alpha})^2}{N}}, \tag{3}$$

where $\bar{\alpha} = \frac{\sum_{i=1}^{N}\hat{\alpha}_i}{N}$ is the mean of $\hat{\alpha}$ over all the N runs of the algorithm. A smaller instability, indicates that there is less variation in the regression coefficient estimates, and thus that the value of the coefficient is less influenced by the specific run of the LDA algorithm. A smaller instability is preferred as this means

that the LDA algorithm provides interpretable and generalizable results.

In the next section of this Thesis we will discuss the results of our instability analysis. Before we turn our focus to these results, we will cover the methodology of the paper. We will first describe how we will implement the LDA analysis and the regression equation (Equation 2). After this we will shortly explain our considerations when deciding on how many times we will run the LDA algorithm. Then we will cover some issues that we have to deal with when we are determining the stability of LDA. Furthermore, we explain how we will implement Lasso regression to the CEO diary data set.

4.2 Implementation of LDA and the Regression Equation

The first step in the analysis is to apply LDA to the CEO diary data set. For the implementation of LDA, we use the code as used by Bandiera et al. (2020) in their paper. They implement LDA in Python using the library topicmodels.

The second step in the analysis is to run regression equation 2. Bandiera et al. (2020) make use of the programming language Stata and use the command areg for running the regression. This command is used for running a regression on a data set with a large number of dummy variables. As running a statistical analysis on multiple runs of a regression equation is difficult in Stata, we use R in our analysis of the regression equation. The regression equation is run using the function felm from the library lfe. Also for the implementation of the Lasso regression we use R; more specifically we use the function glmnet from the package glmnet.

In Table 10 in Appendix A we compare the Stata regression output to the R regression output on the baseline model as is also discussed by Bandiera et al. (2020). We observe that there is no difference in the estimated regression coefficients between the two different programming languages. However, there is some difference in terms of the standard errors. The standard errors as reported by Stata are for each variable slightly higher than the standard errors as produced by R. This is likely due to the fact that standard errors are clustered in different ways in both programming languages and functions. This small differences makes some variables not statistically significant in Stata which are statistically significant in R. In our analysis we are only interested in the estimate of the regression coefficient α and not in its standard error. For our analysis it is thus not important that the standard errors in R are different than those produced by the Stata code.

4.3 Number of LDA Runs

The main objective of this Thesis is to determine how stable the results produced by LDA are, when we run the LDA algorithm multiple times. Running the LDA algorithm once on the CEO diary data takes approximately seven to eight minutes in Python. Based on this running time we decide to run the baseline LDA model for 250 runs. This takes us approximately 30 hours and ensures us that we have a sample of LDA estimates that is large enough for statistical analysis. In the discussion of this Thesis we will also discuss how our results are influenced by choosing less than 250 runs.

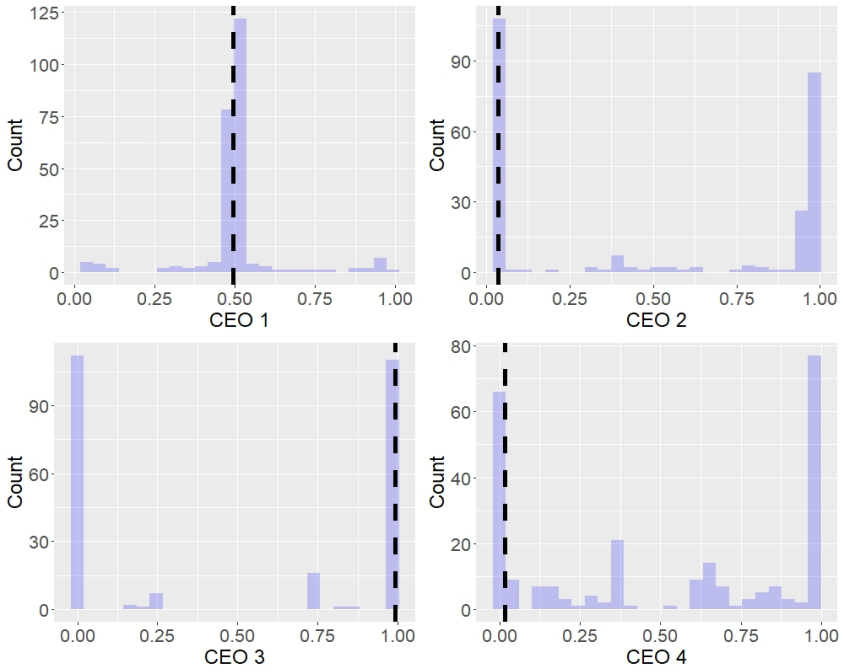
4.4 Random Indexing

A first issue related to running the LDA algorithm multiple times, is that in each run of the algorithm the labelling of the behaviors is random. This is a problem we have already encountered in Section 3.5. In that section we have seen that if we run the LDA algorithm multiple times not each run of the LDA algorithm provides us with a positive regression coefficient (see Table 2). In each run of the LDA algorithm, the labelling of a behavior as either behavior β^0 or β^1 is random. By the LDA algorithm itself it is not fixed which behavior gets which label as the labelling is determined by a human interpreter.

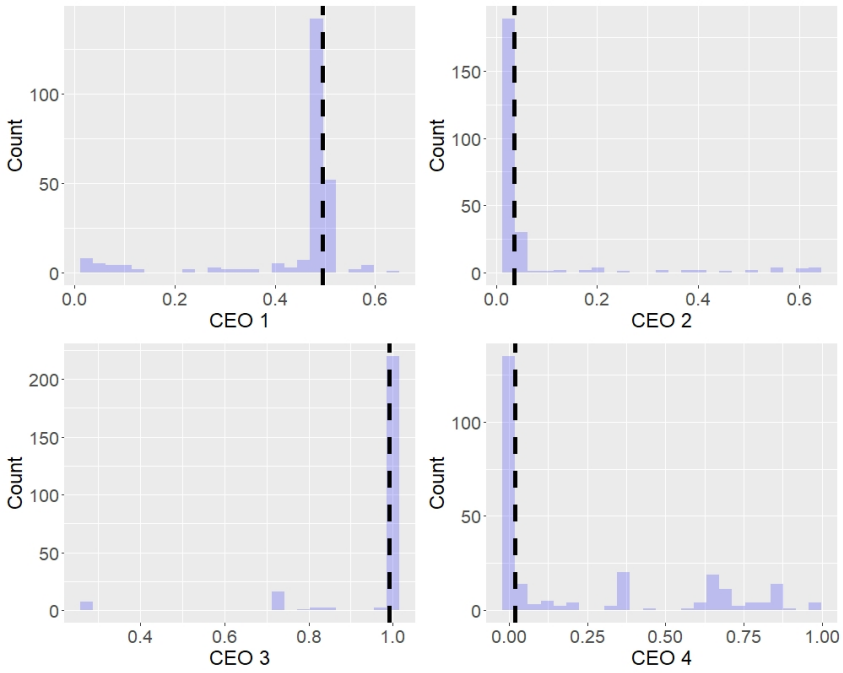
The problem with the labelling becomes clear when we look at Figure 4a. In this figure, we show four histograms. Each histogram corresponds to a different CEO in our CEO diary data set and we show the 250 estimates of the CEO behavior index $\hat{\theta}$. Looking at CEO 3 (this is the histogram in the bottom left), we observe that the histogram of the estimated CEO behaviour index is roughly symmetric: about half of the estimates are equal to zero, whereas the other half of the estimates is equal to one. Besides these estimates, we indicate by the dashed line the value of the CEO behavior index as used in the paper of Bandiera et al. (2020). We see that half of the estimates are in accordance with the value as used by Bandiera et al. (2020), whereas the other half of the estimates are equal to roughly one minus the value of the estimate. This indicates that the labelling of the estimates is random in each run of the algorithm. A similar pattern is observed for the other CEOs in Figure 4a, although maybe less clear for CEO 1 as there almost all estimates are centered around 0.5 instead of around zero and one.

For a correct interpretation of the regression estimates of the 250 runs we would like to have a uniform labeling, such that we can interpret the CEO behavior index in the same way for each run of the algorithm. To obtain a uniform labeling, we follow the following procedure:

1. Denote the vector containing the CEO behavior index of behavior β^1 as used by Bandiera et al. (2020) in their regression equation as $\mathbf{x_{b1}}$ (this corresponds to values that are shown by the dashed lines in



(a) Random labeling of the CEO behavior index.



(b) Uniform labeling of the CEO behavior index.

Figure 4: For 250 runs of the LDA algorithm, the CEO behavior index is determined. Histograms of the value of the CEO behavior index for four different CEOs in the CEO diary data set are shown. The dashed lines indicate the value of the baseline CEO behavior index for each CEO. Figure (a) shows histograms of the CEO behavior index using random labeling; figure (b) shows histograms of the CEO behavior index using uniform labeling.

Figure 4a). We can now denote the CEO behavior index of behavior β^0 as $\mathbf{x_{b0}} = \mathbf{1} - \mathbf{x_{b1}}$.

2. Run the LDA algorithm 250 times. Denote the CEO behavior index of run $i \in \{1, 2, ..., 250\}$ by $\mathbf{x_i}$.

3. For each run $i \in \{1, 2, ..., 250\}$ calculate both $d(\mathbf{x_{b0}}, \mathbf{x_i})$ and $d(\mathbf{x_{b1}}, \mathbf{x_i})$, where d denotes the Euclidean distance.

4. Now for each run $i \in \{1, 2, ..., 250\}$, if

   - $d(\mathbf{x_{b0}}, \mathbf{x_i}) > \mathbf{d}(\mathbf{x_{b1}}, \mathbf{x_i})$ the labeling in run i is the same as the one used by Bandiera et al. (2020), thus we do not have to change anything about the vector $\mathbf{x_i}$,
   - $d(\mathbf{x_{b0}}, \mathbf{x_i}) < \mathbf{d}(\mathbf{x_{b1}}, \mathbf{x_i})$ the labeling in run i is the opposed to the one used by Bandiera et al. (2020), thus we do have to correct for the wrong labeling by taking $\mathbf{x_i} \leftarrow \mathbf{1} - \mathbf{x_i}$.

For each run of the LDA algorithm we thus compare the distance between the CEO behavior index obtained in that run to the CEO behavior index as used in the paper by Bandiera et al. (2020). Choosing the index in each run to correspond to the labeling in the paper will result in a labeling that is uniform across all runs of the algorithm.

In Figure 4b, the result of the relabeling is shown. We observe that, where we previously had symmetric histograms, we now have histograms in which there is one big peak. This indicates that the labeling is now uniform across all runs. However, for CEO 4 we observe that there are quite some CEO behavior index estimates that are not in line with the big peak. The LDA algorithm thus not always produces a clear distinction between the two behaviors, and this is an indication that there is instability in the LDA algorithm when applied to CEO 4. Furthermore, for CEO 1 we observe that almost all estimates are centered around 0.5, which is also the value used by Bandiera et al. (2020). Here we thus observe that CEO 1 is almost a perfect mix between behavior β^0 and β^1 and that some CEOs thus not belong to one of the two groups of CEOs but are really a mixture. Recall that the main assumption of the LDA model is that a CEO can be seen as a mixture between two pure behaviors. This is something we clearly observe for all CEOs in Figure 4b, which indicates that LDA is an appropriate clustering technique for this data set.

Figure 4 shows us that if we apply LDA with only two topics, we can set up a procedure that ensures us that we use a uniform topic labeling in each run of the LDA algorithm. Now assume that we would have chosen $K = 3$ topics in our model. Doing this, would result in $K! = 3! = 6$ different ways to label each of topics. In terms of the histogram, we would then expect there to be six peaks in the histogram before relabeling. Correcting for the random labeling then becomes difficult, as the procedure that we described before can not be generalised for $K > 2$. So we expect that when $K > 2$, the random labeling of LDA becomes a huge issue when analysing multiple runs of the algorithm. However, this is not studied in this Thesis.

4.5 Bootstrapping Statistics

As described in Section 4.3 we will run the baseline LDA model for 250 runs, which will result in 250 estimates of the regression coefficient $\hat{\alpha}$. For each of the regression estimates we would like to determine if the regression estimate is statistically different from zero or not. We check if they are different from zero, as a statistically significant regression coefficient can then be interpreted as CEO behavior having an effect on firm performance. We run our regression in R (see section 4.2) and R produces information about whether or not a regression estimate is significant. However, the output produced by R on the significance cannot be used as. This is because in our regression equation (Equation 2), we use the estimated CEO behavior index $\hat{\theta}$ which is not fixed. As this estimate is random in itself we need to use another measure to determine the statistical significance. So instead, we opt for using a bootstrapping procedure to determine the critical value of $\hat{\alpha}$ for which our regression estimates are statistically significant from 0.

The standard **residual bootstrap** algorithm is described by Cameron and Trivedi (2005). The algorithm consists of the following steps:

1. For the regression model $y_i = \alpha\hat{\theta}_i + X_i\beta + \varepsilon_i$ form the fitted residuals $\hat{\varepsilon}_i$.
2. Bootstrap from the residuals $\{\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_S\}$ new residuals $\{\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, ..., \tilde{\varepsilon}_S\}$.
3. Construct new data (under the null hypothesis that $\alpha = 0$): $\tilde{y}_i = X_i\beta + \tilde{\varepsilon}_i$.
4. Run a regression of $\tilde{y}_i$ on $\hat{\theta}_i$ and X_i to obtain a bootstrap estimate $\hat{\alpha}_B$.
5. Repeat steps 2, 3 and 4 M times.
6. Calculate bootstrap statistics on $\hat{\alpha}$.

This standard residual bootstrapping algorithm, however, cannot be applied to the case of Bandiera et al. (2020). This is because for that specific case $\hat{\theta}$ is not fixed but random as well. We thus opt for using a **complete bootstrap** algorithm that changes step 2 and 4 in the residual bootstrap algorithm. Compared to the residual bootstrap algorithm, this complete bootstrap algorithm not only samples new residuals in each step, but also samples a new CEO behavior index in each run of the bootstrap. The algorithm is the following:

2. Bootstrap from the residuals $\{\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_S\}$ new residuals $\{\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, ..., \tilde{\varepsilon}_S\}$ and from the behavior indices $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N\}$ new behavior indices $\{\tilde{\theta}_1, \tilde{\theta}_2, ..., \tilde{\theta}_N\}$.
4. Run a regression of $\tilde{y}_i$ on $\tilde{\theta}_i$ and X_i to obtain a bootstrap estimate $\hat{\alpha}_B$.

With our bootstrap analysis we would like determine two things. The first thing we want to determine is the p-value corresponding to the baseline regression estimate of $\hat{\alpha} = 0.343$. Running the complete bootstrap analysis for M times gives us M estimates of $\hat{\alpha}$. Counting how many of these estimates are smaller than -0.343 or larger than 0.343 gives us the p-value corresponding to $\hat{\alpha} = 0.343$. This p-value should be in line with the p-value as reported by running the baseline regression in R. The second thing we want to determine using the bootstrapping procedure is a critical region for which we reject the null hypothesis that $\alpha = 0$. The bootstrapping procedure yields a standard normally distributed sample, thus a 95% confidence interval can be found by determining the values of the bootstrapping sample corresponding to the 2.5 percentile and 97.5 percentile. This rejection region can be used to determine which of the 250 runs of the LDA algorithm are statistically significant and which are not.

4.6 Measure of Fit

The different runs of the LDA algorithm will be compared to each other in terms of the instability as defined in Equation 3. This measure yields information about the total instability of all the runs, but it does not shed light on the accuracy of each run of the LDA model in itself. The problem is that for the LDA inference problem, there are multiple local maxima (Griffiths & Steyvers, 2004). Each of these maxima yields a proper solution to the maximum-likelihood problem, but not each of these solutions is satisfactory for the topics modeling problem (Koltcov et al., 2014). In other words, if the LDA algorithm finds a solution, the solution in itself is not per se interesting to a researcher. We thus would like to include a measure of model fit in our analysis, such that we can compare the performance of different runs of the LDA algorithm to each other.

Within topic modeling, one well known measure of fit is perplexity. According to Koltcov et al. (2014), perplexity shows how well a model predicts new test samples. Perplexity is defined (Griffiths & Steyvers, 2004) as

$$\text{perplexity} = e^{\frac{-\ln(P(\mathbf{w}_{\text{test}})|\phi)}{n_{\text{test}}}}, \tag{4}$$

where ϕ is a Dirichlet prior placed on β (the per-topic word distribution), $\mathbf{w}_{\text{test}}$ are the words in our test set and $|\mathbf{w}_{\text{test}}| = \mathbf{n}_{\text{test}}$. Giving the definition, we observe that the perplexity indicates the uncertainty in predicting a single word: the smaller this uncertainty is, the better the model predicts the new test sample. A researcher thus aims for finding a LDA model with a as low as possible perplexity.

The perplexity has been used by Bandiera et al. (2020) to determine which of five initial runs of the baseline LDA model will be used for further analysis (see Section 3.5). We will deepen this study by establishing the relationship between the perplexity and the estimates of the different runs in more detail and by relating this to the instability of the LDA results.

However, there are some caveats regarding the perplexity. As Zhao et al. (2015) denotes, the perplexity is not constant when the number of topics changes: if the number of topics grows, the value of the perplexity decreases (Koltcov et al., 2014). Comparing the perplexity across two models with two different topic sizes is thus not useful in itself. As in this Thesis we only look at a case in which there are two topics, this issue is not of importance. Besides, perplexity is increasing in the vocabulary size (De Waal & Barnard, 2008), but, again, this is not of importance to this Thesis, as vocabulary size remains constant.

4.7 Lasso Regression

Lasso regression is a form of linear regression that is useful when a data set contains many regressors. Lasso is an abbreviation for *least absolute shrinkage and selection operator* and it has been developed by Tibshirani (1996). The special feature of Lasso regression is that for a regression equation some of the coefficient estimates are shrunk towards zero (James, Witten, Hastie, & Tibshirani, 2013). By doing this, Lasso produces regression results that are easy to interpret. However, if two independent variables are strongly correlated, Lasso will pick just one of the two variables. As it is not known beforehand which of the two variables will be chosen, a zero coefficient estimate does not necessarily indicate that the variable has no effect on the dependent variable. Lasso is a method to perform both variable selection and shrinkage at the same time (Tibshirani, 2011). As Lasso regression selects a part of the variables from the complete set of regressors, it yields sparse models (James et al., 2013).

For the description of Lasso regression we consider a data set consisting of N observations. We denote the dependent variable by y and the set of p independent variables by $X = \{x_1, x_2, ..., x_p\}$, thus the ith observation of the data set is given by $\{y_i, x_{1i}, x_{2i}, ..., x_{pi}\}$. For Lasso regression all variables need to be scaled such that the mean is equal to zero and the variance is equal to one (Tibshirani, 2011). For the regression equation

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \varepsilon_i,$$

the Lasso coefficient estimates are given by (James et al., 2013):

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{5}$$

where $\lambda \geqslant 0$ is a tuning parameter, that has be determined. The tuning parameter λ can be determined using k-fold cross-validation (Emmert-Streib & Dehmer, 2019; James et al., 2013).

For the explanation of the k-fold cross-validation principle (this is based on the explanation in James et al., 2013) assume that λ is fixed. Divide the data set $\{y, X\}$ into k groups, called folds, of approximately equal size. Now assume that we look at fold $i \in \{1, 2, ..., k\}$. We treat fold i as the validation set, and

we fit the Lasso regression on the data set excluding the ith fold. The fitted model can be applied to the validation set to produce a test mean squared error MSE_i for the ith fold. The procedure is repeated for each $i \in \{1, 2, ..., k\}$, such that we obtain a set of test errors $\{\text{MSE}_1, \text{MSE}_2, ..., \text{MSE}_k\}$. The k-fold cross-validation estimate is than given by

$$CV_k(\lambda) = \frac{1}{k} \sum_{i=1}^{K} \text{MSE}_i.$$

The value of λ that is used to determine the Lasso coefficients (see Equation 5) is now given by (Emmert-Streib & Dehmer, 2019)

$$\hat{\lambda} = \underset{\lambda \in Y}{\text{argmin}} CV_k(\lambda),$$

where Y denotes the set of test values of λ.

The Lasso regression procedure can also be applied to the CEO diary data set. In the CEO diary data set, there is data about the activities a CEO undertakes in one week. Each activity in the data set is a string and this string consists of information about five features (see Section 3.2 for more detail about the CEO diary data set). These five features are categorical (e.g., a meeting is either with people from production/strategy/marketing/etc.). For the Lasso regression we turn all the unique feature variables into binary variables (e.g., production is equal to 1 if a meeting was with people from production and it is equal to 0 if this is not the case) such that a Lasso regression can be run on the CEO diary data set with the features as regressors.

As we have seen in Section 3, Bandiera et al. (2020) conduct a two-step analysis in which they first form a CEO behavior index using LDA and then run the regression as specified in Equation 2 using this CEO behavior index. Using Lasso regression, we can perform a similar analysis in only one step. To do this we denote by X our regressor data set that not only consists of the 63 features in our CEO diary data set, but also of the 94 control variables that are mentioned in the regression equation. By applying Equation 5, where λ is determined using the k-fold cross-validation procedure, we can determine the Lasso regression coefficients.

The output of the Lasso regression will be compared to the output of the LDA algorithm in two ways. First, we compare the interpretation of the Lasso results to the interpretation of the LDA results. We have already seen in Section 3 that the LDA results on the CEO diary data set indicate that CEOs can be characterised as leaders and managers. As Lasso regression is a way to conduct variable selection, the Lasso regression estimates are an indication which variables are of importance to the firm performance. By comparing the Lasso coefficients to the definition of the leader and manager CEOs as given by Bandiera et al. (2020), we can determine if there is agreement in interpretation between both methods.

A second way to compare the Lasso regression output to the results produced by the LDA model is to determine the correlation between the CEO behavior index implied by both methods. As said before we

denote by X the set of regressors that we use for our Lasso regression. We can write $X = \{X_1, X_2\}$, where X_1 are the features from the CEO diary data set and X_2 are the control variables. Similarly, we can write $\hat{\beta}_{\text{Lasso}} = \{\hat{\beta}_1, \hat{\beta}_2\}$, where $\hat{\beta}_1$ are the coefficient estimates corresponding to X_1 (the features of the CEO diary data set) and $\hat{\beta}_2$ are the coefficient estimates corresponding to X_2 (the control variables). Now $X_1\hat{\beta}_1$ gives a vector containing information about the CEO classification according to Lasso. If we denote the CEO behavior index computed with LDA by $\hat{\theta}$, the correlation between $X_1\hat{\beta}_1$ and $\hat{\theta}$ is an indication of how comparable the two methods are. If there is much agreement between the two methods, then this correlation should be high.

5 Results

5.1 Introduction

This section will form the main part of this Thesis. We will first determine the stability of LDA on the CEO diary data set and then we will compare the performance of LDA on this data set to the performance of a Lasso regression on the same data set. More specifically, this section will consist of three parts. In Section 5.2, we will run the baseline LDA model as described in Section 3.3 of this Thesis 250 times. The 250 different regression estimates that are produced by these 250 runs of the LDA algorithm will be analysed. In Section 5.3, we will run two robustness checks on the baseline model. We will see how changing the vocabulary size and the hyperparameters will (1) affect the main results as found for the baseline model and (2) how this influences the stability of LDA. In their work, Bandiera et al. (2020) do run some robustness checks on the vocabulary size, but not on the hyperparameters. However, for the vocabulary size they do not take into account the stability which may have consequences for their findings. In Section 5.4, we will compare the performance of a Lasso regression to the performance of the LDA model on the CEO diary data set.

5.2 Running the Baseline Model 250 Times

We will look at the instability of the LDA algorithm when we run the baseline LDA algorithm 250 times to the CEO diary data set. Our analysis will consist of four steps after which we will conclude what we learn from this analysis.

Step 1: First Look at Instability

The first step in the analysis of the 250 LDA runs, is to determine for each run $i \in \{1, 2, ..., 250\}$ of the algorithm the estimated regression coefficient $\hat{\alpha}_i$ (see Equation 2 for the regression equation). After this we compute the instability over all the estimated regression coefficients using Equation 3.

Figure 5 shows a histogram of regression estimates of the 250 runs of the LDA algorithm. In the histogram, the dashed line denotes the value corresponding to the baseline model regression estimate. We observe that the majority of the runs of the algorithm yield a regression coefficient centered around the baseline regression estimate. However, we observe that there is quite some dispersion in the values of the regression coefficient, which is an indication of instability in the results. What is more interesting is that some of the regression coefficients are negative. A negative regression coefficient would yield a result contradicting the one found by Bandiera et al. (2020): namely that manager CEOs make more firm profit than leader CEOs. It is possible that these negative regression coefficients are just caused by a "bad" LDA model. We will check for this when studying the perplexity in step 4 of this analysis. Besides, we also observe that there are regression coefficients that are almost equal to zero, which would indicate that there is no difference in firm profits between leader CEOs and manager CEOs.

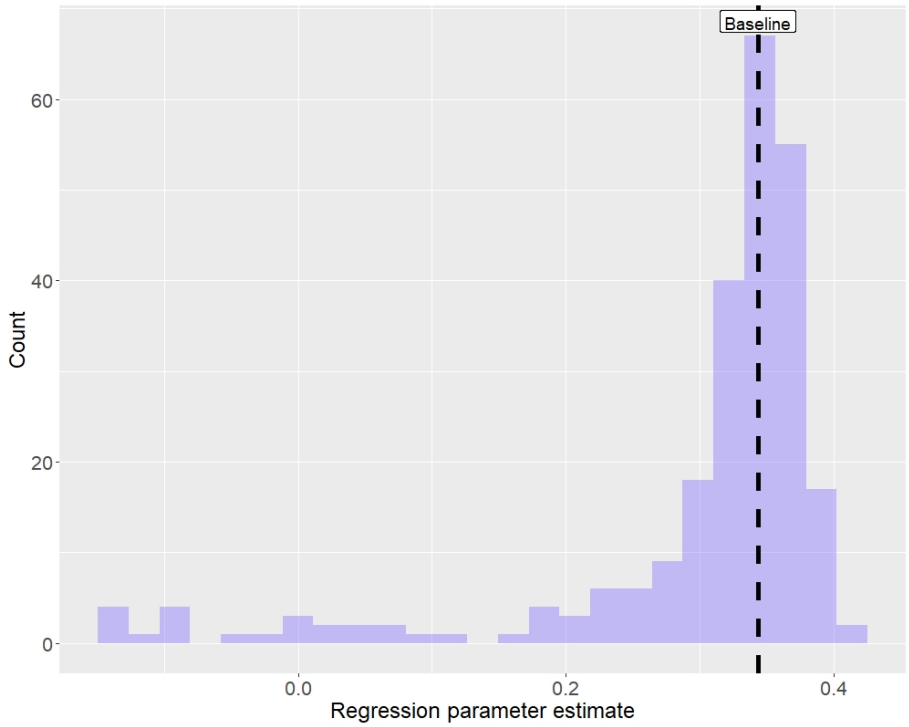


Figure 5: For 250 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined. A histogram of the 250 parameter estimates is shown. The dashed line shows the baseline regression estimate.

Based on the histogram in Figure 5 we observe that there is quite some instability in the regression estimates. The instability has also been calculated using Equation 3 and found to be equal to 0.113 (see the first column of Table 4 for all summary statistics). Given that the range of regression coefficient estimates is from -0.143 to 0.408, an instability of 0.113 can be seen as high as it covers roughly 25% of this range. We also denote that the mean value of 0.301 over all the 250 runs of the model is slightly lower than the value 0.343 of the baseline model. The 250 runs produce a mean lower than the baseline model due to the small and negative regression estimates.

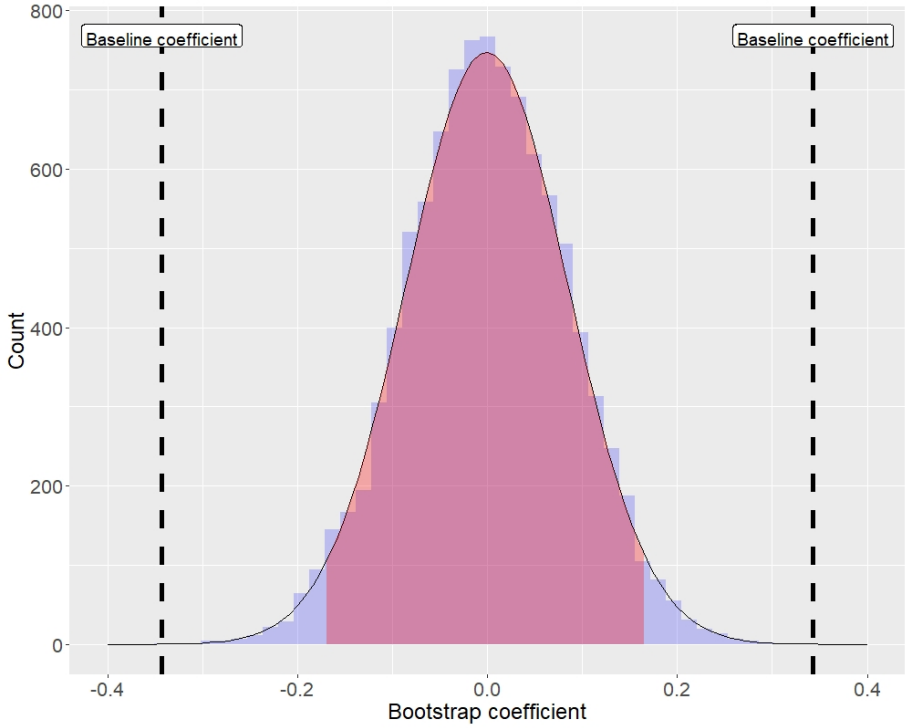
Step 2: Bootstrapping Analysis

In step 1 of the analysis we have seen both graphically (see Figure 5) and numerically (see Table 4) that there is quite some instability in the results of the baseline LDA model. This instability was mostly caused by very small and even negative regression estimates. From an economical point of vies those estimates do not make sense, as they are conflicting the interpretation of the results of the baseline model. It may be the case that the estimates that cause the results to be unstable are statistically insignificant and thus that the existence of these estimates is not contradicting the main results of Bandiera et al. (2020). So as a second step in our analysis we want to determine a rejection region for $\hat{\alpha}$ that tells us which regression estimates are statistically different from zero and which are not. To find the rejection region we apply both the residual and complete bootstrapping procedure as explained in section 4.5.

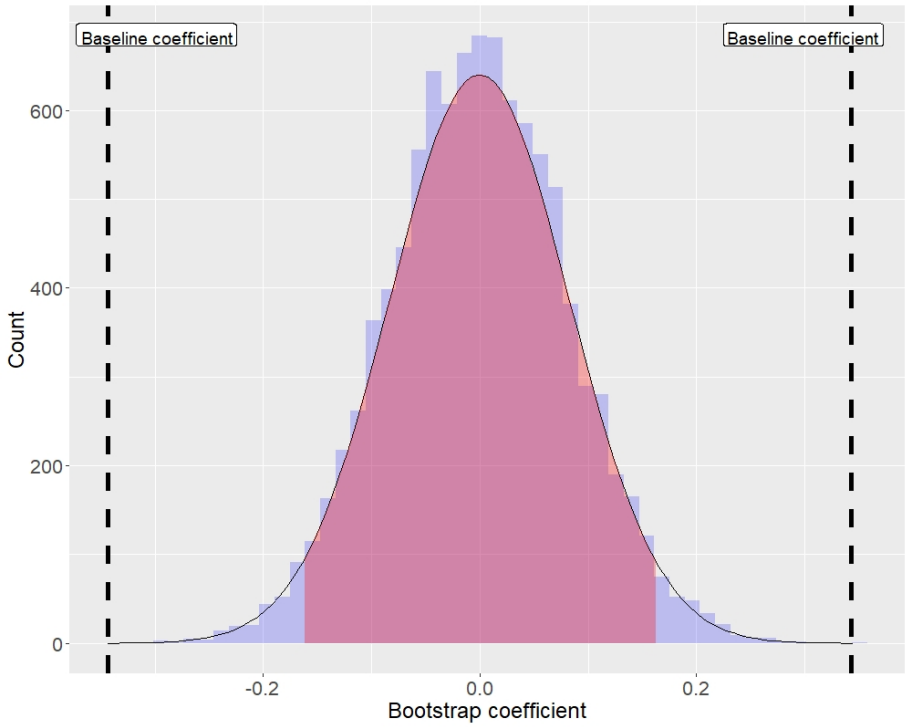
The residual bootstrap analysis can be found in Figure 6a. This figure shows a histogram with the results of 10.000 bootstrap steps. The dashed lines correspond to the value found in the baseline model. The red area corresponds to the 95% confidence interval. We find a 95% confidence interval of $(-0.169, 0.166)$. The values of -0.169 and 0.166 are our critical values: if a estimate is not in the interval $(-0.169, 0.166)$ it is statistically significant, and thus not statistically different from zero. Observing the histogram we denote that the bootstrap results follow a normal distribution closely. This is interesting as bootstrapping is often applied to estimates for which the underlying distribution is not a normal distribution.

For the complete bootstrap analysis, shown in Figure 6b, we find a 95% confidence interval of $(-0.162, 0.163)$. Also for the complete bootstrap analysis we find that the bootstrap results closely follow a normal distribution. The complete bootstrap confidence interval is roughly similar to the one found for the residual bootstrap, but it is more symmetrically centered around zero. This is due to the fact that in the complete bootstrap case we also take the randomness in the CEO behavior index $\hat{\theta}$ into account.

For both the residual and complete bootstrap procedure we find a p-value equal to $p = 0.0001$, thus for both bootstrapping procedures only 1 of the 10.000 bootstrap estimates was found to be larger than 0.343. This p-value is roughly equal to the p-value found for the baseline model in OLS, which was equal to 0.000634.



(a) Residual bootstrapping.



(b) Complete bootstrapping.

Figure 6: For 250 runs of the LDA algorithm, a bootstrapping procedure is applied with 10.000 runs to the 250 coefficient estimates $\hat{\alpha}$. Histogram (a) shows the residual bootstrap results; histogram (b) shows the complete bootstrap results. The dashed lines indicate the baseline model coefficient estimate.

For the subsequent steps in the analysis we will make use of the confidence interval found by the complete bootstrap analysis and not of the interval produced by the residual bootstrap analysis. We opt for using the complete bootstrap results as this better takes into account the randomness in the CEO behavior index.

Step 3: Significance Results

The next step in our analysis is to implement the bootstrap results that we found in the previous step of the analysis. We found a 95% confidence interval equal to $(-0.162, 0.163)$, so the coefficient estimates that lie in this interval are not statistically different from zero. In Figure 7 we observe a similar histogram as in Figure 5. Compared to the previous histogram, the new histogram in Figure 7 also makes clear which estimates are statistically significant (these estimates are shown in blue) and which estimates are not statistically significant (these estimates are shown in red). As before, the dashed line indicates the value of the baseline regression model. Now the dotted lines show the border of the rejection region.

In Figure 7 we observe that the majority of the regression estimates is statistically significant. In column 2

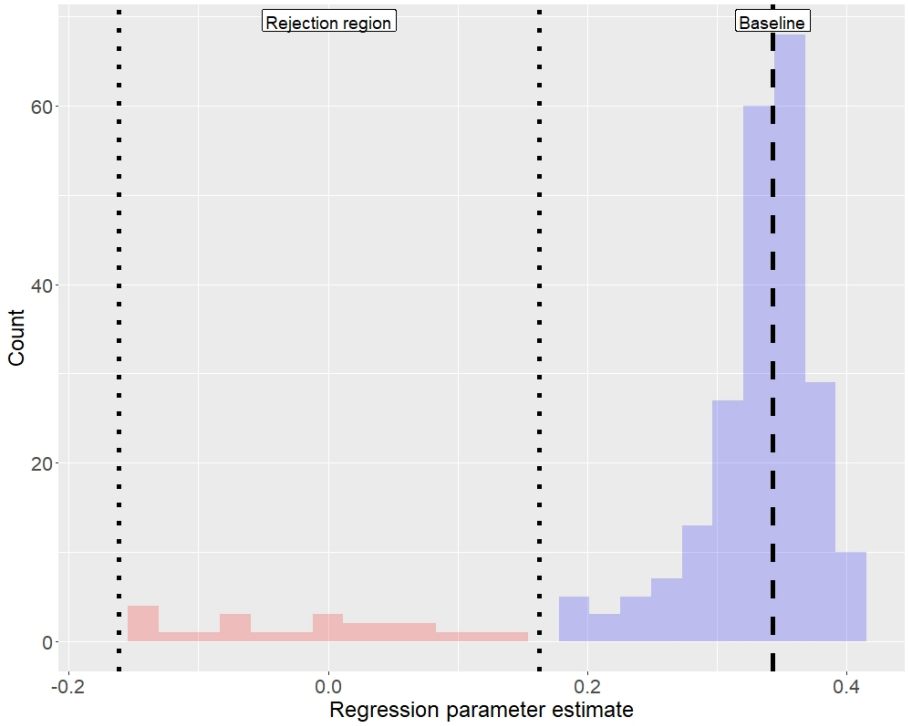


Figure 7: For 250 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined. A histogram of the 250 parameter estimates is shown. The dashed line shows the baseline regression estimate. The dotted lines show the border of the rejection region determined by a complete bootstrap procedure. The estimates in blue are statistically significant, whereas those in red are not statistically significant.

of Table 4 we find that 227 out of the 250 regression estimates are statistically significant. The estimates that are statistically insignificant, are low in size and even negative. The interpretation of the complete set of statistically significant values is thus is in agreement with the results found by Bandiera et al. (2020).

If we only look at the set of regression estimates that are statistically significant, we observe that the instability drops from 0.113 (this is the instability over all the 250 regression estimates) to 0.043 (see column 2 in Table 4 for all summary statistics). The set of statistically significant estimates is thus more stable. For this set we find a mean value of 0.334. This value is more in line with the baseline model estimate of 0.343, compared to the value of 0.301 that was found for the complete set of 250 regression estimates.

We have seen that the set of statistical significant regression estimates is stable. However, if we want to obtain a set of regression estimates that is stable making a selection based on the statistical significance is not a good idea. Only considering the estimates that are statistically significant can be considered "cherry picking", as we do not know beforehand if regression results are significant or not.

Step 4: Perplexity Results

In step 4 of our analysis we would like to take the perplexity of the 250 LDA models into account. In Figure 5 we have seen that some of the regression estimates provide little economic intuition: they are close to zero or even negative. In step 3 of this analysis we have seen that these regression estimates can be removed from the sample of estimates based on the statistical significance. However, this is some form of cherry picking: we disregard all those values that are not interesting to us. This is bad practise as we do not check the quality of the underlying LDA model. It is better to remove values based on the accuracy of the underlying LDA model then to just remove those values from the sample that are not in line with what we want to show. As the perplexity can be seen as a measure of how well a model fits data, we are going to compare the perplexity to the values of the regression estimates. Remember hereby from Section 4.6 that models with a low perplexity are preferred to models with a high perplexity.

Figure 8 shows a scatter plot with the regression parameter estimates on the horizontal axis and the perplexity on the vertical axis. We observe that for the 250 runs of the algorithm there seems to be a cluster of runs that are centered around the baseline parameter estimate. This cluster of runs has a low perplexity value compared to the other runs of the algorithm. We also observe that the negative and small regression estimates all have a high perplexity. This indicates that the "off" estimates as observed in the histogram in Figure 5 are all caused by "bad" LDA models that will never be chosen when the perplexity is checked first. In the scatter plot we observe that there is also some connection between the statistically significance and the perplexity. The statistically insignificant values all have a relatively high perplexity. However, there are also multiple runs of the algorithm that have a high perplexity, but are still statistically significant.

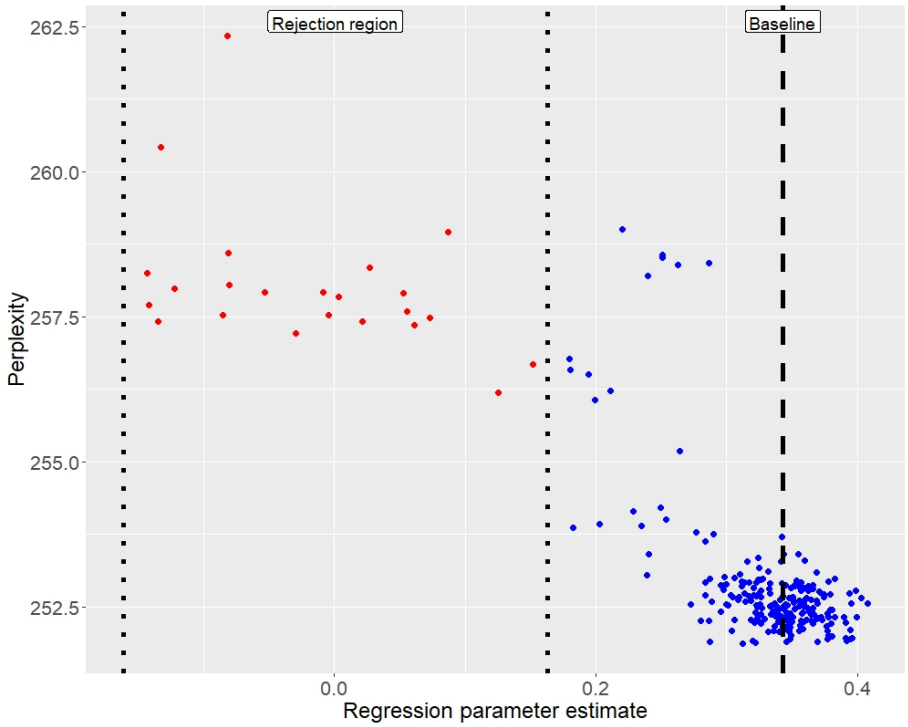


Figure 8: For 250 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined. A scatter plot of the 250 parameter estimates is shown with the regression estimates on the horizontal axis and the value of the perplexity on the vertical axis. The dashed line shows the baseline regression estimate. The dotted lines show the border of the rejection region determined by a complete bootstrap procedure. The estimates in blue are statistically significant, whereas those in red are not statistically significant.

The scatter plot in Figure 8 shows us something interesting: if we choose a critical value for the perplexity we can ensure that we get a sample that is nicely centered around the baseline regression estimate (i.e., the cluster of values that is in the bottom right of the scatter plot). What deviates this approach from the approach in which we look at the statistically significance, is that with the perplexity approach we take into account the performance of the LDA model. With the statistical significance approach we only try to find a set of estimates that is as stable as possible, where we do not control for the model performance. So opposed to finding only a set of estimates that is stable, we also want this set to be the result of good LDA models. This can be achieved by taking into account the perplexity.

To implement the perplexity in our analysis, we study two different methods of choosing a critical value for the perplexity. In the first method, we are interested in what happens if we obtain a sample that is equal in size to the number of statistically significant values that we have. Having equal sample sizes across two methods ensures that we can fairly compare the two samples in terms of stability. A perplexity of 257.421 corresponds to the critical perplexity in this first case, which we will denote as perplexity 1 in Table 4 that shows the summary statistics of this sample.

In Figure 9a we show graphically what happens for the case of perplexity 1. The figure again shows a scatter plot with the regression estimate on the horizontal axis and the value of the perplexity on the vertical axis. The vertical dashed line indicates the baseline regression estimate, whereas the horizontal vertical line denotes the critical perplexity. We observe that by choosing this value for the critical perplexity, most of the negative and small regression parameters are removed from the sample. However, this is not the case for all of those observations. Besides, we observe that some observations that were previously considered to be statistically significant are now removed from the sample.

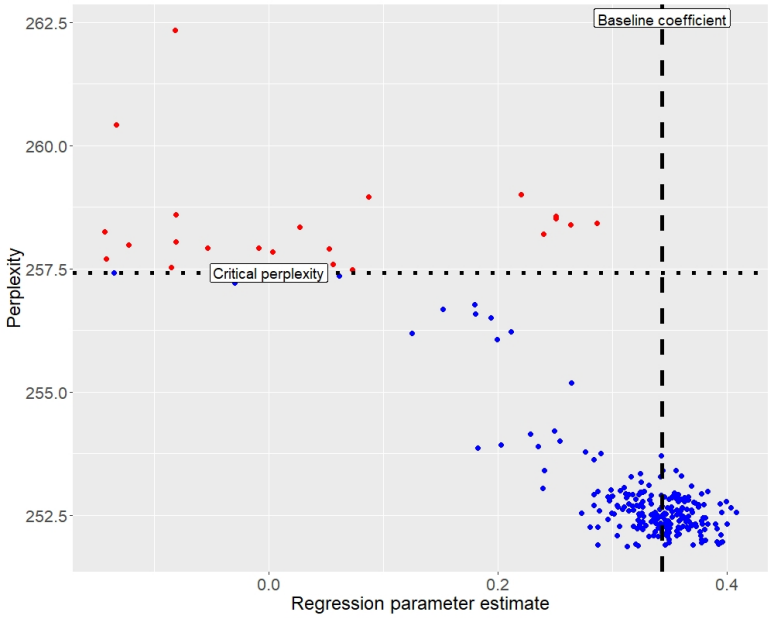
In Table 3 we show a comparison between the set of estimates in the perplexity 1 case and in the statistical significant case. We observe that for 238 of 250 runs of the algorithm there is agreement between both methods. However, there are also 6 coefficient estimates that are accepted based on the significance, but rejected based on the perplexity 1. Furthermore, there are 6 coefficient estimates that are accepted based on perplexity 1, but rejected based on the significance.

We have seen that the set of estimates that are formed by the perplexity 1 case is in close agreement with the set of estimates formed by the statistical significance case. However, comparing the summary statistics of both cases (see column 3 and 4 of Table 4), we observe that the instability is smaller for the significant value case compared to the perplexity 1 case. This is due to the fact that there are still some very small values of the regression estimate accepted based under perplexity 1. So this way of choosing a critical perplexity is not appropriate for removing the instability in the LDA results.

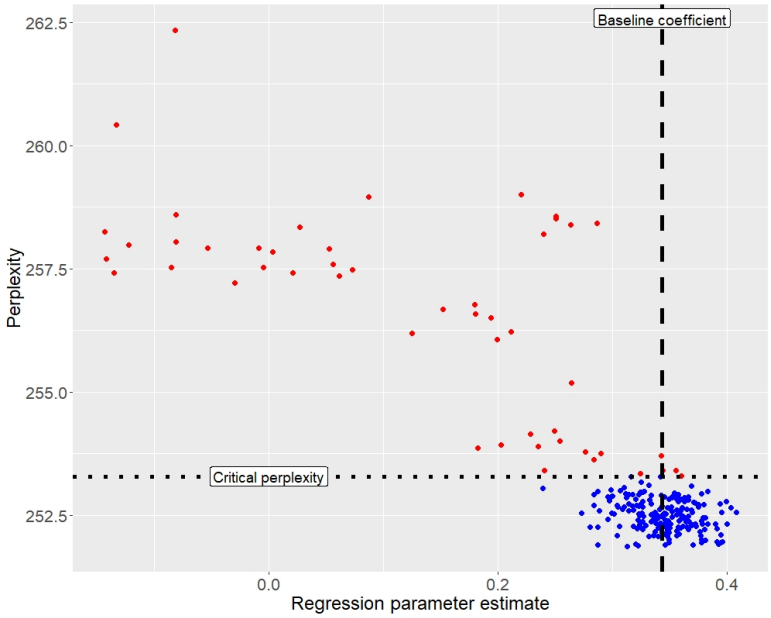
Table 3: A comparison is made between the case in which we select estimates based on the significance and the case in which we select estimates based on the perplexity. The top table shows the perplexity 1 case, and the bottom table the perplexity 2 case. The values indicate how many of the 250 runs of the algorithm are in accordance between both cases.

	Reject significance	Accept significance
Reject perplexity 1	17	6
Accept perplexity 1	6	221

	Reject significance	Accept significance
Reject perplexity 2	23	27
Accept perplexity 2	0	200



(a) Perplexity 1.



(b) Perplexity 2.

Figure 9: For 250 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined. Two scatter plots of the 250 parameter estimates are shown with the regression estimates on the horizontal axis and the value of the perplexity on the vertical axis. The dashed line shows the baseline regression estimate. The dotted lines show the critical perplexity. In scatter plot (a) the critical perplexity is chosen in such a way that the number of observations is equal to the number of statistically significant observations (this is the perplexity 1 case); in scatter plot (b) the critical perplexity is chosen to best mimic the cluster of estimates (this is perplexity 2 case). The estimates in blue are accepted based on the critical perplexity, whereas those in red are rejected.

A second way of choosing a value for the critical perplexity is to choice a value of the perplexity to get visually as close as possible to the cluster of regression estimates that we observe in the scatter plot in Figure 8. We refer to this way as the perplexity 2 case. By doing this, we get a smaller set of regression estimates. Obviously, this set of values has a small instability, as all those estimates belong to the same cluster. Besides, this choice of the critical perplexity ensures that the small and negative regression estimates are not part of the set of estimates. The perplexity value that corresponds to this case is 253.287, which is lower than the critical perplexity of 257.421 we had in the perplexity 1 case.

In Figure 9b we implement the case of perplexity 2. The scatter plot shows again the regression parameter estimates on the horizontal axis and the perplexity on the vertical axis. We observe that the critical perplexity (this is the horizontal dotted line) now creates a sample of estimates that is in much agreement with the cluster of regression estimates in the bottom right of the plot.

In Table 3 we show a comparison between the set of estimates in the perplexity 2 case and in the statistical significant case. There is agreement in both methods for 223 out of 250 runs. This is a lower number of estimates that are in agreement than the 238 estimates for the perplexity 1 case. So, although perplexity 2 better represents the cluster of coefficient estimates, there is a higher cost in falsely rejecting values that are statistically significant.

Looking at the summary statistics in Table 4, we see that with perplexity 2 we find the smallest instability

Table 4: Summary statistics (minimum value, mean value, maximum value, the instability, and the number of observations) of four different groups of regression coefficient estimates. In the first column the statistics are shown over all the 250 runs of the algorithm. In the second column the statistics are shown for the set of statistically significant coefficient estimates. In the third column the statistics are shown for those observations that have the lowest perplexity, where we ensure that the sample size is equal to that of the set of statistically significant coefficient estimates (this is the perplexity 1 case). In the fourth column the statistics are shown for the 80% observations with the lowest perplexity (this is the perplexity 2 case).

	All values	Significant values	Perplexity 1	Perplexity 2
Min	−0.143	0.180	−0.135	0.239
Mean	0.301	0.334	0.328	0.344
Max	0.408	0.408	0.408	0.408
Instability	0.113	0.043	0.066	0.028
N	250	227	227	200

over all cases considered in this analysis. However, the sample of estimates is also the smallest. Furthermore, the mean value of 0.344 over all these estimates is higher than the value reported in the baseline model. The mean value for this case is also higher than the mean value of the set of statistically significant estimates.

Conclusion

From this section, we can conclude three things. The first thing is that LDA in itself in this case is not very stable. There is quite some dispersion in the regression estimates and some of the regression estimates do not have a logical economic intuition. The second thing learned is that the instability is almost completely resolved for if we do correct for statistical significance. However, the statistical significance cannot be used as a rejection criterion to form a new sample as doing this can be considered cherry picking. The third and most interesting result from this analysis is that there is a connection between the coefficient estimates and the perplexity. We have shown that the perplexity can act like a "rejection tool": runs of the LDA algorithm that have a higher perplexity than a post-determined perplexity are rejected, whereas other runs are accepted. By doing this we end up with good models that can be interpreted in economical way.

5.3 Stability of the Robustness Checks

5.3.1 Different Cut-Off Values

Introduction

In Section 3.2. we have described the CEO diary data set in detail. This data set consists of $4,253$ unique activities (this is the vocabulary size). However, not all of these activities are informative to a researcher. As Bandiera et al. (2020) denote: "*LDA identifies pure behaviors by finding patterns of co-occurrence among activities across CEOs, so infrequently occurring activities are not informative.*" For this reason Bandiera et al. (2020) drop all the activities in the data set that are found in fewer than 30 activities. This drop of activities reduces the number of unique activities in the data set to 654. In their paper, Bandiera et al. (2020) do not go into much detail about why this cut-off value of 30 is chosen. They do run some robustness checks in their paper by alternatively choosing a cut-off value of 15 and 45. However, also about these results they do not get into much detail. They only mention that they "*find little effect on the main results*".

Regarding the cut-off value of the data set we are interested in two things: do the results from the baseline model change if an another cut-off value is chosen, and does the cut-off value affect the stability of the LDA results? To study these two things, we will create LDA samples with five different cut-off values:

$\{0, 15, 30, 45, 425\}$. We study the three cut-off values that have also been studied by Bandiera et al. (2020) and add to that a case in which none of the activities is removed from the data set (this corresponds to a cut-off value of 0) and a case in which a lot of activities are removed from the data set (this corresponds to the cut-off value of 425).

Table 5 provides us with some insights in the number of activities in each of the data sets created by the different cut-off scenarios. We observe that if we increase the cut-off value, the number of unique activities as a percentage of the total activities drops quickly. For a cut-off value of 0 we find that the number of unique activities accounts for approximately 3.3% of the total number of activities, whereas this number is only equal to 0.1% for a cut-off value of 425: a higher cut-off value thus results in more mundane activities. Also, the number of unique activities as a percentage of the total vocabulary size of the complete CEO diary data set (this size is $4,254$) drops quickly. Already for a cut-off value of 15 this percentage is equal to approximately 28.3% which indicates that 71.7% of the activities (these are the more specialist activities) are removed.

The analysis of the five different cut-off scenarios, follows the previous section closely. However, we will go into less detail in this section (and even skip step 1 from the analysis) compared to the previous section, as this section is a repetition of the previous section with different numbers. We will only cover the main steps in the analysis that do provide interesting results. For each cut-off value we run the LDA algorithm 100 times. This number is smaller compared to the one in the previous section as here we need to run five different LDA specifications which is computationally expensive.

Table 5: For each cut-off value in the set $\{0, 15, 30, 45, 425\}$ the number of unique activities and the total number of activities is shown. Furthermore, it is shown for each cut-off value how big the proportion of unique activities is compared to the total number of activities and how big the proportion of unique activities is compared to the total vocabulary size of of the CEO diary data set (this size is equal to $4,253$).

Cut-off value	Unique activities	Total activities	Unique as a % of total	Unique as a % of vocabulary
0	$4,253$	$127,660$	3.332	100.000
15	$1,204$	$110,039$	1.094	28.309
30	654	$98,347$	0.665	15.377
45	459	$91,134$	0.504	10.792
425	46	$42,763$	0.108	1.082

Step 2: Bootstrapping Analysis

Table 6 provides us with the bootstrap results of the five different cut-off scenarios. We observe that the p-value is small for all cut-off values for both the residual and the complete bootstrap case. However, as the p-value reported for the standard baseline OLS regression was also small (around 0.0006) this is not surprising. Comparing the residual and complete bootstrap confidence intervals to each other shows us that the residual bootstrap confidence interval are more right-skewed, whereas the complete bootstrap confidence interval are more left-skewed.

Just like in the previous section, we will use the 95% confidence intervals produced by the complete bootstrap analysis in the subsequent steps of this analysis. We observe that as the the cut-off value is increasing, the confidence interval seems to get wider. However, there is a turning point (in this case at the cut-off value of 30) after which the confidence interval gets more narrow again. The cut-off values thus seems to have an impact on the critical values of the different cut-off cases. In Appendix B the histograms corresponding to the 10.000 runs of both bootstrapping procedures can be found and we observe that they are in close alignment with the normal distribution.

Table 6: For each cut-off value in the set $\{0, 15, 30, 45, 425\}$ the bootstrap results are presented. For both the residual and complete bootstrap procedure (10.000 runs) the 95% confidence interval and the p-value are shown.

Cut-off value	Residual 95% CI	Residual p-value	Complete 95% CI	Complete p-value
0	(-0.169, 0.164)	0.0001	(-0.145, 0.147)	0.0000
15	(-0.164, 0.162)	0.0001	(-0.157, 0.158)	0.0000
30	(-0.169, 0.164)	0.0000	(-0.162, 0.164)	0.0000
45	(-0.167, 0.162)	0.0000	(-0.157, 0.161)	0.0001
425	(-0.165, 0.167)	0.0002	(-0.162, 0.159)	0.0000

Step 3: Significance Results

Figure 10 shows five histograms. Each histogram corresponds to one of the cut-off scenarios $\{0, 15, 30, 45, 425\}$. In each histogram the regression estimates that are statistically significant are shown in blue, whereas the estimates that are statistically not different from zero are shown in red. The dashed line denotes the baseline regression estimate and the dotted lines indicate the critical region that was found according to the complete bootstrap analysis. The histograms for the cut-off values of $\{15, 30, 45\}$ show what we have already observed in the previous section: there is a wide dispersion in the result of the regression estimates, but this instability

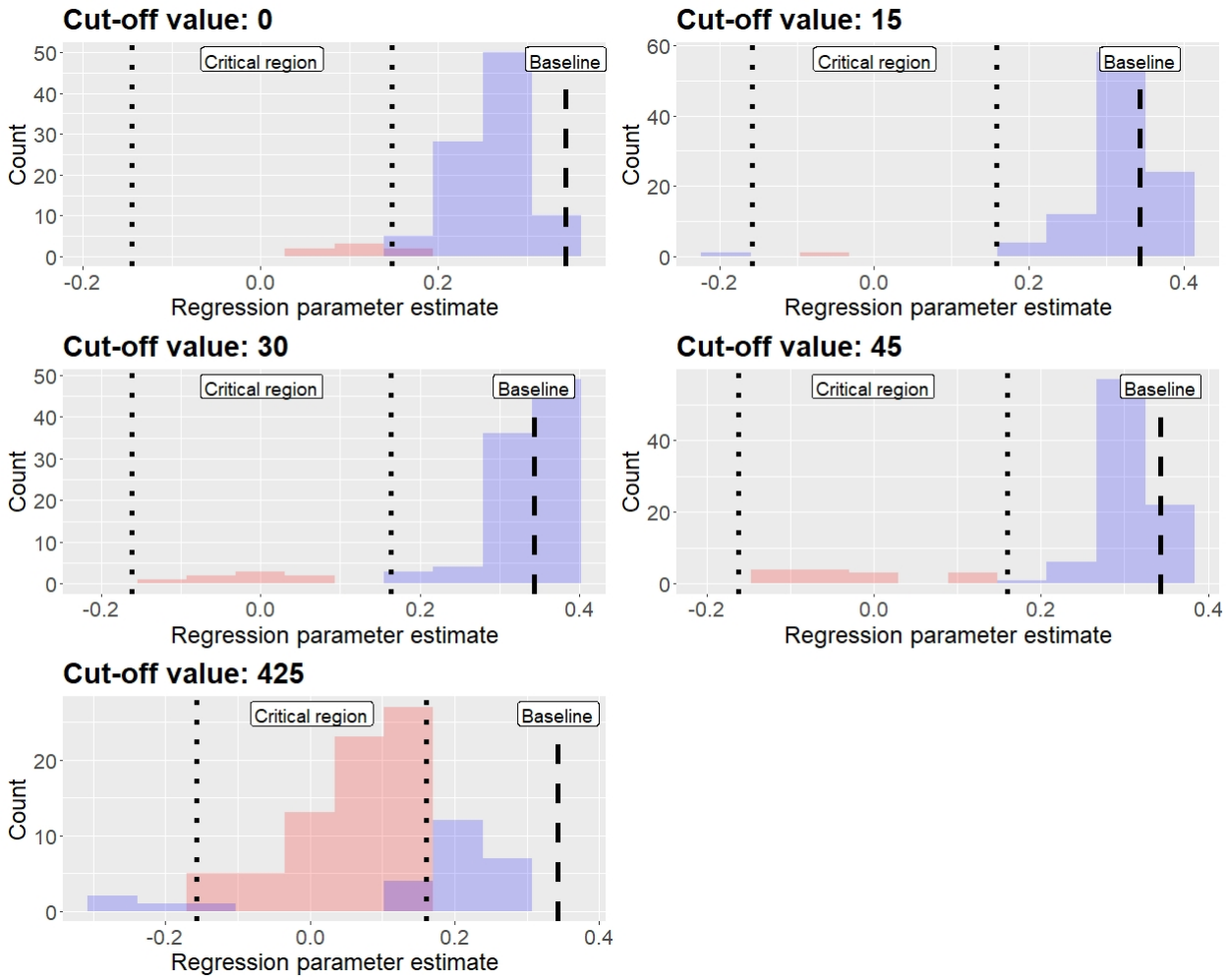


Figure 10: For 100 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined for each cut-off value in the set $\{0, 15, 30, 45, 425\}$. Histograms of the 100 parameter estimates are shown. The dashed lines shows the baseline regression estimate. The dotted lines show the border of the rejection region determined by a complete bootstrap procedure. The estimates in blue are statistically significant, whereas those in red are not statistically significant.

is largely removed if we only focus on the statistically significant values. For the cut-off values of 15 and 425 we observe something strange: there are negative observations that are statistically significant. This finding is problematic, as the conclusion of the paper of Bandiera et al. (2020) that CEO leader perform better than CEO managers is not valid for this case. The histogram for the cut-off value of 0 and 425 are not centered around the baseline coefficient estimate. Besides, they seem to consist of more insignificant values, especially for the case with a cut-off of 425.

Table 7 provides us with summary statistics for the different cases of the cut-off value. If we do not account for the statistical significance (this is column 1 in the table), we observe that the instability in the regression

estimates is increasing in the cut-off value. This means that if the number of unique activities in the data set is smaller, the LDA algorithm produces more dispersed results. This is likely due to the fact that if the number of activities is decreasing, the activities in itself are more routine and thus do not provide the researcher with additional information.

Besides, Table 7 shows us that that those values that are statistically significant (this is column 2 in the table) have a lower instability compared to the instability when we do not correct for the significance (this is column 1). This results holds true for all the different cut-off scenarios except for those with a cut-off value equal to 425. As we have seen in the histogram in Figure 10, the case in which there is a cut-off value of 425 is really unstable and many of the coefficient estimates are insignificant: Table 7 shows us that only 26 of the 100 coefficient estimates are statistically significant from 0 in this case.

Step 4: Perplexity Results

Just like in the previous section, we will now study the relationship between the perplexity and the instability of the regression estimates. Figure 11 shows for each cut-off scenario a scatter plot in which the regression estimates are denoted on the horizontal axis and the perplexity is shown on the vertical axis. The regression estimates that are accepted based on the critical perplexity are shown in blue, whereas those that are rejected based on the critical perplexity are shown in red. The vertical dashed line denotes the baseline regression estimate. In the figure we observe that for each cut-off scenario, expect for the case with a cut-off of 425, there is a cluster of regression estimates.

For each cut-off scenario, we visually determine a value for the critical perplexity to approach the cluster of coefficient estimates as good as possible. We find the following critical perplexity values for the cut-off cases $\{0, 15, 30, 45, 425\}$: $\{674.140, 368.702, 253.131, 200.114, 31.886\}$. These values are shown in the scatter plots in Figure 11 by the horizontal dotted lines. We observe that for all cut-off values except for the cases 0 and 425, the clusters are nicely obtained this way. In Table 11 in Appendix C we show a comparison between the significant values and the values chosen based on their perplexity. We observe that in all cases there is much agreement between both methods. This agreement is, however, not true for a cut-off value of 425, but this cut-off scenario seems completely off.

We compare the summary statistics for the case in which we only take the estimates with a low perplexity value into account (this is the third column in Table 7) to the case in which we take all the estimates into account (the first column of the table) and the case in which we only take those estimates into account that are statistically significant (the second column of the table). We observe that across the three different cases, the instability is smallest for the low perplexity case. However, the number of estimates that are in the sample is also much smaller for the low perplexity case compared to the other two cases. It seems that

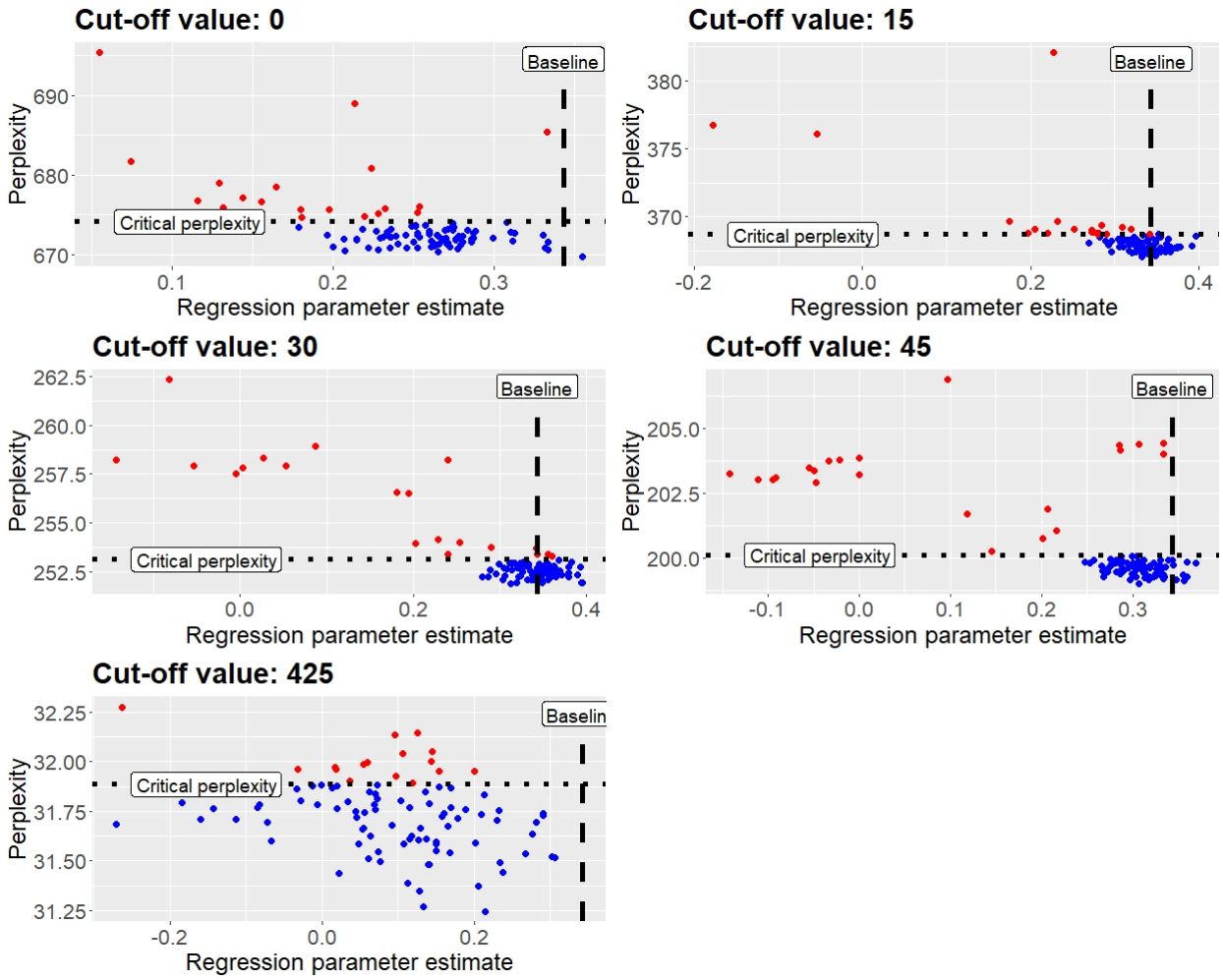


Figure 11: For 100 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined for each cut-off value in the set $\{0, 15, 30, 45, 425\}$. Scatter plots of the 100 parameter estimates are shown with the regression estimates on the horizontal axis and the value of the perplexity on the vertical axis. The dashed line shows the baseline regression estimate. The dotted lines show the critical perplexity. This critical perplexity is chosen to best mimic the cluster of estimates that is visually observed. The estimates in blue are accepted based on the critical perplexity, whereas those in red are rejected.

the instability can be reduced for all cut-off values, except for a cut-off value of 425: the instability for this cut-off value is high for all three cases. We also observe that the mean values of the low perplexity case are in good alignment with the baseline regression estimate.

Conclusion

To conclude, we summarise the three main results found from this robustness analysis. The first result is that the value of the coefficient estimate is influenced by the cut-off value, but that this effect is small. The

Table 7: Summary statistics (minimum value, mean value, maximum value, the instability, and the number of observations) of three different groups of regression coefficient estimates for each cut-off value in the set $\{0, 15, 30, 45, 425\}$. In the first column the statistics are shown over all the 100 runs of the algorithm. In the second column the statistics are shown for the set of statistically significant coefficient estimates. In the third column the statistics are shown for those observations that best mimic the cluster of values with a low perplexity.

Cut-off value: 0	All values	Significant values	Low perplexity values
Minimum	0.055	0.156	0.179
Mean	0.247	0.257	0.263
Maximum	0.355	0.355	0.355
Instability	0.053	0.038	0.033
Number observations	100	93	80

Cut-off value: 15	All values	Significant values	Low perplexity values
Minimum	−0.178	−0.178	0.270
Mean	0.313	0.317	0.336
Maximum	0.396	0.396	0.396
Instability	0.074	0.065	0.025
Number observations	100	99	80

Cut-off value: 30	All values	Significant values	Low perplexity values
Minimum	−0.143	0.181	0.280
Mean	0.306	0.334	0.344
Maximum	0.396	0.396	0.396
Instability	0.105	0.042	0.026
Number observations	100	92	80

Cut-off value: 45	All values	Significant values	Low perplexity values
Minimum	−0.142	0.202	0.248
Mean	0.258	0.304	0.307
Maximum	0.370	0.370	0.370
Instability	0.121	0.032	0.027
Number observations	100	86	78

Cut-off value: 425	All values	Significant values	Low perplexity values
Minimum	−0.271	−0.271	−0.271
Mean	0.085	0.162	0.096
Maximum	0.306	0.306	0.306
Instability	0.120	0.168	0.119
Number observations	100	27	80

main conclusion of the paper of Bandiera et al. (2020) (leader CEOs have higher firm profits than manager CEOs) is supported by each cut-off scenario, although the size of the effect is a bit different. The second result from this analysis is that although the instability can be removed by looking at the statistical significance, this is not a good way to do. We have seen that by only considering the statistical significant results, there may be negative results that are still statistically significant. Those results are rejected by the perplexity method, which thus proves to be a more robust way to obtain a stable LDA sample. The third result from this analysis is that the cut-off value has an effect on the stability. The instability is small (when corrected for the perplexity) for the cases $\{0, 15, 30, 45\}$. For the case in which the cut-off value is 425 the instability is large and is not resolved for by looking at the perplexity. As a cut-off value of 425 corresponds to a sample with only very mundane tasks, a small vocabulary size may thus not be preferred when the goal is to obtain stable LDA results.

5.3.2 Different Prior Settings

Introduction

In the baseline model, a symmetric prior is placed on both the hyperparameter of θ and β. Bandiera et al. (2020) give no justification for choosing this type of hyperparameters. They also do not run robustness checks to see what happens if they choose a different hyperparameter. In Section 2.3 we have read that, according to Wallach et al. (2009), choosing either a symmetric or asymmetric hyperparameter does not influence the topic coherence. In this section we will check if this is indeed the case. Besides, we are interested if the main results of the baseline model can also be found by using different hyperparameter specifications.

In the baseline model the following prior settings are chosen: $(\theta, \beta) = (1.0, 0.1)$. As these are two symmetric priors we refer to this scenario as SS. To check the robustness we run four additional prior scenarios. Each specification will be run for 100 times. We study the following specifications, where S stands for symmetric, A for asymmetric, the subscript h for high compared to the baseline scenario, and the subscript l for low compared to the baseline scenario:

- the scenario SA_l with $(\theta, \beta) = (1.0, 0.01)$,
- the scenario SA_h with $(\theta, \beta) = (1.0, 0.5)$,
- the scenario A_lS with $(\theta, \beta) = (0.5, 0.1)$,
- the scenario A_hS with $(\theta, \beta) = (1.5, 0.1)$,
- the scenario SS with $(\theta, \beta) = (1.0, 0.1)$.

Compared to the previous two sections, the robustness and stability checks for the different hyperparameters do not yield any interesting results. We will thus only look at a table that shows the summary statistics for three different cases: a case with all regression estimates, a case in which we only take into account the estimates that are statistically significant based on a complete bootstrap analysis, and a case in which only those estimates with a low enough perplexity value are taken into account. More detailed results for each case can be found in the Appendices D to H.

Results and Conclusion

In Table 8 we show the results of the stability and robustness analysis for the five different prior scenarios: $\{SA_l, SA_h, A_lS, A_hS, SS\}$. The first column shows the summary statistics if we take into account all the 100 runs of the LDA algorithm. We observe that for each prior specification there is quite some instability in the regression estimate. Although the minimum, mean and maximum values seems to be a bit different across the five prior specifications, this difference is only small. Besides, the main conclusion of the paper of Bandiera et al. (2020) is not affected by the choice of the prior hyperparameters.

The second column of Table 8 shows for each prior specification the results if we only take into account the set of regression estimates that is statistically significant based on a complete bootstrap analysis (see for more details on this complete bootstrap analysis Appendix D and E). Yet again, the exact prior specification seems to have little impact on the summary statistics. For all five specifications, the instability is removed by taking into account the statistical significance.

However, instead of focusing on the statistical significance it is better to focus on a critical perplexity value (see Appendix G and H for a look at how the critical perplexity is determined). Column three of Table 8 reveals that the instability is indeed lower for the case in which we check for the perplexity. However, now we do for the first time observe a small difference across the five prior specifications as the number of estimates in the sample of SA_l is considerably lower compared to the other specifications. To conclude this section, we have shown that the choice of prior specification has little effect on the regression estimates. Besides, the instability is same over all five prior specifications studied.

5.4 Comparison with Lasso

We want to compare the LDA analysis on the CEO diary data set, to a Lasso regression on the same data set. To do this, we first recall the main things that we have learned from the LDA analysis of Bandiera et al. (2020). In Section 3.3 we discussed how LDA has been applied to the CEO diary data set by Bandiera et al. (2020). The results indicated that a CEO can be characterised as a mixture between a leader CEO and

Table 8: Summary statistics (minimum value, mean value, maximum value, the instability, and the number of observations) of three different groups of regression coefficient estimates for each prior scenario in the set $\{SA_l, SA_h, A_lS, A_hS, SS\}$. In the first column the statistics are shown over all the 100 runs of the algorithm. In the second column the statistics are shown for the set of statistically significant coefficient estimates. In the third column the statistics are shown for those observations that best mimic the cluster of values with a low perplexity.

Prior scenario: SA_l	All values	Significant values	Low perplexity values
Minimum	−0.194	−0.194	0.292
Mean	0.286	0.322	0.356
Maximum	0.428	0.428	0.417
Instability	0.128	0.084	0.033
Number observations	100	87	60

Prior scenario: SA_h	All values	Significant values	Low perplexity values
Minimum	−0.068	0.206	0.292
Mean	0.312	0.332	0.338
Maximum	0.381	0.381	0.381
Instability	0.081	0.032	0.021
Number observations	100	93	85

Prior scenario: A_lS	All values	Significant values	Low perplexity values
Minimum	−0.006	0.158	0.250
Mean	0.288	0.307	0.317
Maximum	0.380	0.380	0.380
Instability	0.080	0.043	0.028
Number observations	100	92	77

Prior scenario: A_hS	All values	Significant values	Low perplexity values
Minimum	−0.084	0.173	0.300
Mean	0.317	0.350	0.361
Maximum	0.411	0.411	0.411
Instability	0.113	0.047	0.025
Number observations	100	90	83

Prior scenario: SS	All values	Significant values	Low perplexity values
Minimum	−0.143	0.181	0.280
Mean	0.306	0.334	0.344
Maximum	0.396	0.396	0.396
Instability	0.105	0.042	0.026
Number observations	100	92	80

a manager CEO. Table 1 shows us which features belong to a leader CEO and which features belong to a manager CEO. A leader CEO is characterised as having many meetings at the C-suite level. Those meetings are with both outsiders and insiders and many meetings consisted of conference calls, video conferences and phone calls. For the manager CEOs, the meeting are mostly plants visits. Besides, manager CEOs often speak with people from production and suppliers. Most of their meetings take place with only outsiders. The regression analysis indicated that leader CEOs have higher firm profits than manager CEOs.

To be able to compare Lasso to LDA we have to run a Lasso regression. In Section 4.7, we have described how the Lasso regression looks like. However, in that section we did not discuss which variables we will add to our regression equation. We will run the Lasso regression with two different sets of variables. The first set of variables, which we will refer to as the basis variables, consists of those variables that directly affect the characterisation of CEO behavior. These eight variables are {type_conference_call, type_phone_call, type_video_conference, bunits, groupcom, ins, out, production, suppliers, type_site_visit} and this selection of variables is based on Table 1, which is taken from Bandiera et al. (2020). Of the basis variables, the first six variables correspond to a leader CEO, whereas the last four variables correspond to a manager CEO. The second set of variables, which we will refer to as the complete variables, consists of all the variables that we have in our CEO diary data set.

The CEO diary data set that has been used for the LDA analysis consists of information about $1,114$ CEOs. Information about all these CEOs is taken into account when constructing a CEO behavior index. This CEO behavior index is then used as input for a regression analysis that uses data from 920 CEOs. This drop in the number of CEOs is caused by missing values in the firm data. As the OLS regression cannot be performed using missing values, some of the CEOs that were used in the LDA analysis have been removed from the OLS analysis. For our Lasso regression, we also cannot have missing values in our data. Merging the firm data set with the CEO diary data set and performing some data cleaning, yields a Lasso data set consisting of information of 919 CEOs. An important step in the data cleaning for the Lasso regression, is that missing values are being replaced by zeros. This is done to ensure that our data set has a large enough size (removing missing values would give a data set with only 236 CEOs). Replacing missing values by zero may cause bias in the Lasso regression results. However, we have conducted three robustness checks in which we alternatively (1) replaced the missing values by 0.5 (as we have binary variables), (2) replaced the missing values of a variable by the mean value over those observations that are present, and (3) did remove all the observations for which we do have a missing value. In Appendix I the results of these checks are presented and they do not deviate from the results found in the current section.

To compare LDA to Lasso we use two different methods. Our first way of comparing both methods, is to look at the interpretation of the Lasso coefficients. If Lasso and LDA agree on the CEO diary data set, the coefficients of the Lasso regression should be interpreted in the following way:

- the positive Lasso coefficients should correspond to the features a leader CEO has (these are the variables {type_conference_call, type_phone_call, type_video_conference, bunits, groupcom, ins} in the basis variables data set),
- the negative Lasso coefficients should correspond to the features a manager CEO has (these are the variables {out, production, suppliers, type_site_visit} in the basis variables data set),
- the variables that Lasso shrinks towards zero should correspond to neither a leader or a manager CEO.

A second way to compare both methods is to look at the correlation that exists between the two methods. We can form a CEO behavior index for the Lasso regression (see the methodology in Section 4.7) and see how this correlates with the CEO behavior index formed by the LDA regression. If the two methods perform the same on the data set, there should be a high correlation between the two indices.

Method 1: Interpretation

In Table 9 we present the results of the Lasso regression. However, this table only shows part of the variables that are in our Lasso regression: only the basis CEO diary variables and the complete CEO diary variables that are non-zero are shown. Table 17 in Appendix J shows the complete Lasso regression results.

Table 9 gives the results of two distinct regressions. The final two columns in this table correspond to an OLS regression of the baseline model CEO behavior index on the CEO diary variables. This OLS regression can be used as a "sanity check": variables related to a leader CEO should have a positive OLS coefficient, whereas variables related to a manager CEO should have a negative OLS coefficient. In the table, column (1) relates to the basis CEO diary variables and column (2) relates to the complete CEO diary variables. In the table we observe something unexpected. The first six basis variables in the table correspond to a leader CEO, and we thus expect to find positive coefficients: this is not the case for the variables type_conference_call and type_phone_call. For the last four basis variables (that correspond to a manager CEO), we expect to find negative coefficients, but this is not the case for the variable out. An explanation for this finding is that Bandiera et al. (2020) merged some of the variables to construct Table 1. E.g., the first three basis variables are all related to the type of meeting a CEO has, and according to Bandiera et al. (2020) a merged variable of these three variables is a characteristic of a leader CEO. However, our analysis shows that if the merged variable is related to a leader CEO, that this is not an indication that the distinct variables are also related to a leader CEO. In our comparison of the Lasso regression to the LDA analysis, we will thus compare the signs of the Lasso coefficients to the signs of the OLS coefficients that are in the last two columns of Table 9.

The first two columns of Table 9 present the results of the Lasso regression. Column (1) relates to the basis

CEO diary variables and column (2) relates to the complete CEO diary variables. Looking in column (1), we see that four out of ten of the basis variables are shrunk towards zero. This also holds true for column (2), where we include the complete CEO diary variables. This indicates that not all of the basis variables have a large effect on firm performance according to the LDA analysis. This contradicts what we find with the OLS regression, as three out of four of these variables have a significant effect on the CEO behavior index. Furthermore, we see that not all of the Lasso coefficients of the basis variables in column (1) have the same signs as the OLS coefficients in column (1). This indicates, that the distinction between the manager and leader CEO as found by the LDA analysis is not found by the Lasso analysis. This becomes even more clear when comparing the Lasso regression in column (2) to the OLS regression in column (2): of the 22 variables in Table 9 that have a non-zero Lasso coefficient, 6 variables have a different sign across the two regression methods. Based on the Lasso regression results, a concrete classification of CEO behavior cannot be made.

Table 9: The results of two distinct regressions. In the first two columns the results of a Lasso regression are shown, whereas the last two columns show the results of an OLS regression. The dependent variable for the Lasso regression is the firm performance and the dependent variable for the OLS regression is the baseline model CEO behavior index. Both regressions have the same independent variables: CEO diary and firm control variables. Columns (1) show the results for the basis CEO diary variables, whereas columns (2) show the results for the complete CEO diary variables. In the Lasso regression, coefficients that are equal to zero, indicate that the variable is shrunk to zero. In the OLS regression, * indicates significance at 10% level, ** indicates significance at 5% level, and *** indicates significance at 1% level. This table only shows part of the results (the complete results can be found in Appendix K): only the basis CEO variables and the complete CEO variables that are non-zero are shown.

	Lasso regression:		*OLS regression:*	
	firm_performance		CEO_behavior	
	(1) Basis	(2) Complete	(1) Basis	(2) Complete
type_conference_call	0	0	0.031	−0.026
type_phone_call	−0.016	−0.005	−0.028	−0.009
type_video_conference	−0.019	−0.018	0.040*	0.010
bunits	0.012	0.008	0.282***	0.128
groupcom	0.025	0.022	0.442***	0.200
ins	0.014	0.011	0.118***	0.231***
out	0	0	0.216***	0.220***
production	0	0	−0.201***	−0.430
suppliers	0	0	−0.196***	−0.239**
type_site_visit	−0.044	−0.041	−0.115***	−0.162***
strategy		0.006		−0.012
hr		0.008		0.092

board	0.042	0.032
compliance	−0.002	−0.052
it	−0.008	−0.062**
legal	0.030	−0.031
retail	−0.021	−0.025
clients	−0.021	−0.243
investors	0.016	−0.061
lawyers	−0.007	−0.096*
n_functions	0.011	0.512
type_business_meal	0.024	0.015
F_duration_1hr	−0.004	0.035
F_planned_planned	0.018	0.049**
F_participants_missing	0.013	0.030
F_participants_one_ppl	−0.002	0.006

Method 2: Correlation

In Table 9 we discussed the Lasso regression results. Based on the Lasso regression estimates, we can form a Lasso CEO behavior index (see Section 4.7 for the exact methodology). The correlation between the Lasso CEO behavior index and the LDA CEO behavior index is 0.451 when we only take into account the basis CEO diary variables, and 0.479 when we take into account the complete CEO diary variables. We observe that both correlations are almost equal to each other, indicating that there is not a big difference between using either the basis CEO diary variables or the complete CEO diary variables. Furthermore, we observe that the Lasso and LDA CEO behavior indices are only moderately correlated to each other. This indicates that there is a substantial difference between the two indices. We did check if different values of λ, gave rise to a higher correlation between the two methods and a different variable selection procedure. This was not the case, indicating that our Lasso results are robust (see Appendix K for some graphical robustness checks).

Conclusion

In this section, we tried to check if the results found by the LDA analysis are in agreement with results produced by a Lasso regression. As both methods serve as a dimensionality reduction tool, we would expect that both methods would find similar results. However, our analysis indicates that on the CEO diary data set, LDA and Lasso do not produce similar results. Both in terms of interpretation and a quantitative measure, we could not find any evidence that the two methods perform the same on this data set.

6 Conclusion and Discussion

In this Thesis we have studied the working of the machine learning algorithm LDA on structured text data. The algorithm is stochastic, thus each new run of the algorithm on the same data set provides us with new results. Previous research has indicated that LDA does not produce stable results on non-structured text data: a single run of LDA does not produce valid to interpret results. No researchers have yet looked at the stability of LDA on structured text data, which is what we did in this Thesis. Furthermore, we compared the performance of LDA to the performance of Lasso regression. These two methods both serve as a dimensionality reduction tool, so knowing if the two methods can be used interchangeably on structured text data is important if LDA turns out to be non-stable. In this Thesis, we made use of the data set and framework by Bandiera et al. (2020). In their study, Bandiera et al. (2020) determined an index for CEO behavior using LDA on structured text data.

We have shown that LDA is not stable on a structured text data set: two runs of the algorithm on the same data set produce different and sometimes even contradicting results. As this shows that a single run of the LDA algorithm cannot be interpreted in a safe way, we conclude that LDA is not stable on this type of data. However, we have shown that there is reliable way to ensure that we end up with LDA results that are stable. The perplexity, which is a statistical measure of model fit, can be used as a rejection criterion for different LDA runs. By defining a critical level of perplexity, we can ensure that runs of the LDA algorithm that have a perplexity lower than the critical level of perplexity are stable. Furthermore, we have shown that runs of the LDA algorithm that are statistically significant are stable. However, as only taking into account statistical significant estimates can be seen as cherry picking (we do not know beforehand if there is a statistical significant effect) this is not a reliable rejection criterion to produce stable LDA results.

Besides, we did two robustness checks: on the vocabulary size and on the prior hyperparameters. These robustness checks served two goals: we wanted to test if choosing a different model specification changes the main conclusions of Bandiera et al. (2020), and we wanted to test if the stability of the LDA algorithm was affected by choosing a different model specification. It turned out that changing the model specifications did not change the main conclusions of Bandiera et al. (2020), which shows that their results are robust. Regarding the stability, we found that the different robustness checks did not produce stable results. However, the perplexity again serves as a measure to fix the stability of the results. The only conflicting result was that when the vocabulary size becomes too small, the LDA results are not stable and cannot be made stable by taking the perplexity into account.

Furthermore, we have shown that LDA does not produce similar results to a Lasso regression on the studied data set. The interpretation of the Lasso coefficients was not in line with the interpretation of the topics as produced by LDA. Also a CEO behavior index formed by Lasso turned out to be only moderately correlated

to a CEO behavior index formed by LDA. This all indicates that, although both Lasso and LDA can be used to reduce the dimensionality of the data set, they do not do this in a similar way. The two methods thus cannot be used interchangeably and when a researcher wants to study structured text data, he needs to make a supported choice for either of the two methods.

Regarding the scope and generalisability of the Thesis there are some concerns. In this Thesis we have studied only one example of a structured text data set. It may be the case that the results we found for this data set, do not hold true for other applications of the LDA algorithm. Besides, we studied a data set that is relatively small in size (there are $1,114$ CEOs and $42,233$ activities). LDA is normally applied to much larger data sets (e.g. Azqueta-Gavaldón, 2017: $40,454$ observations; Hansen et al., 2018: $26,645$ observations; Larsen & Thorsrud, 2019: $459,745$ observations), so for further research it may be interesting to study the stability of LDA on a larger structured text data set.

Also related to the Lasso-LDA comparison, it may be the case that Lasso and LDA give comparable results on a different data set. This is something that can be studied by analysing more data sets using both methods. Besides, we did not come up with an explanation as to why Lasso and LDA perform differently on the CEO diary data set. By systematically studying more structured text data sets using both methods, we may develop guidelines that tell us when which method can best be used.

The results we found in relation to the perplexity are interesting, but it is not yet known how we can implement the perplexity method into our analysis. We have seen that choosing a critical level of perplexity can ensure us that an LDA algorithm provides stable results, but we do not know if this critical level can be determined beforehand. If a researcher first needs to run the LDA algorithm 100 times in order to determine a critical perplexity, we may wonder if this method will ever be efficient. However, running the algorithm 100 times is doable: it just takes much time. A possible solution could be to run the LDA algorithm for a small number of times, e.g. 10 times, and then pick the run of the LDA algorithm with the smallest perplexity. Or alternatively, a researcher could run the algorithm 100 times parallel in the cloud.

In our main analysis in Section 5.2 we used 250 runs of the LDA algorithm, but we have seen in Section 5.3, where we looked at the robustness checks, that 100 runs of the LDA algorithm gave the same results. It is interesting to study if the same results could also be found using even a smaller number of runs. This is also related to the previous paragraph, as running the algorithm less times makes this method more efficient.

Furthermore, in this Thesis we have only studied two robustness checks of the LDA algorithm: the vocabulary size and the hyperparameter specifications. There are, however, many more options for the robustness. It could be the case that, e.g., the choice of inference method or the computer packages that are used have an influence on the stability, but we have not checked for this. This is something for further researchers to do.

References

Adragni, K. P., & Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4385–4405.

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology*, *98*, 74–88.

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *6*(1).

Alhawarat, M., & Hegazi, M. (2018). Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access*, *6*, 42740–42749.

Ambrosino, A., Cedrini, M., Davis, J. B., Fiori, S., Guerzoni, M., & Nuccio, M. (2018). What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology*, *25*(4), 329–348.

Anandkumar, A., Foster, D. P., Hsu, D., Kakade, S. M., & Liu, Y.-K. (2015). A spectral algorithm for latent dirichlet allocation. *Algorithmica*, *72*(1), 193–214.

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

Athey, S. (2019). 21. the impact of machine learning on economics. In *The economics of artificial intelligence* (pp. 507–552). University of Chicago Press.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725.

Azqueta-Gavaldón, A. (2017). Developing news-based economic policy uncertainty index with unsupervised machine learning. *Economics Letters*, *158*, 47–50.

Bandiera, O., Prat, A., Hansen, S., & Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, *128*(4), 1325–1369.

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*(6), 1371–1391.

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

Blei, D. M., & McAuliffe, J. D. (2010). Supervised topic models. *arXiv preprint arXiv:1003.0783*.

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

Bui, Q. V., Sayadi, K., Amor, S. B., & Bui, M. (2017). Combining latent dirichlet allocation and k-means for documents clustering: Effect of probabilistic based distance measures. In *Asian conference on intelligent information and database systems* (pp. 248–257).

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge

university press.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Chen, Z., & Doss, H. (2019). Inference for the number of topics in the latent dirichlet allocation model via bayesian mixture modeling. *Journal of Computational and Graphical Statistics*, *28*(3), 567–585.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, *75*(3), 1371–1415.

Crain, S. P., Zhou, K., Yang, S.-H., & Zha, H. (2012). Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In *Mining text data* (pp. 129–161). Springer.

De Waal, A., & Barnard, E. (2008). Evaluating topic models with stability.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, *64*(2-3), 221–245.

Emmert-Streib, F., & Dehmer, M. (2019). High-dimensional lasso-based computational regression models: Regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction*, *1*(1), 359–383.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74.

Gissler, S., Oldfather, J., & Ruffino, D. (2016). Lending on hold: Regulatory uncertainty and bank lending standards. *Journal of Monetary Economics*, *81*, 89–101.

Greenberg, E. (2012). *Introduction to bayesian econometrics*. Cambridge University Press.

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? stability analysis for topic models. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 498–513).

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

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801–870.

Hermalin, B. E. (1998). Toward an economic theory of leadership: Leading by example. *American Economic Review*, 1188–1206.

Hofmann, T. (2013). Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*.

Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production: evidence from the collapse of international science. *The Quarterly Journal of Economics*, *133*(2), 927–991.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211.

Koltcov, S., Koltsova, O., & Nikolenko, S. (2014). Latent dirichlet allocation: stability and applications

to studies of user-generated content. In *Proceedings of the 2014 acm conference on web science* (pp. 161–165).

Kotter, J. P. (2001). What leaders really do. *Harvard business review*, *79*(11).

Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, *210*(1), 203–218.

Larsen, V. H., Thorsrud, L. A., & Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, *117*, 507–520.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Lee, S., Song, J., & Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computer Information Systems*, *51*(1), 1–10.

Limsettho, N., Hata, H., & Matsumoto, K.-i. (2014). Comparing hierarchical dirichlet process with latent dirichlet allocation in bug report multiclass classification. In *15th ieee/acis international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (snpd)* (pp. 1–6).

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Mantyla, M. V., Claes, M., & Farooq, U. (2018). Measuring lda topic stability from clusters of replicated runs. In *Proceedings of the 12th acm/ieee international symposium on empirical software engineering and measurement* (pp. 1–4).

Mazarura, J., & De Waal, A. (2016). A comparison of the performance of latent dirichlet allocation and the dirichlet multinomial mixture model on short text. In *2016 pattern recognition association of south africa and robotics and mechatronics international conference (prasa-robmech)* (pp. 1–6).

Minka, T. P., & Lafferty, J. (2012). Expectation-propogation for the generative aspect model. *arXiv preprint arXiv:1301.0588*.

Mintzberg, H. (1973). The nature of managerial work.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Nakano, T., Yoshii, K., & Goto, M. (2014). Vocal timbre analysis using latent dirichlet allocation and cross-gender vocal timbre similarity. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5202–5206).

Newman, D., Porteous, I., & Triglia, S. (2011). Distributed gibbs sampling for latent variable models. *Scaling Up Machine Learning: Parallel and Distributed Approaches*.

Péladeau, N., & Davoodi, E. (2018). Comparison of latent dirichlet modeling and factor analysis for topic extraction: A lesson of history. In *Proceedings of the 51st hawaii international conference on system*

sciences.

Rieger, J., Koppers, L., Jentsch, C., & Rahnenführer, J. (2020). Improving reliability of latent dirichlet allocation by assessing its stability using clustering techniques on replicated runs. *arXiv preprint arXiv:2003.04980*.

Rieger, J., Rahnenführer, J., & Jentsch, C. (2020). Improving latent dirichlet allocation: On reliability of the novel method ldaprototype. In *International conference on applications of natural language to information systems* (pp. 118–125).

Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Understanding text pre-processing for latent dirichlet allocation. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics* (Vol. 2, pp. 432–436).

Stantcheva, S. (2021). Understanding tax policy: How do people reason? *The Quarterly Journal of Economics*, *136*(4), 2309–2369.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952–961).

Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 ieee international conference on data science and advanced analytics (dsaa)* (pp. 165–174).

Syed, S., & Spruit, M. (2018). Selecting priors for latent dirichlet allocation. In *2018 ieee 12th international conference on semantic computing (icsc)* (pp. 194–202).

Tang, H., Shen, L., Qi, Y., Chen, Y., Shu, Y., Li, J., & Clausi, D. A. (2012). A multiscale latent dirichlet allocation model for object-oriented clustering of vhr panchromatic satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(3), 1680–1692.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282.

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

Weigel, J. L. (2020). The participation dividend of taxation: How citizens in congo engage more with the state when it tries to tax them. *The Quarterly Journal of Economics*, *135*(4), 1849–1903.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. In *Bmc bioinformatics* (Vol. 16, pp. 1–10).

Zhao, W., Zou, W., & Chen, J. J. (2014). Topic modeling for cluster analysis of large biological and medical datasets. In *Bmc bioinformatics* (Vol. 15, pp. 1–11).

7 Appendix

Appendix A - Comparison of the Stata and R Regression Output of the Baseline Model

Table 10: Regression results of the baseline regression equation (Equation 2). Column (1) shows the Stata output as used by Bandiera et al. (2020) in their paper. Column (2) shows the R output.

	Dependent variable:	
	ly	
	(1) Stata	(2) R
ceo_behavior	0.343***	0.343***
	(0.108)	(0.098)
lemp	0.889***	0.889***
	(0.040)	(0.036)
cons	0.245**	0.245**
	(0.123)	(0.111)
active	−0.101	−0.101
	(0.179)	(0.163)
X_Iyear_2008	−0.133	−0.133
	(0.276)	(0.251)
X_Iyear_2009	0.019	0.019
	(0.533)	(0.483)
X_Iyear_2010	0.394	0.394
	(0.271)	(0.246)
X_Iyear_2011	0.800***	0.800***
	(0.265)	(0.240)
X_Iyear_2012	0.719***	0.719***
	(0.221)	(0.200)
X_Iyear_2013	0.423**	0.423**
	(0.190)	(0.172)
X_Icty_2	0.887***	0.887***
	(0.315)	(0.286)
X_Icty_3	1.049***	1.049***

	(0.210)	(0.191)
X_Icty_4	0.716**	0.716**
	(0.310)	(0.281)
X_Icty_5	−0.141	−0.141
	(0.659)	(0.597)
X_Icty_6	0.176	0.176
	(0.474)	(0.429)
emp_imputed	−0.149	−0.149
	(0.274)	(0.248)
pa	0.169**	0.169**
	(0.086)	(0.078)
reliability	−0.020	−0.020
	(0.029)	(0.027)
ww1	−0.271	−0.271
	(0.376)	(0.341)
ww2	0.188	0.188
	(0.458)	(0.415)
ww3	−0.143	−0.143
	(0.380)	(0.345)
ww4	0.154	0.154
	(0.333)	(0.302)
ww5	−0.134	−0.134
	(0.354)	(0.321)
ww6	−0.025	−0.025
	(0.332)	(0.301)
ww7	0.011	0.011
	(0.375)	(0.340)
ww8	0.242	0.242
	(0.346)	(0.313)
ww9	−0.096	−0.096
	(0.352)	(0.319)
ww10	0.279	0.279
	(0.364)	(0.330)
ww11	0.118	0.118
	(0.335)	(0.304)
ww12	0.244	0.244

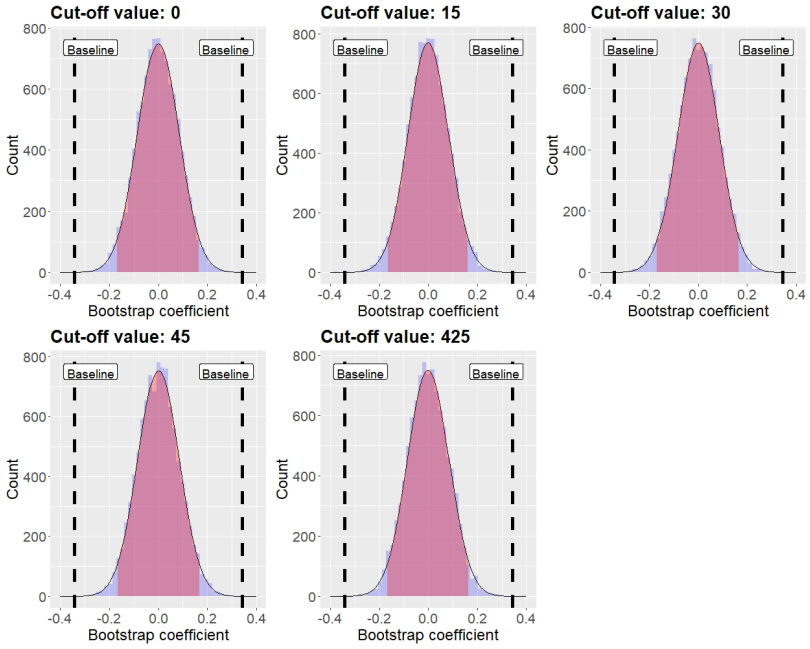
	(0.361)	(0.327)
ww13	−0.135	−0.135
	(0.397)	(0.359)
ww14	0.149	0.149
	(0.413)	(0.374)
ww15	0.284	0.284
	(0.402)	(0.364)
ww16	0.485	0.485
	(0.400)	(0.362)
ww17	0.305	0.305
	(0.331)	(0.299)
ww18	0.232	0.232
	(0.315)	(0.286)
ww19	0.301	0.301
	(0.397)	(0.360)
ww20	0.575**	0.575***
	(0.222)	(0.201)
ww21	0.582***	0.582***
	(0.200)	(0.181)
ww22	0.762	0.762
	(0.537)	(0.487)
ww24	−0.366	−0.366
	(0.249)	(0.226)
ww25	0.079	0.079
	(0.305)	(0.276)
ww26	−0.081	−0.081
	(0.313)	(0.283)
ww27	0.207	0.207
	(0.276)	(0.250)
ww28	0.162	0.162
	(0.341)	(0.309)
ww29	0.068	0.068
	(0.330)	(0.299)
aa1	−0.146	−0.146
	(0.230)	(0.208)
aa2	−0.029	−0.029

	(0.410)	(0.371)
aa3	0.363	0.363
	(0.462)	(0.419)
aa4	−0.834**	−0.834**
	(0.417)	(0.378)
aa5	−0.001	−0.001
	(0.384)	(0.348)
aa6	−0.013	−0.013
	(0.341)	(0.309)
aa7	−0.491	−0.491
	(0.526)	(0.477)
aa8	0.130	0.130
	(0.318)	(0.288)
aa9	0.039	0.039
	(0.293)	(0.265)
aa10	−0.079	−0.079
	(0.269)	(0.244)
aa11	−0.211	−0.211
	(0.414)	(0.375)
aa12	−0.259	−0.259
	(0.424)	(0.384)
aa13	−0.333	−0.333
	(0.360)	(0.327)
aa14	0.653**	0.653**
	(0.287)	(0.262)
aa15	−0.082	−0.082
	(0.389)	(0.353)
aa16	−0.121	−0.121
	(0.302)	(0.274)
aa17	−0.161	−0.161
	(0.216)	(0.196)
aa18	0.074	0.074
	(0.247)	(0.224)
aa19	−0.411	−0.411
	(0.505)	(0.458)
aa20	0.129	0.129

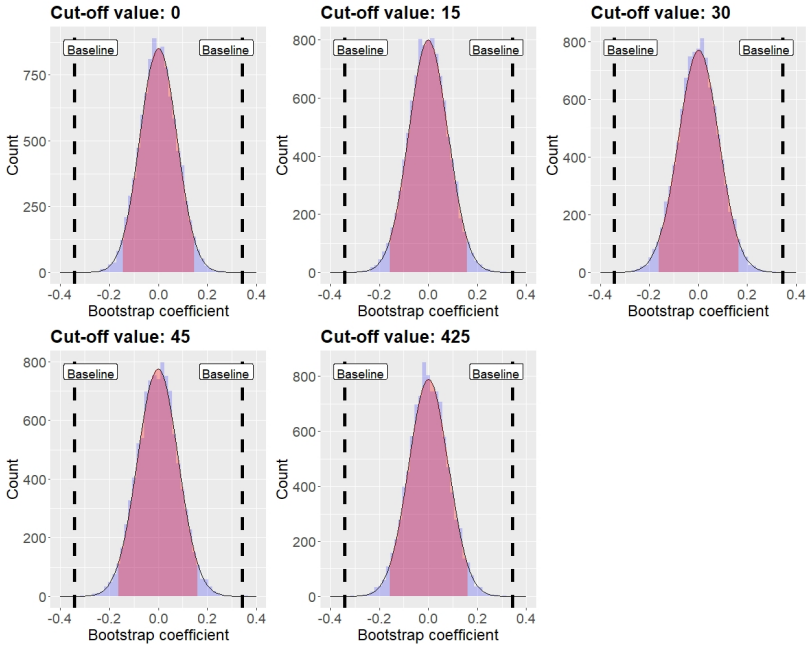
	(0.305)	(0.277)
aa21	−0.337	−0.337
	(0.256)	(0.232)
aa22	0.042	0.042
	(0.174)	(0.158)
aa23	−0.052	−0.052
	(0.251)	(0.228)
aa24	0.257	0.257
	(0.208)	(0.189)
aa25	−0.298	−0.298
	(0.424)	(0.384)
aa26	−1.070***	−1.070***
	(0.275)	(0.250)
aa27	−0.147	−0.147
	(0.265)	(0.240)
aa28	−0.385	−0.385
	(0.439)	(0.398)
aa29	−0.405	−0.405
	(0.468)	(0.424)
aa30	0.247	0.247
	(0.265)	(0.240)
aa31	−0.184	−0.184
	(0.424)	(0.384)
aa32	−0.336	−0.336
	(0.358)	(0.324)
aa33	−0.120	−0.120
	(0.343)	(0.311)
aa34	−0.015	−0.015
	(0.395)	(0.358)
aa35	1.402***	1.402***
	(0.365)	(0.331)
aa36	0.630*	0.630*
	(0.360)	(0.326)
aa37	0.342	0.342
	(0.471)	(0.427)
aa38	0.194	0.194

	(0.505)	(0.458)
aa39	0.168	0.168
	(0.188)	(0.170)
aa41	0.065	0.065
	(0.482)	(0.436)
aa42	−0.446	−0.446*
	(0.293)	(0.266)
aa43	0.235	0.235
	(0.256)	(0.232)
aa44	0.047	0.047
	(0.443)	(0.401)
aa45	0.072	0.072
	(0.289)	(0.262)
aa46	0.118	0.118
	(0.423)	(0.384)
aa47	−0.057	−0.057
	(0.187)	(0.169)
aa48	0.407	0.407
	(0.366)	(0.332)
aa49	−0.063	−0.063
	(0.211)	(0.191)
aa50	−0.285	−0.285
	(0.494)	(0.448)
Observations	920	920
R^2	0.829	0.829
Adjusted R^2	0.768	0.767
Note:	*p<0.1; **p<0.05; ***p<0.01	

Appendix B - Histograms of the Bootstrapping Results for the Different Cut-Off Scenarios



(a) Residual bootstrapping.



(b) Complete bootstrapping.

Figure 12: For 100 runs of the LDA algorithm, a bootstrapping procedure is applied with 10.000 runs to the 100 coefficient estimates $\hat{\alpha}$ for each cut-off value in the set $\{0, 15, 30, 45, 425\}$. Histograms (a) show the residual bootstrap results; histograms (b) show the complete bootstrap results. The dashed lines indicate the baseline model coefficient.

Appendix C - Comparison Between the Significance and Perplexity Case for Different Cut-Off Value

Table 11: A comparison is made between the case in which we select estimates based on the significance and the case in which we select estimates based on the perplexity. The top table shows the perplexity 1 case, and the bottom table the perplexity 2 case. The values indicate how many of the 100 runs of the algorithm are in accordance between both cases for each cut-off scenario in the set $\{0, 15, 30, 45, 425\}$.

Cut-off value: 0	Reject significance	Accept significance
Reject perplexity	7	13
Accept perplexity	0	80

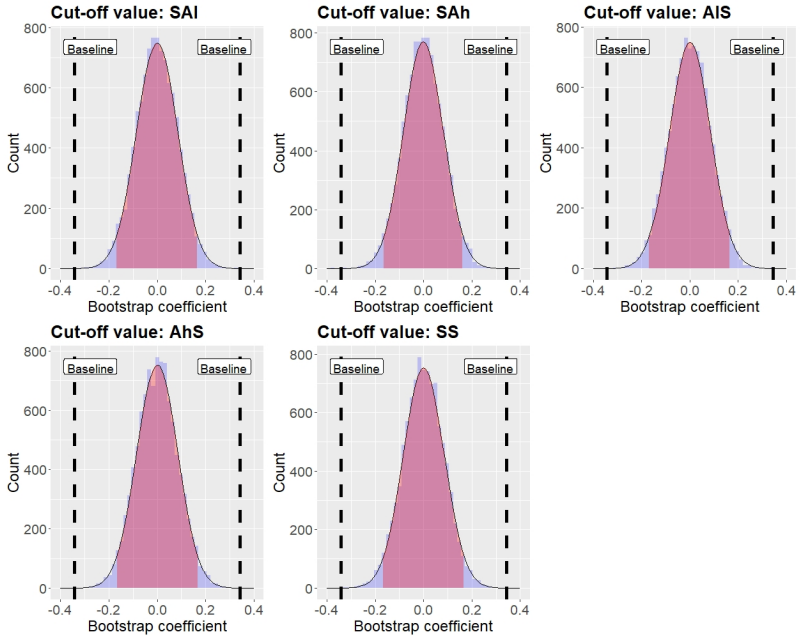
Cut-off value: 15	Reject significance	Accept significance
Reject perplexity	1	19
Accept perplexity	0	80

Cut-off value: 30	Reject significance	Accept significance
Reject perplexity	8	12
Accept perplexity	0	80

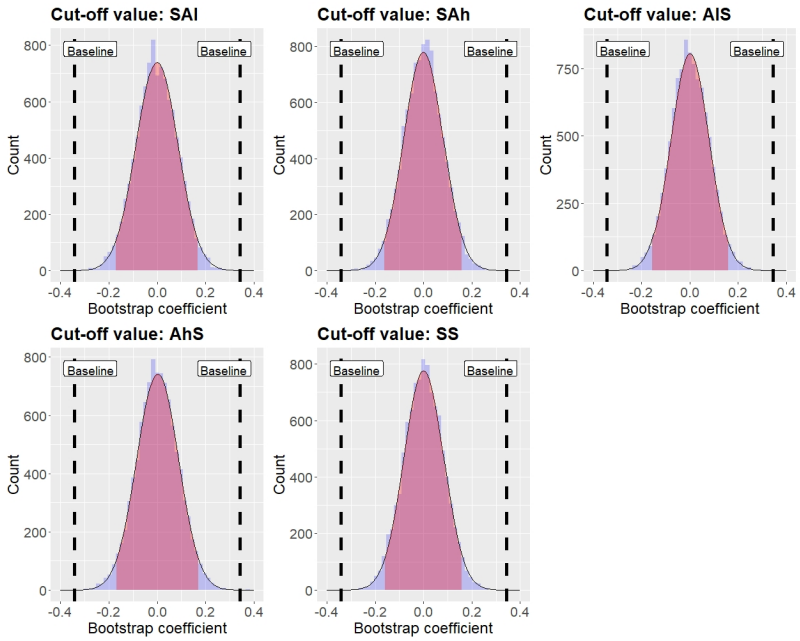
Cut-off value: 45	Reject significance	Accept significance
Reject perplexity	14	8
Accept perplexity	0	78

Cut-off value: 425	Reject significance	Accept significance
Reject perplexity	18	2
Accept perplexity	55	25

Appendix D - Histograms of the Bootstrapping Results for the Different Prior Scenarios



(a) Residual bootstrapping.



(b) Complete bootstrapping.

Figure 13: For 100 runs of the LDA algorithm, a bootstrapping procedure is applied with 10.000 runs to the 100 coefficient estimates $\hat{\alpha}$ for each prior scenario in the set $\{SA_l, SA_h, A_lS, A_hS, SS\}$. Histograms (a) show the residual bootstrap results; histograms (b) show the complete bootstrap results. The dashed lines indicate the baseline model coefficient.

Appendix E - Bootstrapping Results for the Different Prior Scenarios

Table 12: For each prior scenario in the set $\{SA_l, SA_h, A_lS, A_hS, SS\}$ the bootstrap results are presented. For both the residual and complete bootstrap procedure (10.000 runs) the 95% confidence interval and the p-value are shown.

Prior scenario	Residual 95% CI	Residual p-value	Complete 95% CI	Complete p-value
SAl	(-0.169, 0.164)	0.0001	(-0.172, 0.167)	0.0001
SAh	(-0.164, 0.162)	0.0001	(-0.163, 0.158)	0
AlS	(-0.168, 0.164)	0	(-0.155, 0.158)	0
AhS	(-0.166, 0.167)	0	(-0.169, 0.170)	0.0001
SS	(-0.168, 0.166)	0	(-0.160, 0.159)	0

Appendix F - Histograms of the Regression Estimates for Different Prior Specifications

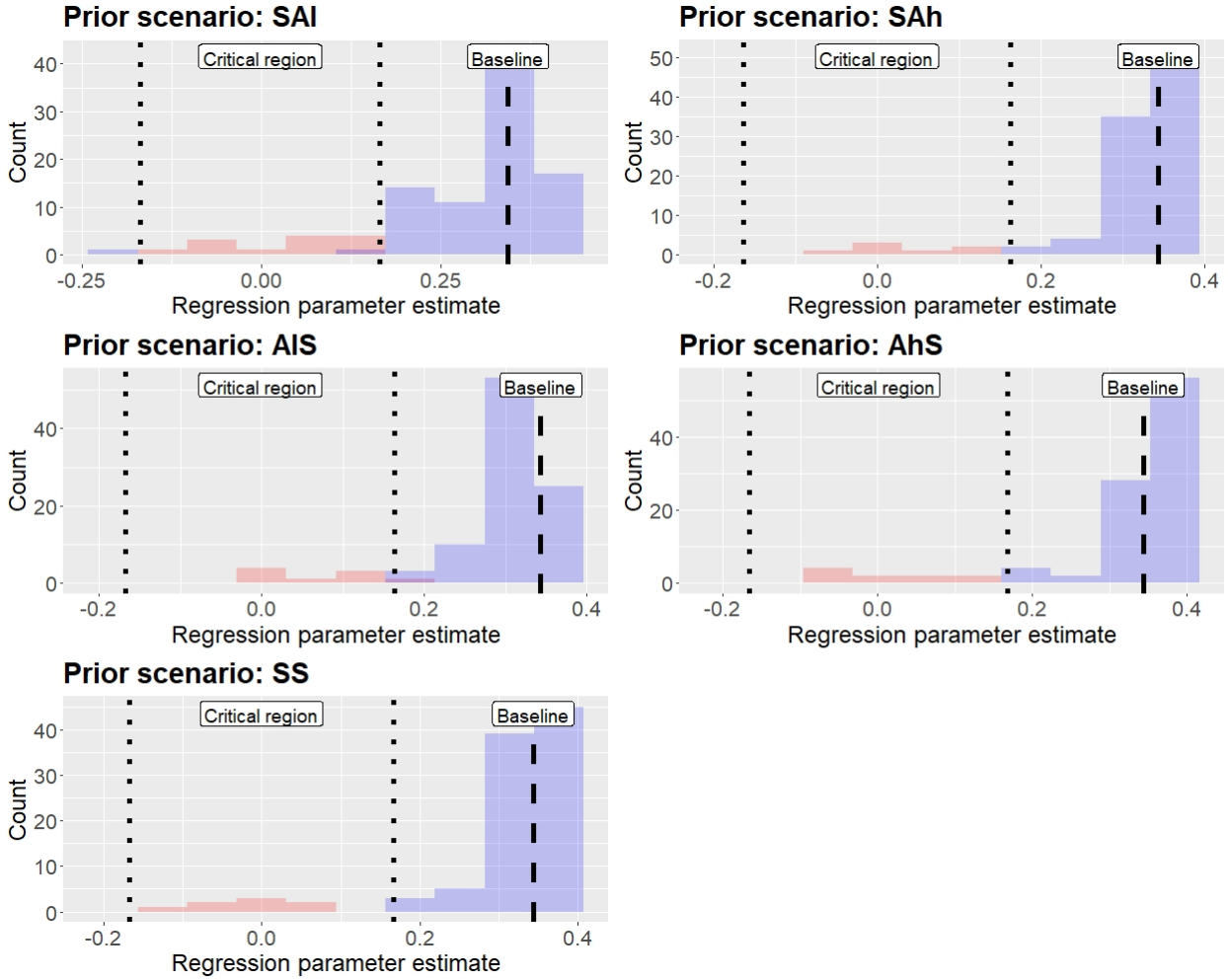


Figure 14: For 100 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined for each prior scenario in the set $\{SA_l, SA_h, A_lS, A_hS, SS\}$. Scatter plots of the 100 parameter estimates are shown with the regression estimates on the horizontal axis and the value of the perplexity on the vertical axis. The dashed line shows the baseline regression estimate. The dotted lines show the critical perplexity. This critical perplexity is chosen to best mimic the cluster of estimates that is visually observed. The estimates in blue are accepted based on the critical perplexity, whereas those in red are rejected.

Appendix G - Scatter Plots of the Regression Estimates Against the Perplexity for Different Prior Specifications

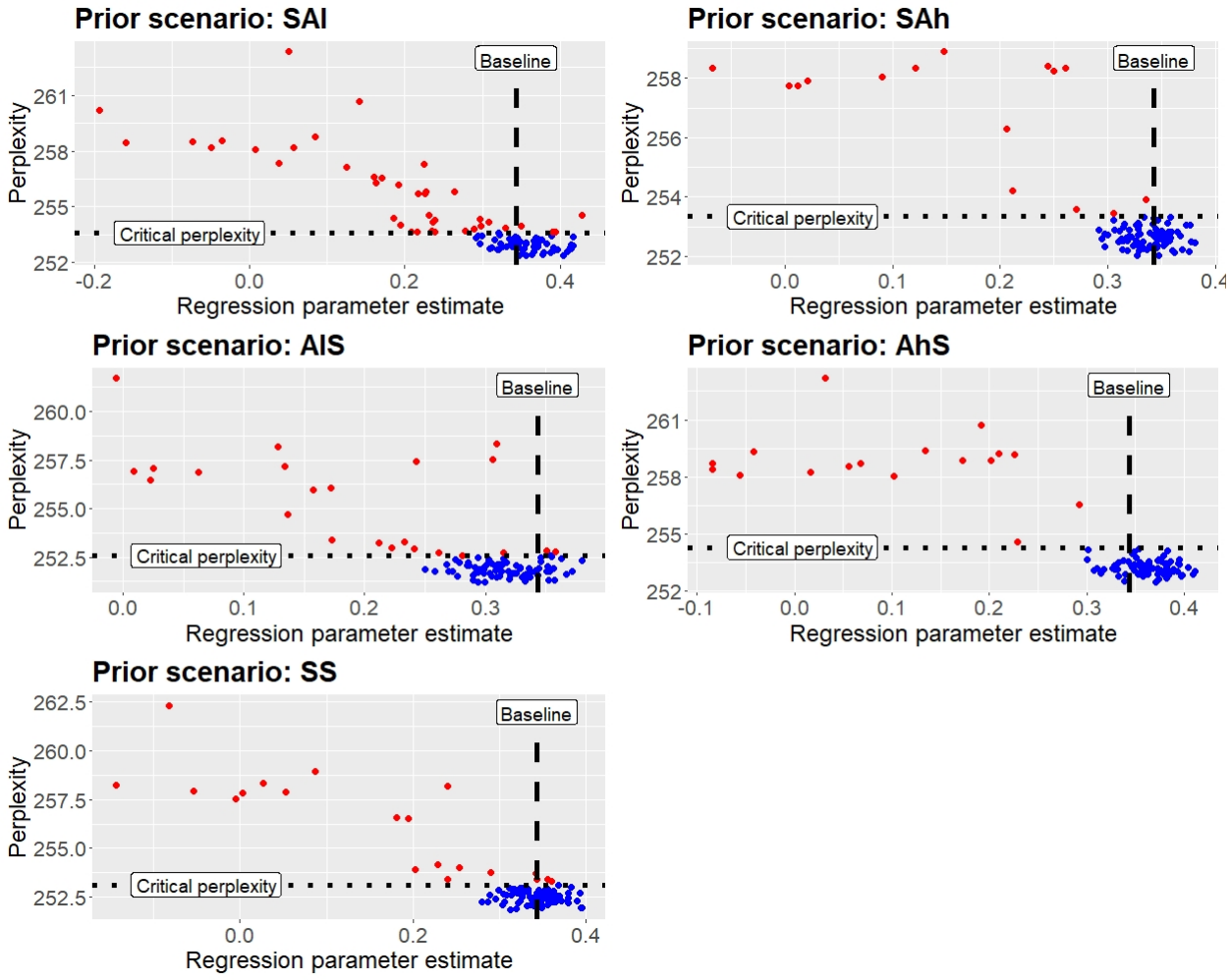


Figure 15: For 100 runs of the LDA algorithm, the regression parameter estimate $\hat{\alpha}$ (see Equation 2 for the regression equation) is determined for each prior scenario in the set $\{SA_l, SA_h, A_lS, A_hS, SS\}$. Scatter plots of the 100 parameter estimates are shown with the regression estimates on the horizontal axis and the value of the perplexity on the vertical axis. The dashed line shows the baseline regression estimate. The dotted lines show the critical perplexity. This critical perplexity is chosen to best mimic the cluster of estimates that is visually observed. The estimates in blue are accepted based on the critical perplexity, whereas those in red are rejected.

Appendix H - Comparison Between the Significance and Perplexity Case for Different Prior Scenarios

Table 13: A comparison is made between the case in which we select estimates based on the significance and the case in which we select estimates based on the perplexity. The top table shows the perplexity 1 case, and the bottom table the perplexity 2 case. The values indicate how many of the 100 runs of the algorithm are in accordance between both cases for each prior scenario in the set $\{SA_l, SA_h, A_lS, A_hS, SS\}$.

Prior scenario: SA_l	Reject significance	Accept significance
Reject perplexity	13	27
Accept perplexity	0	60

Prior scenario: SA_h	Reject significance	Accept significance
Reject perplexity	7	8
Accept perplexity	0	85

Prior scenario: A_lS	Reject significance	Accept significance
Reject perplexity	8	15
Accept perplexity	0	77

Prior scenario: A_hS	Reject significance	Accept significance
Reject perplexity	10	7
Accept perplexity	0	83

Prior scenario: SS	Reject significance	Accept significance
Reject perplexity	10	7
Accept perplexity	0	83

Appendix I - Robustness Check for the Lasso Regression: Dealing with Missing Values

Option 1: Replacing Missing Values with 0.5

Table 14: The results of two distinct regressions on a data set where missing values are set equal to 0.5. In the first two columns the results of a Lasso regression are shown, whereas the last two columns show the results of an OLS regression. The dependent variable for the Lasso regression is the firm performance and the dependent variable for the OLS regression is the baseline model CEO behavior index. Both regressions have the same independent variables: CEO diary and firm control variables. Columns (1) show the results for the basis CEO diary variables, whereas columns (2) show the results for the complete CEO diary variables. In the Lasso regression, coefficients that are equal to zero, indicate that the variable is shrunk to zero. In the OLS regression, * indicates significance at 10% level, ** indicates significance at 5% level, and *** indicates significance at 1% level.

	Lasso regression:		*OLS regression:*	
	firm_performance		CEO_behavior	
	(1) Basis	(2) Complete	(1) Basis	(2) Complete
(Intercept)	0	0	−0.000	−0.000
finance		0		−0.133***
mkting		0		−0.038
production	0	0	−0.199***	−0.258***
strategy		0.0005		0.028
hr		0.004		0.179***
bunits	0.009	0.007	0.279***	0.219***
other		−0.003		0.032
admin		0		−0.003
board		0.042		0.111***
chairman		0		−0.121***
compliance		0		−0.110
coo		0		0.139***
cao		0		−0.019
it		0		−0.215
legal		0.025		−0.008
retail		0		0.450*
groupcom	0.023	0.019	0.440***	0.309***
clients		−0.027		−0.123***

suppliers	0	0	−0.201***	−0.151***
banks		0		−0.068***
investors		0.010		−0.035*
lawyers		−0.017		−0.075***
consultants		0		−0.031
politicians		0		0.013
govoff		0		−0.089***
journalists		0		−0.093***
unions		0		−0.035
compts		0		0.010
others		0		0.162***
dealers		0		−0.049
associations		0		0.169***
pemployee		0		0.030
ins	0.012	0.009	0.113***	0.222***
out	0	0	0.212***	0.208***
n_functions		0.016		0.144**
type_business_meal		0.026		0.007
type_conference_call	0	0	0.031	−0.031
type_email		0		–
type_meeting		0		−0.211**
type_other		0		–
type_personal_family		0		–
type_phone_call	−0.014	−0.004	−0.029	−0.009
type_public_event		0		0.023
type_site_visit	−0.044	−0.043	−0.121***	−0.159***
type_travelling		0		–
type_video_conference	−0.017	−0.021	0.040*	0.008
type_working_alone		0		–
type_workrelated_leisure		0		–
level1_alone		0		–
level1_interacting		0		–
level1_personal		0		–
level1_travel		0		–
F_duration_15m		0		−0.007
F_duration_1hr		−0.005		0.037

F_duration_1hrplus		0	0.075**
F_duration_30m		0	–
F_planned_missing		0	–
F_planned_planned		0.022	0.060***
F_planned_unplanned		0	–
F_participants_blank		0	–
F_participants_missing		0.012	0.027
F_participants_one_ppl		02	0.007
F_participants_two_plus_ppl		0	–
lemp	0.775	0.757	
cons	0	0	
active	0	0	
X_Iyear_2008	0	0	
X_Iyear_2009	−0.007	−0.007	
X_Iyear_2010	0	0	
X_Iyear_2011	0.034	0.042	
X_Iyear_2012	0.049	0.052	
X_Iyear_2013	0	0	
X_Icty_2	0.150	0.153	
X_Icty_3	0.152	0.151	
X_Icty_4	0.096	0.094	
X_Icty_5	0	0	
X_Icty_6	0	0	
emp_imputed	−0.080	−0.080	
pa	0.021	0.021	
reliability	−0.011	−0.014	
ww1	0	0	
ww2	0	0	
ww3	0	0	
ww4	0	0	
ww5	0	−0.006	
ww6	0	−0.003	
ww7	0	0	
ww8	0	0	
ww9	0	0	
ww10	0.009	0.010	

ww11	0	0
ww12	0	0
ww13	0	0
ww14	0	0
ww15	0	0
ww16	0.005	0.010
ww17	0	0
ww18	0.006	0.0001
ww19	0	0
ww20	0	0
ww21	0	0.001
ww22	0.001	0.001
ww24	−0.037	−0.039
ww25	0	0
ww26	−0.012	−0.011
ww27	0	0
ww28	0	0
ww29	0	0
aa1	0	0
aa2	0	0
aa3	0	0
aa4	0	−0.001
aa5	0	0
aa6	0	0
aa7	−0.019	−0.021
aa8	0	0
aa9	0	0
aa10	0	0
aa11	0	0
aa12	0	0
aa13	0	0
aa14	0.014	0.017
aa15	0	0
aa16	0	0
aa17	−0.007	−0.004
aa18	0	0

aa19	0	−0.004
aa20	0	0
aa21	0	0
aa22	0	0
aa23	−0.001	−0.007
aa24	0	0
aa25	0	0
aa26	0	0
aa27	0	0
aa28	0	0
aa29	0	0
aa30	0	0
aa31	0	−0.0003
aa32	−0.012	−0.017
aa33	0	0
aa34	0	0
aa35	0.010	0.012
aa36	0.021	0.026
aa37	0	0
aa38	0	0
aa39	0	0
aa41	0	0
aa42	0	0
aa43	0	0
aa44	0	0
aa45	0	0
aa46	−0.001	−0.005
aa47	0	0
aa48	0.008	0.011
aa49	0	0
aa50	0	−0.005

Correlation between the LDA CEO behavior index and the Lasso CEO behavior index based on the basis variables: 0.447.

Correlation between the LDA CEO behavior index and the Lasso CEO behavior index based on the complete variables: 0.466.

Option 2: Replacing Missing Values with the Column Mean

Table 15: The results of two distinct regressions on a data set where missing values for a variable are set equal to the mean of the non-missing observations of that variable. In the first two columns the results of a Lasso regression are shown, whereas the last two columns show the results of an OLS regression. The dependent variable for the Lasso regression is the firm performance and the dependent variable for the OLS regression is the baseline model CEO behavior index. Both regressions have the same independent variables: CEO diary and firm control variables. Columns (1) show the results for the basis CEO diary variables, whereas columns (2) show the results for the complete CEO diary variables. In the Lasso regression, coefficients that are equal to zero, indicate that the variable is shrunk to zero. In the OLS regression, * indicates significance at 10% level, ** indicates significance at 5% level, and *** indicates significance at 1% level.

	Lasso regression:		*OLS regression:*	
	firm_performance		CEO_behavior	
	(1) Basis	(2) Complete	(1) Basis	(2) Complete
(Intercept)	0	0	−0.000	0.000
finance		0		−0.116
mkting		0		−0.0005
production	0	0	−0.197***	−0.215*
strategy		0.007		0.042
hr		0.008		0.199***
bunits	0.012	0.009	0.282***	0.236***
other		0		0.060
admin		0		0.013
board		0.044		0.120**
chairman		0		−0.121***
compliance		−0.001		−0.015
coo		0		0.156***
cao		0		−0.009
it		−0.008		−0.033
legal		0.031		0.004
retail		−0.020		−0.010
groupcom	0.025	0.023	0.444***	0.351***
clients		−0.024		−0.094
suppliers	0	0	−0.196***	−0.144**
banks		0		−0.059*
investors		0.017		−0.029

lawyers		−0.009		−0.056**
consultants		0		−0.027
politicians		0		0.020
govoff		0		−0.088*
journalists		0		−0.091***
unions		0		−0.026
compts		0		0.014
others		0		0.195***
dealers		0		−0.009
associations		0		0.186***
pemployee		0		0.034
ins	0.011	0.009	0.104***	0.205***
out	0	0	0.208***	0.189***
n_functions		0.011		0.076
type_business_meal		0.026		0.017
type_conference_call	0	0	0.031	−0.026
type_email		0		−
type_meeting		0		−0.206**
type_other		0		−
type_personal_family		0		−
type_phone_call	−0.016	−0.006	−0.029	−0.009
type_public_event		0		0.026
type_site_visit	−0.045	−0.043	−0.120***	−0.164***
type_travelling		0		−
type_video_conference	−0.019	−0.020	0.040	0.010
type_working_alone		0		−
type_workrelated_leisure		0		−
level1_alone		0		−
level1_interacting		0		−
level1_personal		0		−
level1_travel		0		−
F_duration_15m		0		−0.011
F_duration_1hr		−0.005		0.032
F_duration_1hrplus		0		0.069*
F_duration_30m		0		−
F_planned_missing		0		−

F_planned_planned		0.018	0.047**
F_planned_unplanned		0	–
F_participants_blank		0	–
F_participants_missing		0.015	0.017
F_participants_one_ppl		−0.002	0.006
F_participants_two_plus_ppl		0	–
lemp	0.775	0.756	
cons	0	0	
active	0	0	
X_Iyear_2008	0	0	
X_Iyear_2009	−0.008	−0.004	
X_Iyear_2010	0	0	
X_Iyear_2011	0.037	0.042	
X_Iyear_2012	0.051	0.052	
X_Iyear_2013	0	0	
X_Icty_2	0.151	0.154	
X_Icty_3	0.154	0.152	
X_Icty_4	0.098	0.094	
X_Icty_5	0	0	
X_Icty_6	0	0	
emp_imputed	−0.079	−0.081	
pa	0.023	0.020	
reliability	−0.012	−0.012	
ww1	0	0	
ww2	0	0	
ww3	0	0	
ww4	0	0	
ww5	−0.001	−0.005	
ww6	−0.0004	−0.003	
ww7	0	0	
ww8	0	0	
ww9	0	0	
ww10	0.011	0.011	
ww11	0	0	
ww12	0	0	
ww13	0	0	

ww14	0	0
ww15	0	0
ww16	0.007	0.009
ww17	0	0
ww18	0.008	0
ww19	0	0
ww20	0	0
ww21	0	0.001
ww22	0.002	0.001
ww24	−0.038	−0.039
ww25	0	0
ww26	−0.013	−0.011
ww27	0	0
ww28	0	0
ww29	0	0
aa1	0	0
aa2	0.001	0
aa3	0	0
aa4	0	−0.0004
aa5	0	0
aa6	0	0
aa7	−0.021	−0.020
aa8	0	0
aa9	0	0
aa10	0	0
aa11	0	0
aa12	0	0
aa13	0	0
aa14	0.016	0.018
aa15	0	0
aa16	0	0
aa17	−0.008	−0.005
aa18	0	0
aa19	0	−0.004
aa20	0	0
aa21	0	0

aa22	0	0
aa23	−0.003	−0.006
aa24	0	0
aa25	0	0
aa26	0	−0.0004
aa27	0	0
aa28	0	0
aa29	0	0
aa30	0	0
aa31	0	−0.003
aa32	−0.013	−0.014
aa33	0	0
aa34	0	0
aa35	0.012	0.012
aa36	0.022	0.025
aa37	0	0
aa38	0	0
aa39	0	0
aa41	0	0
aa42	0	0
aa43	0	0
aa44	0	0
aa45	0	0
aa46	−0.002	−0.005
aa47	0	0
aa48	0.009	0.010
aa49	0	0
aa50	−0.001	−0.006

Correlation between the LDA CEO behavior index and the Lasso CEO behavior index based on the basis variables: 0.456.

Correlation between the LDA CEO behavior index and the Lasso CEO behavior index based on the complete variables: 0.461.

Option 3: Removing Observations with a Missing Value

Table 16: The results of two distinct regressions on a data set where observations with missing values are removed (the data set only contains information about 236 CEOs). In the first two columns the results of a Lasso regression are shown, whereas the last two columns show the results of an OLS regression. The dependent variable for the Lasso regression is the firm performance and the dependent variable for the OLS regression is the baseline model CEO behavior index. Both regressions have the same independent variables: CEO diary and firm control variables. Columns (1) show the results for the basis CEO diary variables, whereas columns (2) show the results for the complete CEO diary variables. In the Lasso regression, coefficients that are equal to zero, indicate that the variable is shrunk to zero. In the OLS regression, * indicates significance at 10% level, ** indicates significance at 5% level, and *** indicates significance at 1% level.

	Lasso regression:		*OLS regression:*	
	firm_performance		CEO_behavior	
	(1) Basis	(2) Complete	(1) Basis	(2) Complete
(Intercept)	0	0	0.000	−0.000
finance		0		0.797*
mkting		0		0.933*
production	0	0	−0.294***	0.963
strategy		0		0.202
hr		0		0.766**
bunits	0.057	0.041	0.506***	1.128***
other		0		0.306***
admin		0		0.353
board		0.020		0.757**
chairman		0		−0.109**
compliance		0		0.346*
coo		0		0.284***
cao		0		0.378*
it		0		0.107
legal		0		0.117**
retail		−0.001		0.034
groupcom	0	0	0.283***	0.636***
clients		0		0.674
suppliers	0.017	0	−0.143***	0.496
banks		0		0.237
investors		0		0.164

lawyers		0		0.132
consultants		0		0.453*
politicians		0.007		0.029
govoff		0		0.405*
journalists		0		0.023
unions		0		–
compts		0		0.290
others		0		0.516***
dealers		0		0.339*
associations		0		0.692***
pemployee		0.006		0.212**
ins	0	0	0.035	0.258***
out	0	0	0.128	0.242
n_functions		0		−2.098*
type_business_meal		0.005		0.026
type_conference_call	0	0	0.028	0.034
type_email		0		–
type_meeting		0		0.009
type_other		0		–
type_personal_family		0		–
type_phone_call	−0.029	−0.007	−0.105**	0.045
type_public_event		0		0.226***
type_site_visit	−0.059	−0.035	−0.022	−0.031
type_travelling		0		–
type_video_conference	0	−0.020	−0.020	−0.029
type_working_alone		0		–
type_workrelated_leisure		0		–
level1_alone		0		–
level1_interacting		0		–
level1_personal		0		–
level1_travel		0		–
F_duration_15m		0		−0.071
F_duration_1hr		0		−0.035
F_duration_1hrplus		0		−0.040
F_duration_30m		0		–
F_planned_missing		0		–

F_planned_planned		0.056	0.053
F_planned_unplanned		0	–
F_participants_blank		0	–
F_participants_missing		0	–
F_participants_one_ppl		0	0.008
F_participants_two_plus_ppl		0	–
lemp	0.695	0.679	
cons	0.025	0.006	
active	0	0	
X_Iyear_2008	0	0	
X_Iyear_2009	−0.089	−0.063	
X_Iyear_2010	−0.053	−0.034	
X_Iyear_2011	0	0	
X_Iyear_2012	0.020	0.013	
X_Iyear_2013	0	0	
X_Icty_2	0	0	
X_Icty_3	0	0	
X_Icty_4	0	0	
X_Icty_5	0	0	
X_Icty_6	0	0	
emp_imputed	−0.035	−0.037	
pa	0	0	
reliability	−0.043	−0.017	
ww1	0	0	
ww2	0	0	
ww3	0	0	
ww4	0	0	
ww5	0	0	
ww6	0	0	
ww7	0	0	
ww8	0	0	
ww9	0	0	
ww10	0	0	
ww11	0	0	
ww12	0	0	
ww13	0	0	

ww14	0	0
ww15	0	0
ww16	0	0
ww17	0	0
ww18	0	0
ww19	0	0
ww20	0	0
ww21	0	0
ww22	0	0
ww24	−0.045	−0.023
ww25	0	0
ww26	−0.015	0
ww27	0	0
ww28	0	0
ww29	0	0
aa1	0	0
aa2	0.001	0
aa3	0	0
aa4	0	0
aa5	0	0
aa6	0	0
aa7	0	0
aa8	0	0
aa9	0	0
aa10	0	0
aa11	0	0
aa12	0	0
aa13	0	0
aa14	0	0
aa15	0	0
aa16	0	0
aa17	0	0
aa18	0	0
aa19	0	0
aa20	0	0
aa21	0	0

aa22	0	0
aa23	0	0
aa24	0	0
aa25	0	0
aa26	0	0
aa27	0	0
aa28	0	0
aa29	0	0
aa30	0	0
aa31	0	0
aa32	−0.021	0
aa33	0	0
aa34	0	0
aa35	0	0
aa36	0.037	0.026
aa37	0	0
aa38	0	0
aa39	0	0
aa41	0	0
aa42	0	0
aa43	0	0
aa44	0	0
aa45	0	0
aa46	0	0
aa47	0	0
aa48	0.018	0
aa49	0	0
aa50	0	0

Correlation between the LDA CEO behavior index and the Lasso CEO behavior index based on the basis variables: 0.474.

Correlation between the LDA CEO behavior index and the Lasso CEO behavior index based on the complete variables: 0.493.

Appendix J - Lasso Regression Results for the Basis Variables and for Complete Variables

Table 17: The results of two distinct regressions on a data set where missing values are set equal to zero. In the first two columns the results of a Lasso regression are shown, whereas the last two columns show the results of an OLS regression. The dependent variable for the Lasso regression is the firm performance and the dependent variable for the OLS regression is the baseline model CEO behavior index. Both regressions have the same independent variables: CEO diary and firm control variables. Columns (1) show the results for the basis CEO diary variables, whereas columns (2) show the results for the complete CEO diary variables. In the Lasso regression, coefficients that are equal to zero, indicate that the variable is shrunk to zero. In the OLS regression, * indicates significance at 10% level, ** indicates significance at 5% level, and *** indicates significance at 1% level.

	Lasso regression:		*OLS regression:*	
	firm_performance		CEO_behavior	
	(1) Basis	(2) Complete	(1) Basis	(2) Complete
(Intercept)	0	0	0.000	−0.000
finance		0		−0.262
mkting		0		−0.188
production	0	0	−0.201***	−0.430
strategy		0.006		−0.012
hr		0.008		0.092
bunits	0.012	0.008	0.282***	0.128
other		0		−0.027
admin		0		−0.066
board		0.042		0.032
chairman		0		−0.121***
compliance		−0.002		−0.052
coo		0		0.091
cao		0		−0.087
it		−0.008		−0.062**
legal		0.030		−0.031
retail		−0.021		−0.025
groupcom	0.025	0.022	0.442***	0.200
clients		−0.021		−0.243
suppliers	0	0	−0.196***	−0.239**
banks		0		−0.100*

investors		0.016		−0.061
lawyers		−0.007		−0.096*
consultants		0		−0.092
politicians		0		0.017
govoff		0		−0.148*
journalists		0		−0.087***
unions		0		−0.027
compts		0		−0.036
others		0		0.085
dealers		0		−0.051
associations		0		0.101
pemployee		0		−0.001
ins	0.014	0.011	0.118***	0.231***
out	0	0	0.216***	0.220***
n_functions		0.011		0.512
type_business_meal		0.024		0.015
type_conference_call	0	0	0.031	−0.026
type_email		0		–
type_meeting		0		−0.204**
type_other		0		–
type_personal_family		0		–
type_phone_call	−0.016	−0.005	−0.028	−0.009
type_public_event		0		0.026
type_site_visit	−0.044	−0.041	−0.115***	−0.162***
type_travelling		0		–
type_video_conference	−0.019	−0.018	0.040*	0.010
type_working_alone		0		–
type_workrelated_leisure		0		–
level1_alone		0		–
level1_interacting		0		–
level1_personal		0		–
level1_travel		0		–
F_duration_15m		0		−0.009
F_duration_1hr		−0.004		0.035
F_duration_1hrplus		0		0.072**
F_duration_30m		0		–

F_planned_missing		0	–
F_planned_planned		0.018	0.049**
F_planned_unplanned		0	–
F_participants_blank		0	–
F_participants_missing		0.013	0.030
F_participants_one_ppl		−0.002	0.006
F_participants_two_plus_ppl		0	–
lemp	0.775	0.757	
cons	0	0	
active	0	0	
X_Iyear_2008	0	0	
X_Iyear_2009	−0.008	−0.002	
X_Iyear_2010	−0	−0.058	
X_Iyear_2011	0.037	0.039	
X_Iyear_2012	0.051	0.050	
X_Iyear_2013	0	0	
X_Icty_2	0.151	0.152	
X_Icty_3	0.153	0.150	
X_Icty_4	0.097	0.092	
X_Icty_5	0	0	
X_Icty_6	0	0	
emp_imputed	−0.079	−0.077	
pa	0.023	0.019	
reliability	−0.012	−0.011	
ww1	0	0	
ww2	0	0	
ww3	0	0	
ww4	0	0	
ww5	−0.001	−0.003	
ww6	−0.001	−0.002	
ww7	0	0	
ww8	0	0	
ww9	0	0	
ww10	0.011	0.010	
ww11	0	0	
ww12	0	0	

ww13	0	0
ww14	0	0
ww15	0	0
ww16	0.007	0.007
ww17	0	0
ww18	0.008	0
ww19	0	0
ww20	0	0
ww21	0	0
ww22	0.002	0
ww24	−0.038	−0.038
ww25	0	0
ww26	−0.013	−0.010
ww27	0	0
ww28	0	0
ww29	0	0
aa1	0	0
aa2	0.001	0
aa3	0	0
aa4	0	0
aa5	0	0
aa6	0	0
aa7	−0.021	−0.018
aa8	0	0
aa9	0	0
aa10	0	0
aa11	0	0
aa12	0	0
aa13	0	0
aa14	0.016	0.016
aa15	0	0
aa16	0	0
aa17	−0.008	−0.003
aa18	0	0
aa19	0	−0.002
aa20	0	0

aa21	0	0
aa22	0	0
aa23	−0.003	−0.005
aa24	0	0
aa25	0	0
aa26	0	0
aa27	0	0
aa28	0	0
aa29	0	0
aa30	0	0
aa31	0	−0.0004
aa32	−0.013	−0.011
aa33	0	0
aa34	0	0
aa35	0.012	0.010
aa36	0.022	0.024
aa37	0	0
aa38	0	0
aa39	0	0
aa41	0	0
aa42	0	0
aa43	0	0
aa44	0	0
aa45	0	0
aa46	−0.002	−0.003
aa47	0	0
aa48	0.009	0.009
aa49	0	0
aa50	−0.001	−0.004

Appendix K - Robustness Check for the Lasso Regression: Value of Lambda

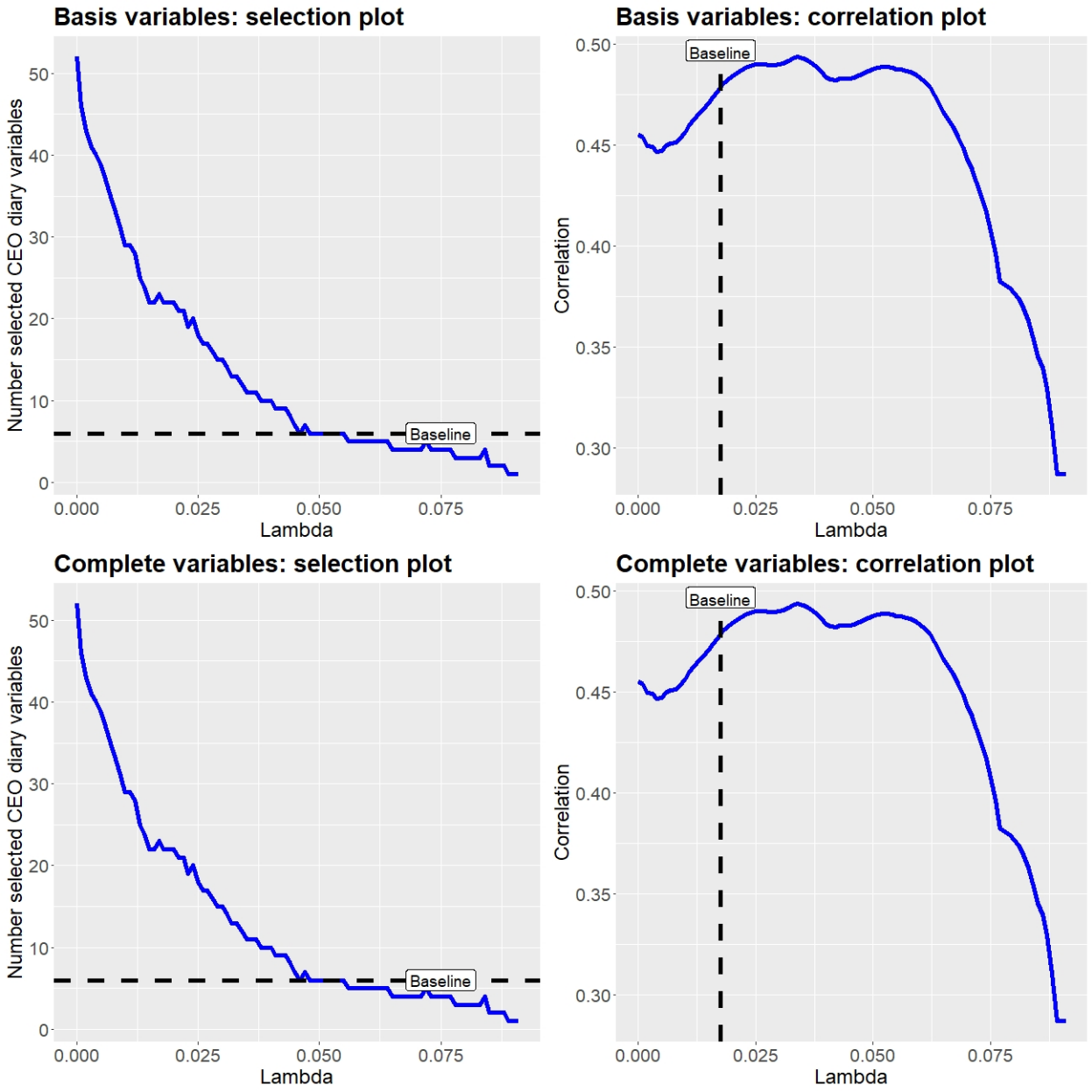


Figure 16: The first row presents two graphs with the robustness checks for the Lasso regression on the basis CEO diary variables, the second row presents two graphs with the robustness checks for the Lasso regression on the complete CEO diary variables. The first column shows two graphs that show the effect of λ on the number of the selected CEO diary variables, whereas the second column shows two graphs that show the effect of λ on the correlation between the LDA and Lasso CEO behavior indices.