Measuring Political Leaders' Professed Priorities Through their Speeches: The Case for Topic Modeling

Salfo Bikienga*
university of nebraska-lincoln, CoB 523 P.O. Box 880489 lincoln, NE 68588, united states

Matthew J. Cushing
university of nebraska-lincoln, CoB 525 P.O. Box 880489 lincoln, NE 68588, united states

Kent M. Eskridge
university of nebraska-lincoln, department of statistics, HARH 343E, lincoln, NE 68583, united states

June 25, 2018

Abstract

The role of political leadership for economic development has recently gained interest in economics. However, empirically testing the influence of leaders remains challenging since there is not an agreed-upon measure of "good political leadership". We show that quantifying leaders' professed agendas through their public statements provides a viable means for studying the role of leadership for economic development. Taking advantage of recent developments in machine learning, we apply Latent Dirichlet Allocation (LDA) to quantify the thematic contents of U.S'governors' speeches, and show a positive association between U.S governors commitment to business promotion and business expansion in their states. We further demonstrate that the thematic contents of leaders' speeches are proxies for their professed priorities. The U.S governors' State of the State Addresses (SoSAs) are used as a test case, since these speeches are by design meant to lay out the governors' priorities. Our findings illustrate the usefulness of topic modeling in analyzing political speech and suggest that speeches may be useful in identifying the role of leadership in economic development. ***(JEL O10, R50, C38)***

Keywords: Political leadership; Development; Topic modeling; Latent Dirichlet Allocation; Canonical Correlation Analysis; State of the State Addresses.

*Corresponding author, e-mail: sbikienga@huskers.unl.edu

1 Introduction

In 2010, the United Nation Commission on Growth and Development published a report on the role of political leadership for economic growth (Brady and Spence, 2010) with the goal of encouraging economists and policy practitioners to recognize the role of political leaders in economic development. That report lamented the paucity of rigorous analyses on the subject and expressed hope that the report would spur research on the role of leadership in economic growth.

Anecdotal evidences on the importance of political leaders for policy success abound. Listing what she terms "Ten lessons for successful Reforms" (lessons learned from leading Nigeria's reforms agenda of 2003-2006) Okonjo-Iweala (2012, p. 129) states that: "Even if a first-rate team is assembled, reform will not occur without the political will and support of the head of state. In fact, there is little point to embarking on reform unless there is a demand for it by the top leadership." Schneider et al. (2018) documents the instrumental role of Ecuador president, Rafael Correa, in pushing for reforms of the country's education system. That reform propelled Ecuador's student's performance in international tests from the worst to one of the best of fifteen Latin American countries.

Formal research on the role of political leaders for economic growth is sparse. Jones and Olken (2005) compares the average growth rate before and after a political leader death, and concluded that the difference is statistically significant, suggesting that leadership does matter for economic growth. Easterly and Pennings (2016) disputes the findings in Jones and Olken (2005), arguing that controlling for exogenous shocks renders their results tenuous at best. Besley et al. (2011) shows that leaders' characteristics, such as education, matter for economic growth. Blinder and Watson (2016) shows that the U.S economy tends to perform better under democratic rather than republican leadership.

A major challenge for testing the role of political leadership and economic development is the lack of an objective measure of leadership. As Easterly and Pennings (2016) points out, the designation of a 'good' or 'bad' leader is typically based on some ex post measures of success. A formal test of the role of leadership requires an ex ante measure of the qualities that make for good leadership. Similarly, Besley et al. (2011) concludes that "the exact mechanism at work in explaining how leadership matter remains opaque."

The existing literature has empirically shown that leadership matters for economic growth. However, the literature falls short in providing the mechanism by which leadership matter. The current paper explores the possibility of using the thematic contents of speeches to identify the mechanism by which leaders affect the economy. By quantifying the thematic contents of leaders' public statements, we can demonstrate whether those speeches reveal leaders' priorities and whether those priorities translate into policy actions that explain the positive role of leadership for

economic growth. Further, by establishing that priorities can be measured in this way, we make the case that topic modeling techniques are potentially useful in measuring leaders' priorities and the association of these priorities and economic growth. Topic modeling provides a systematic and informative way of quantifying political leaders expressed priorities.

The analysis of speeches to study leaders is widely used and accepted in many fields (e.g. Political Science, History, Accounting and Management). Winter (2005) notes that the "one kind of data from political leaders that is produced and preserved in abundance" is their words. Political leaders communicate their agenda, mobilize followers, and research suggests that their public statements reflect what they wants, and what they are pledging to be (Hermann, 2008). Grimmer (2010) identifies the expressed agendas of the U.S senators, using the senate Press Releases. Hermann et al. (2001) shows that political leaders' public statements can be used to study their leadership styles (for example, the goal driven leaders are persistent in what they say, whereas the opportunistic leaders tend to respond to news, and find it difficult to have a consistent message over time). In explaining the role of president Correa of Ecuador in the impressive educational reforms of his country, Schneider et al. (2018) states: "While electoral campaigns in Latin America, and elsewhere, often promise education reform, Correa kept education in the spotlight not only during the campaign but also throughout his three terms in office. Few Ecuadoreans doubted his personal conviction about the importance of education quality and equity." That is, his relentless commitment to his education reform was readily identifiable in his public statements.

Text analytics techniques offer means to systematically quantify these type of relentless commitments of political leaders through the quantification of thematic contents of their speeches over time. For instance, is a particular theme consistently present and at high proportion in a leader's important speeches? Does the consistency over time suggest a commitment to an agenda? is a leader commitment to an agenda conducive to success? Topic modeling algorithms provides a path to answering these questions. These techniques are gaining acceptance in the mainstream economics literature, (Boukus and Rosenberg, 2006; Einav and Levin, 2014; Romer and Romer, 2015; Alexopoulos and Cohen, 2015; Baker et al., 2016; Hansen and McMahon, 2016; Mullainathan and Spiess, 2017; Shapiro et al., 2017; Gentzkow et al., 2017). The use of text analytics offers an opportunity to study economic questions previously thought too nebulous to approach rigorously. Political leadership is an example of such an interesting, but difficult question to study with traditional structured data (Brady and Spence, 2010).

In sum, the goal of this paper is to show that political leaders' priorities can be measured through their public statements; by showing that the U.S. governors' priorities expressed in their SoSAs strongly correlate with their priorities expressed in the structure of their states budgets. The paper uses data on U.S governnors' SoSAs and their states' budgets because these data appear fitting for testing the ability of topic modeling to uncover leaders' professed priorities. By design,

the SoSA is used by the governors to lay out their policy priorities (Ferguson, 2006, p. 36); and in most states, the governor has full budgetary responsibility, conferring him or her the prerogative to reveal his or her priorities by shaping the state budget (Heidbreder, 2012). Thus, we hypothesize a strong correlation between the two channels through which U.S. governors display their priorities. We identify the thematic contents of governors' speeches using text analytics and machine-learning techniques, particularly, the Latent Dirichlet Allocation algorithm (also known as topic modeling). We then use the Canonical Correlation Analysis approach to test the strength of the correlation between the governors' priorities expressed in the speeches and the priorities shown in the structure of the states' budgets. The ability to measure political leaders' priorities through their speeches open a pathway to studying political leaders' commitments to development agendas and national economic progress. In fact, to motivate the relevance of such methodological approach to studying leadership and economic development, after presenting the idea of topic modeling, we illustrate its use to study the association of leaders' commitment to their economic agenda and business expansion in U.S states in Section 2. Section 3 demonstrates that topic modeling is a viable approach to measuring leaders' priorities. We discuss the main findings and their implications in Section 4, and conclude in Section 5.

2 LDA and leadership studies

We start by explaining the idea of topic modeling, then show that U.S governors' professed commitment to their economic agenda is strongly associated with business expansion in their state.

2.1 The State of the State Addresses, and topics extraction

The raw data consists of 596 State of the State addresses given by 153 U.S governors from 2001 to 2013. As is customary in the literature, the speeches are pre-processed to remove link words (such as: a, the, and, for etc.) and common English words (e.g.: during, each, some, very, their, being etc.). Further, longer words are truncated to their roots. For example, education, educated and educating, having the same root are all truncated to educ. We also remove words that are less than 4 characters, and words that are in less than 20% of the speeches (rare words). The preprocessing leaves a total of 1034 unique words. The data then consists of a matrix, W, with 596 rows representing speeches and 1034 columns giving the counts of words in each speech. Each row represents a speech and each column represents the number of occurrences of a particular word.

Because, even after preprocessing, the number of unique words remains large, some kind of dimension reduction procedure is typically deemed necessary. The traditional approach, popular in the linguistics literature, is Latent Semantic Analysis (LSA). LSA achieves the dimension re-

duction by applying Singular Value Decomposition (SVD) methods to the matrix of words counts (Deerwester et al., 1990). As such, the method is equivalent to Principal Components Analysis (PCA) methods (Shlens, 2014). Landauer et al. (2007) present theoretical linguistics arguments demonstrating that meaning can be extracted from this decomposition. In the economics literature, Boukus and Rosenberg (2006) has employed LSA to show that the themes of the Federal Open Market Committee (FOMC) minutes are correlated with current and future economic conditions[1].

Hofmann (1999) criticizes the LSA approach, noting that LSA is not based on an explicit data generative process and hence is not suitable for formal statistical inferences. Hofmann (1999, 2001) developed the Probabilistic Latent Semantic Analysis (PLSA) to address this shortcoming. The approach taken in this paper, Latent Dirichlet Allocation (LDA), can be viewed as a Bayesian approach to the PLSA model. The advantages of a Bayesian approach in the present context is that it addresses the fact that the word count matrix is typically sparse. Maximum likelihood estimates of the probability of a word being used have the unattractive property of assigning an identically zero estimate if the sample count is zero.

Section 2.2 and Appendix A contains an extended formal description of the hierarchical Bayesian matrix factorization method, LDA, used in this paper. Here we present a heuristic development of the approach and describe how the results can be interpreted. Consider decomposing the matrix of word counts, W, as follows:

$$W_{D,V} \simeq \theta_{D,K}\phi_{K,V}$$

where θ is a matrix of topic distribution over documents, and φ is a matrix of word distributions over topics. D represents the number of documents (or speeches), and V is the vocabulary list, i.e. the list of unique words. In this setting, φ contains a collection of words grouped into a small number (K) of topics and θ contains the percentage of each topic contained in each speech. Thus, the first column of θ represents the Topic 1 proportions in each of the documents. The first row of φ represents the words' relative importance for Topic 1. Sorting the first row of φ in decreasing order of the words' relative importance aids in interpreting the meaning of Topic 1. The remaining topics are treated similarly.

Parallel to PCA, in LDA, the θ matrix is the matrix of document components scores, and the φ matrix is the matrix of loadings. A major difference between PCA and LDA is that the components

[1]Factorial Analysis methods, particularly, Principal Component Analysis (PCA) methods are widely used in economics. They are used to construct index of multidimensional measures of wellbeing (Ram, 1982; Bérenger and Verdier-Chouchane, 2007; Decancq and Lugo, 2013). Along the same line, Tabellini (2010), employs among other tools, a PCA method to construct a cultural index to study the importance of culture for economic development; Temple and Johnson (1998), constructs an index of social development.

Moreover, it has been shown that theses factor models methods can help improve forcasts in time series analysis (Stock and Watson, 2011; Bai and Wang, 2016); they also provide efficient instruments (Bai and Ng, 2010; Kapetanios and Marcellino, 2010).

and the loadings values are interpreted as probability values; θ and φ are parameter estimates of probability models. It is also helpful to think of each element of the θ matrix as the value of an index, and the rows of φ as the contribution of each word to the definition of the index (see equation A.1 and A.2 in appendix A).

A few simple examples, using six documents (each, a vector of 1034 unique words), illustrate how LDA results can be interpreted. Beginning with 2 topics model, we collapse the matrix of words distribution into a matrix of two topics distribution (See Table 1 and 2).

Table 1: Example of topics distribution when K, the number of topics imposed, is 2.

	Topic.1	Topic.2
Alabama_2001_D_1.txt	0.75	0.25
Alabama_2002_D_2.txt	0.65	0.35
Alabama_2003_R_3.txt	0.26	0.74
Alabama_2004_R_4.txt	0.38	0.62
Alabama_2005_R_5.txt	0.50	0.50
Alabama_2006_R_6.txt	0.45	0.55

The first row of Table 1 shows that Topic 1 occupies about 75% of Alabama's SoSA of 2001. Topic 2 occupies about 25%. The SoSA of Alabama in 2001 and 2002 are similar (they are very high on Topic 1), and those of 2005 and 2006 are also similar. The approach provides a rigorous and quantifiable method for arriving at these conclusions. Table 2 gives the first few words of the transpose of the φ matrix. Note that in this table, the words relative weights are replaced with the words themselves. Thus, for each column, the first word is the word with the highest weight. Examining these words may be useful in interpreting the topics in Table 1.

Interpreting the individual topics found by LDA can be difficult at times. In the present case, however, identifying Topic 1 as education and Topic 2 as funding appears relatively uncontroversial. Thus, we surmise that the governor's priorities in 2001 and 2002 were education, whereas the governor's priorities in 2003 and 2004 were budgetary issues.

Next, by moving from two to three topics ($K = 3$), that is splitting the speeches into three topics, it appears that the speeches of 2001 and 2002 are not that similar, and those of 2005, and 2006 remain quite similar (see Table 3). The speeches of 2003 and 2004 appear similar with respect to Topic 2. Increasing the number of topics, i.e. the level of detailed decomposition of the speeches, allows for finer exposition of differences or similarities between speeches.

In sum, these examples highlight two points: 1) setting the number of topics determines the level of detail: the higher the number of topics, the higher the level of granularity; and 2), examining topics proportions reveals differences and similarities between speeches or between governors.

Table 2: Descending, ordered list of the most weighted words for each topic

Topic 1	Topic 2
school	budget
educ	fund
work	govern
help	peopl
econom	million
children	work
famili	make
health	public
busi	propos
nation	servic
make	chang
creat	program
student	know
teach	spend
invest	come

Table 3: Example of topics distribution when K, the number of topics imposed, is 3.

	Topic.1	Topic.2	Topic.3
Alabama_2001_D_1.txt	0.13	0.29	0.58
Alabama_2002_D_2.txt	0.11	0.47	0.41
Alabama_2003_R_3.txt	0.48	0.42	0.10
Alabama_2004_R_4.txt	0.37	0.39	0.24
Alabama_2005_R_5.txt	0.28	0.41	0.31
Alabama_2006_R_6.txt	0.30	0.37	0.33

Of course, this leaves open the questions of whether these differences reflect differences in policy priorities and whether these professed priorities translate into actions. Section 3.2.1 addresses these questions.

2.2 Latent Dirichlet Allocation: the data generative process

This section provides a rigorous exposition of topic modeling and presents the data generative process used to estimate the θ and φ matrices. LDA is a generative model that represents documents as being generated by a random mixture over latent variables called topics (Blei et al., 2003). A topic is defined as a distribution over words. For a given corpus (a collection of documents) of D documents each of length N_d , the generative process for LDA is defined as follows:

1. For each topic k, draw a distribution over words $\phi_k \sim Dirichlet(\beta)$ with $k = \{1,2,...K\}$
2. For each document d:
   (a) Draw a vector of topic proportions $\theta_d \sim Dirichlet(\alpha)$
   (b) For each word i
      i. Draw a topic assignment $z_{d,n} \sim multinomial(\theta_d)$ with $z_{d,n} \in \{1,2,...,K\}$
      ii. Draw a word $w_{d,v} \sim multinomial(\phi_{k=z_{d,n}})$ with $w_{d,v} \in \{1,2,...,V\}$

The above generative process allows us to construct an explicit joint likelihood of the observed and hidden variables. Markov Chain Monte Carlo (MCMC), or Variational Bayes methods can then be used to estimate the parameters θ and φ (See Blei et al., 2003; Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2007; Blei, 2012 for further exposition of the method). We derive the posterior distribution of the θ_s and ϕs in appendix A.1. Practical tools for estimating the topics distributions of a corpus exist (see Grun and Hornik (2011); Silge and Robinson (2017, chap. 6))[2].

2.3 Applying LDA for political leadership studies

Before demonstrating that topic modeling can be used to measure leaders' professed priorities, we illustrate the usefulness of topic modeling for leadership studies by applying LDA to explore the association between the consistency with which U.S governors talk about the economy and business expansion in their states.

[2]The tools we refer to are based on the R statistical software; mostly because we are most familiar with these tools. But, similar tools exist for Python users. Mallet is another popular software, and can be found at: http://mallet.cs.umass.edu/topics.php

2.3.1 Leaders' consistency measure

We postulate that talking persistently about an issue is a sign of commitment to an agenda, which we term professed agenda after Grimmer (2010). To capture that idea of professed agenda, we use the log of the inverse of the coefficient of variation. Formally,

Let $\theta_{i,j,l}$ be the relative share of topic j, ($j = \{1,2..K\}$), in governor i, ($i = \{1,2,...,N\}$) speech at year l ($l = \{1,2,...,L\}$). Then

$$\bar{\theta}_{i,j} = \frac{\sum_{l=1}^{L} \theta_{i,j,l}}{L}$$

gives an idea of the overall importance of topic j in governor i combined speeches.

$$s_{i,j} = \sqrt{\frac{\sum_{l=1}^{L} (\theta_{i,j,l} - \bar{\theta}_{i,j})^2}{L-1}}$$

gives the level of variations of topic j in governor i speeches; and

$$C_{i,j} = log(\frac{\bar{\theta}_{i,j}}{s_{i,j}})$$

is what we term consistency measure.

The intuition of our consistency measure is that consistency implies low variance. That idea is captured in the formula by having $s_{i,j}$ in the denominator. However, a low variance alone is not enough to conclude that a particular topic or theme is important for a governor. Hence, we use the mean $\bar{\theta}_{i,j}$ in the numerator. Consequently, our consistency measure is high when on average the governor talks a lot about the topic with low variations from year to year. We take the log of the ratio to temper the effect of potential outliers.

Figure 1 shows the distribution of the consistency measure by Topic[3]. The vertical line in each histogram indicates the mean of the consistency measure.

2.3.2 The outcome variables: establishment net entry rate

The dependent variable was collected from the US Census Bureau, Business Dynamics Statistics website. The annual business establishment entry rate (we are using the net entry rate) was computed using the following formula:

$$entry\_rate_{t,s} = \frac{estab_{t,s} - estab_{t-1,s}}{estab_{t-1,s}} \times 1000,$$

[3]Histograms are used instead of a summary table to show the distribution of the variables

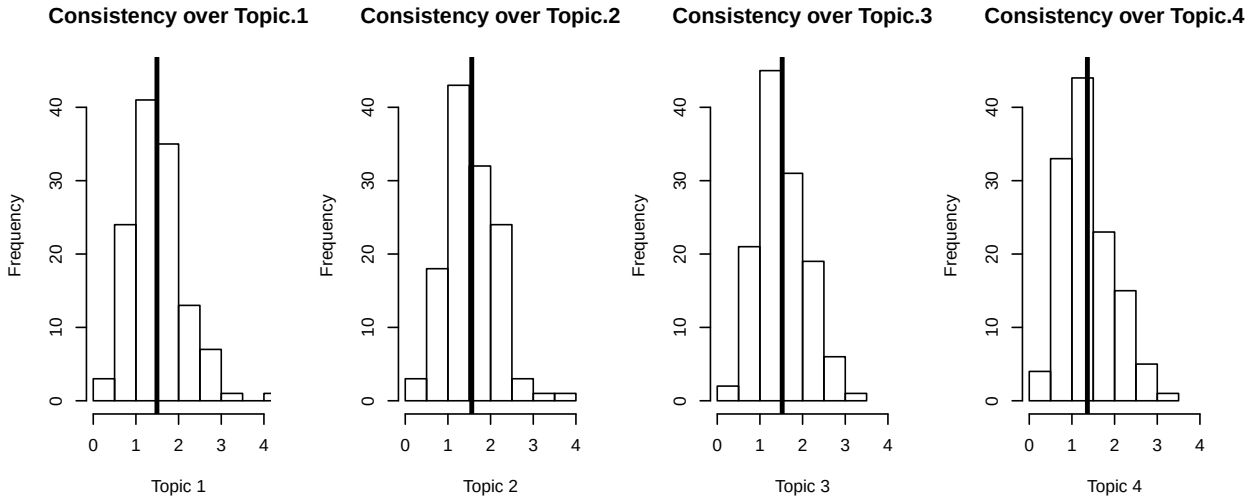


Figure 1: Distributions of the consistency measure for K = 4 topics

where:

- $entry\_rate_{t,s}$ is the business establishment net entry rate at year t in state s.
- $estab_{t,s}$ is the total number of business establishments at year t in state s.
- $estab_{t-1,s}$ is the total number of business establishments at year t-1 in state s
- The entry rate was multiplied by 1000 to scale up the number, for convenience.

Similarly, we compute the US net entry rate as:

$$usa\_entry\_rate_{t,s} = \frac{usa\_estab_t - usa\_estab_{t-1}}{usa\_estab_{t-1}} \times 1000,$$

We postulate that if a governor action has to have an impact in the economy, that impact will be observed in the future. So, we match 2001 SoSA with 2002 establishment net entry rate, which we refer to as one period lead. We call it two period lead if we match 2001 SoSA with 2003 net entry rate, and three period lead if we match 2001 SoSA with 2004 net entry rate.

The data are aggregated by governor's term of four years. Therefore, for a governor of 2001 to 2004, the one period lead average net entry rate is computed as follows:

$$E_i = \frac{\sum_{t=2002}^{2005} NetEntryRate_t}{4}$$

From an initial data set of 596 observations, the aggregated data consists of 125 observations, where a governor of a single term is the unit of observation. We chose a term as the unit of observation because political leaders tend to shift their focus from term to term.

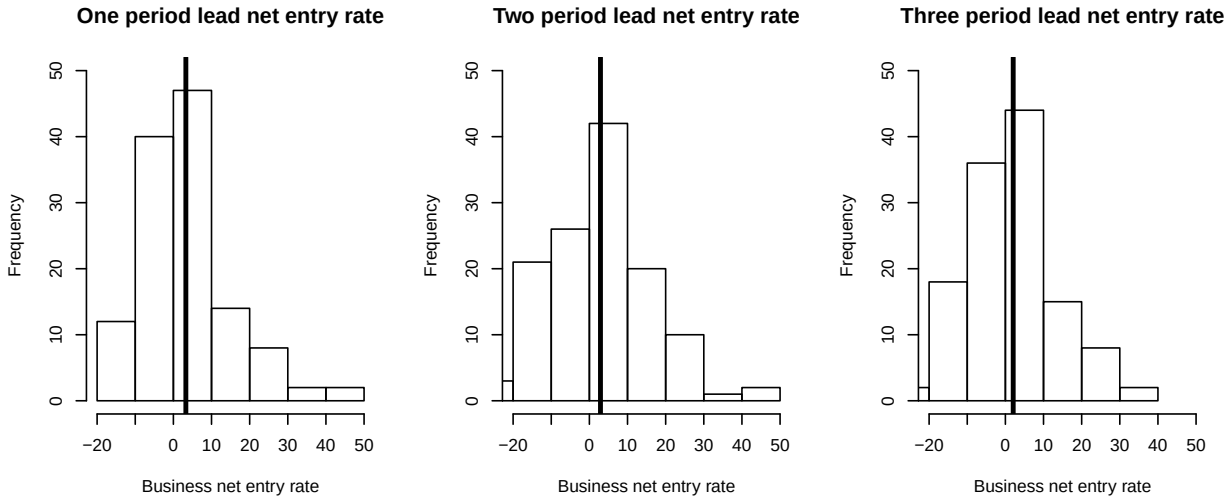


Figure 2: Distribution of establishment entry rate. From left to right, the distribution are for one, two, and three period leads.

Fig. 2 (From left to right) shows the distributions of the average net business entry rate for one, two and three period lead.

K is set to 4, after applying a model selection criterion. There is not a statistically satisfactory method for selecting K. "In practice, however, it is very common to simply start with a number of topics on the order of ten, and then adjust the number of topics in whatever direction seems to improve interpretability."(Gentzkow et al., 2017, p.18). We found $K = 4$ to be a reasonable choice for analyzing the association between our consistency measure and average business entry rate (see Appendix B.1 for an explanation on how we arrive at choosing $K = 4$).

2.3.3 Governors' agendas and average business net entry rate

The basic model regresses the average business net entry rate (E_i) on the consistency over topics $C_{i,j}$, controlling for the U.S entry rate, as follows:

$$
\begin{aligned}
E_i &\sim Normal(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 E_{US_i} + \sum_{j=1}^{K} \tau_j C_{i,j} \\
\beta_0, \beta_1, \tau_j &\sim Normal(0, 100) \\
\sigma &\sim Uniform(0, 25)
\end{aligned}
$$

where $K = 4$ is the number of themes, and E_{US_i} is the average net business establishment entry

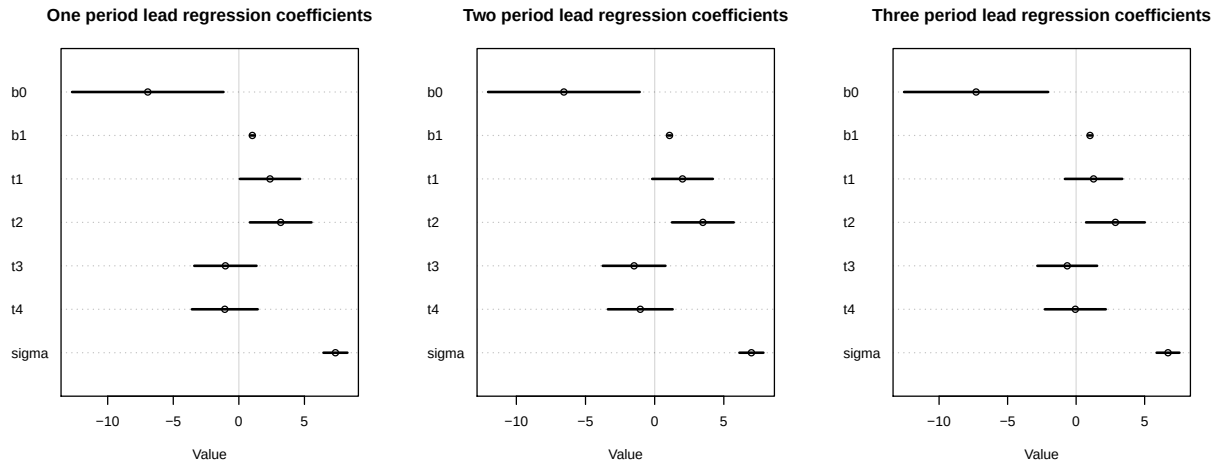


Figure 3: Plots of the regression posterior estimates.

rate of the US during governor i's term. E_{US_i} accounts for the state of the US economy.

We employ a Bayesian regression approach with noncommittal broad priors on the parameters β_0, β_1, τ_j, and σ so that the priors have minimal influence on the posteriors. These assumed prior distributions yield posterior mean estimates very similar to the Maximum Likelihood Estimates (MLE). We opted for a Bayesian regression to get the full posterior distributions of the parameters (see Appendix B.2). Thus, instead of point estimates given by OLS, or MLE, Bayesian regression gives a range of credible parameters estimates consistent with the observed data. The Quadratic Approximation (Maximum a Posteriori) estimation method is used to compute the posteriors estimates of the model parameters. For simple models like ours, this approach works very well, and often yields the exact posterior distribution (McElreath, 2016, p.41).

Fig. 3 shows the plot of the regression coefficients. The results suggest that only Topic 2 is strongly associated with average net business entry rate; this association is positive and the 95% high posterior density interval (HPDI) does not contain zero. The results are similar whether the outcome variable is the one, the two, or the three period lead. The results indicate that the consistency measure of Topic 2 captures something meaningful in explaining the variations in the establishment entry rate variables. What is Topic 2 referring to? Table 4 suggests that Topic.2 refers to an economic agenda, as words related to investment, building, creating, producing, energy, company, business, and economy are prominent in conferring a meaning to topic 2.

Topic 1 is marginally strong and is positively associated with business entry rate, particularly for the one and two period leads. Topic 1 refers to budgetary issues, with expansionary tone, as the words such as increase, fund, provide would suggest. Topic 3 and 4 are weakly and negatively associated with business entry rate. While Topic 4 can be interpreted as a combination of education and health care, it is not clear what Topic 3 is. For large K, it is customary for LDA to yield a

"residual", uninterpretable, topic.

Table 4: List of words, ranked by their relative importance for their respective topics. The list is used to infer the meaning of the topic.

Topic 1	Topic 2	Topic 3	Topic 4
fund	econom	peopl	school
budget	busi	govern	educ
million	work	make	children
increas	creat	work	health
propos	energi	know	teach
program	develop	come	student
servic	nation	just	famili
govern	futur	governor	help
provid	build	chang	make
dollar	help	like	care
revenu	invest	reform	high
system	compani	look	program
legisl	peopl	money	nation
educ	produc	good	colleg
depart	communiti	right	work

In light of Fig. 3 and Table 4, we conclude that there is a strong positive association between the U.S governors' commitment to their economic agendas (as measured by the consistency with which they talk about the economy) and the rate of business entry in their states.

Fig. 4 shows the relationship between consistency on Topic 2 (or economic issues) and net entry rate of business establishments[4].

[4]Note that there are two high leverage values on the plot (the two values at the extreme right of the plot). Removing them would have increased the regression coefficient of Topic 2; but we have no reason to remove them.

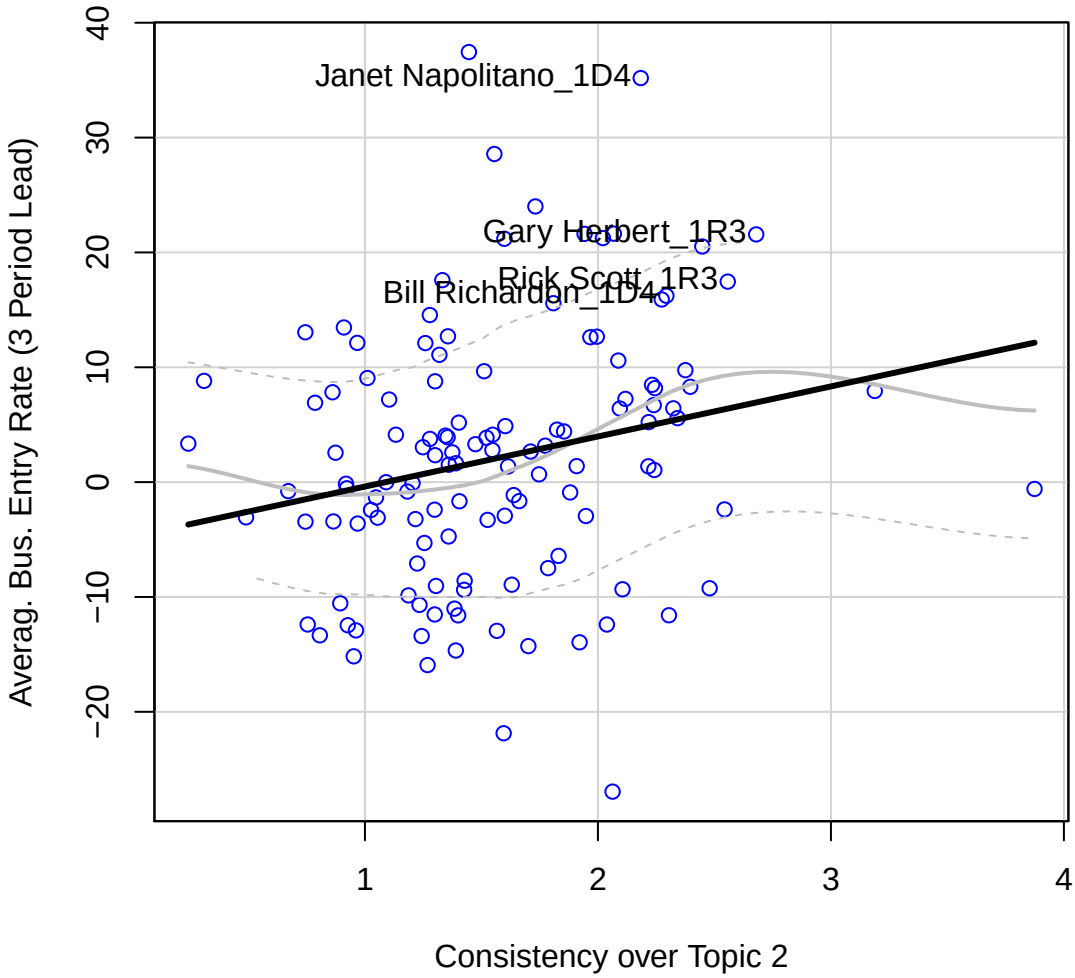


Figure 4: Scatter plot of consistency over topic.2 (economy theme) and average business net entry rate.

A major lacking in the current literature on leadership and economic growth is our inability to explain the mechanism by which leaders may affect the economy. An advantageous feature of topic modeling is that we can refer to the leaders' speeches to identifies what the governors profess to be doing. The excerpts (Fig. 5, and 6) highlight the content of four speeches by four "high achieving" governors, whom we identified on the scatterplot in Fig. 4 (Janet Napolitano of Arizona, Gary Herbert of Utah, Rick Scott of Florida, and Bill Richardson of New Mexico). The excerpts show evidences of policy actions, which are indeed expected to increase the number of establishments in the states if they are successful.

Figure 5: Excerpt of a few state' State of State Addresses (1)

(a) Excerpt of the state of state address (Arizona 2003)

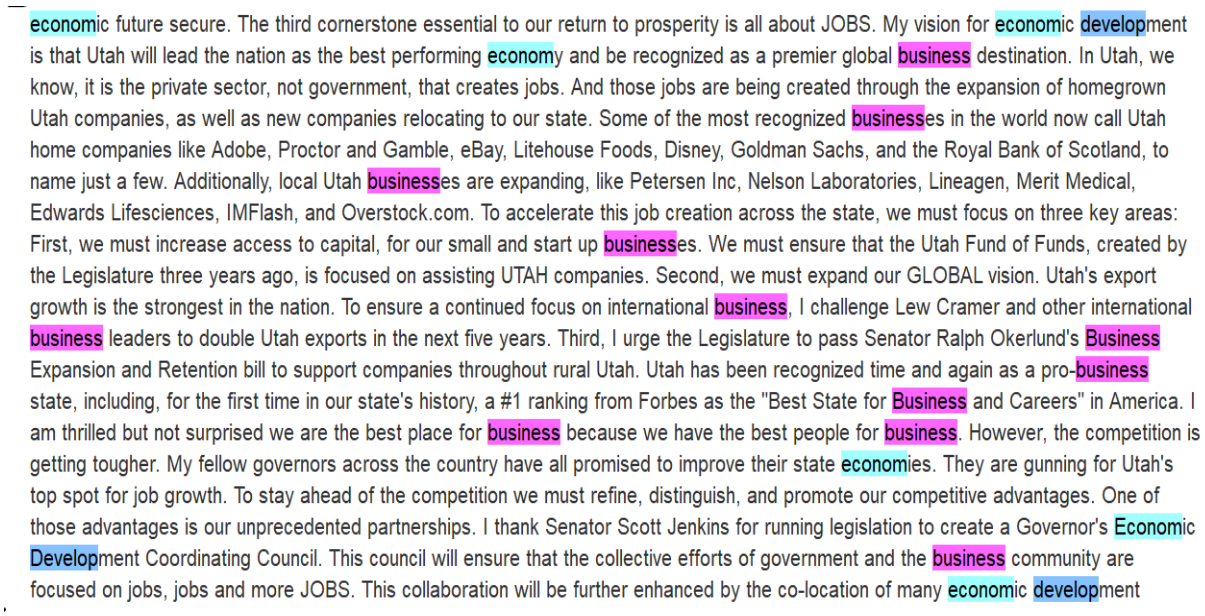
educate the workforces of tomorrow, and their research expands our horizons. Priority 2 Building the new Arizona economy Let's now turn to building the new Arizona economy. Although our unemployment rate is below the national average, we do not offer enough meaningful, high-paying, jobs for our people, whose per capita income ranks 37th among the states. It is time to coordinate our efforts to develop Arizona's promising tech sector. And it is time to take full advantage of our geographic proximity to establish Arizona as America's premier portal to trade with Latin America. Develop tech industries To achieve the economic renaissance I envision, our economy must be powered by innovation, and be driven by the entrepreneurs and tech-based businesses that will create the high-wage jobs and clean industries we seek. Three steps are key. First, our public and private sectors must speak with one economic voice. To do this, I will sign an executive order this week creating the Governor's Council on Technology and Innovation. It will focus on three areas: coordination of technology transfer from universities to the commercial sector, capital formation, and infrastructure development. Second, it is time to remove the single biggest obstacle to smooth technology transfer from our university campuses to the commercial sector. Arizona's constitution prohibits universities from forming or taking equity positions in commercial ventures, which slows down their efforts to convert research innovations into viable commercial applications. Competing states do not have this prohibition, and they enjoy greater success in luring tech start-ups. I will submit to you a ballot referendum to repeal this article of Arizona's constitution. Third, we must do a better job at attracting capital for small and growing high tech businesses. They need this capital to grow their enterprises and create more high-paying jobs for an educated workforce. I will work with the business community to attract more development capital to Arizona, particularly for new companies. International trade My administration will not limit its economic development to Arizona alone. Though many believe that our

(b) New Mexico 2003

and its entrepreneurs who play such a role in New Mexico's economic future. This administration is going to beat the bushes from coast to coast, from Europe to the Pacific Rim, seeking quality companies in need of good workers and a great environment into which to expand their companies. But, our economic development strategy doesn't depend on industrial recruitment alone. We must make the state of New Mexico a hospitable place for entrepreneurs to start and grow their business ideas. Whether from outside or from homegrown visionaries, ground-up business development must be encouraged and nurtured by state policy. I have already said I intend to spend at least 25 percent of my time working on economic development. We will undertake these efforts at many levels and with many concepts. Growing tourism and trade with Mexico is a theater of development with great promise particularly with our sister border state of Chihuahua. In cooperation with Chihuahua Gov. Patricio Martinez, I will work to build up the business and cultural bonds between our people. As New Mexicans learn more about Chihuahua, more will wish to visit and do business there. The same is true the other way to New Mexico. Here at home, we need to strengthen the teaching of business and entrepreneurial skills in our schools. Our young people must be taught the basics of business risk and reward so that more of them will take to improving their lives and building the economy of our state from within. Small business incubator programs in the population centers of the state must be strengthened and improved. Access to capital is critical, and because we lie so far from the money centers of the coasts, we remain below the radar of much of the venture capital market. The Legislature invested $10 million in the New Mexico Small Business Investment Corp., but the money has languished in a bank account for lack of an implementation strategy. I will build that implementation strategy. I further propose that we invest up to $200 million just 2 percent of the total in the state's permanent funds in New Mexico businesses. This will jump-start an entrepreneurial arm of New Mexico's economy. We will work with existing grass-roots business startup organizations such as Accion, Wesst Corp., the New Mexico Community Development Fund. With our new state investment officer, we have ensured the best expertise to manage and control the commitment of state venture funds. We will partner with private capital. While the primary purpose of our state's permanent funds must always be to provide revenue to state government, we must also invest them where fiscally prudent to create jobs and diversify our economy. By stepping up with cash, we will send a signal that New Mexico is serious about business and willing to put our money where our recruitment is. To facilitate all these ambitious development goals, I ask the Legislature for seed money: I would like to add $15 million to the in-plant training fund, bringing it to $20 million when combined with existing funds. $3 million to fun a nonprofit corporation to recruit and market new businesses and jobs. We must tap the skills and leverage the efforts of everyone to grow the economy. $9 million in a one-shot expenditure

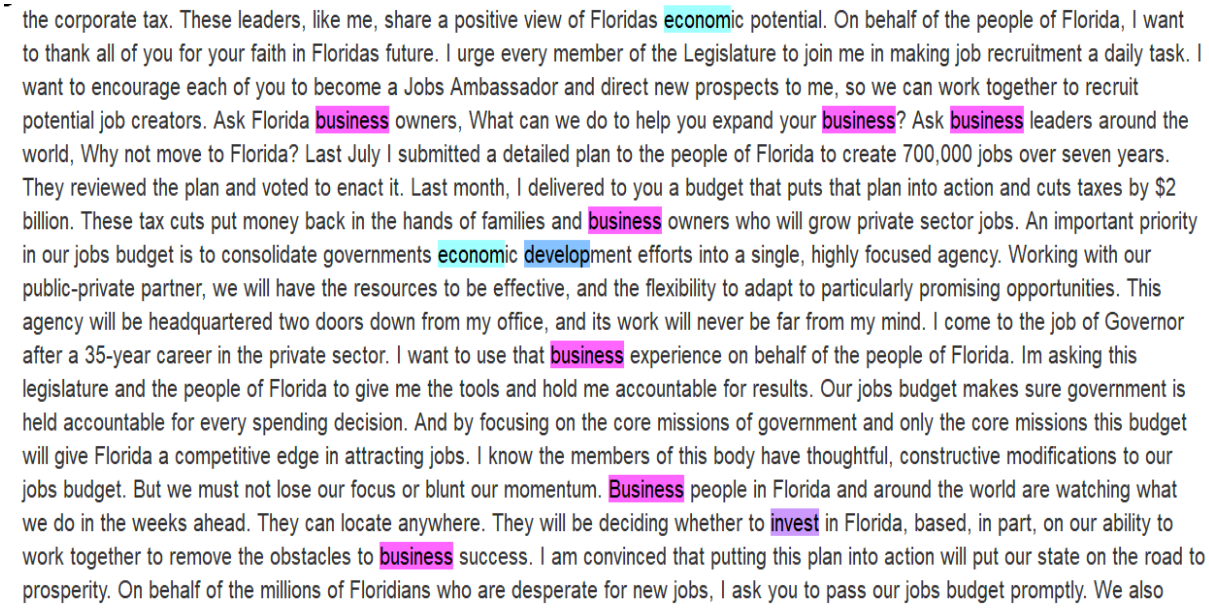
Figure 6: Excerpt of a few state' State of State Addresses (2)

(a) Utah 2011



economic future secure. The third cornerstone essential to our return to prosperity is all about JOBS. My vision for economic development is that Utah will lead the nation as the best performing economy and be recognized as a premier global business destination. In Utah, we know, it is the private sector, not government, that creates jobs. And those jobs are being created through the expansion of homegrown Utah companies, as well as new companies relocating to our state. Some of the most recognized businesses in the world now call Utah home companies like Adobe, Proctor and Gamble, eBay, Litehouse Foods, Disney, Goldman Sachs, and the Royal Bank of Scotland, to name just a few. Additionally, local Utah businesses are expanding, like Petersen Inc, Nelson Laboratories, Lineagen, Merit Medical, Edwards Lifesciences, IMFlash, and Overstock.com. To accelerate this job creation across the state, we must focus on three key areas: First, we must increase access to capital, for our small and start up businesses. We must ensure that the Utah Fund of Funds, created by the Legislature three years ago, is focused on assisting UTAH companies. Second, we must expand our GLOBAL vision. Utah's export growth is the strongest in the nation. To ensure a continued focus on international business, I challenge Lew Cramer and other international business leaders to double Utah exports in the next five years. Third, I urge the Legislature to pass Senator Ralph Okerlund's Business Expansion and Retention bill to support companies throughout rural Utah. Utah has been recognized time and again as a pro-business state, including, for the first time in our state's history, a #1 ranking from Forbes as the "Best State for Business and Careers" in America. I am thrilled but not surprised we are the best place for business because we have the best people for business. However, the competition is getting tougher. My fellow governors across the country have all promised to improve their state economies. They are gunning for Utah's top spot for job growth. To stay ahead of the competition we must refine, distinguish, and promote our competitive advantages. One of those advantages is our unprecedented partnerships. I thank Senator Scott Jenkins for running legislation to create a Governor's Economic Development Coordinating Council. This council will ensure that the collective efforts of government and the business community are focused on jobs, jobs and more JOBS. This collaboration will be further enhanced by the co-location of many economic development

(b) Florida 2011



the corporate tax. These leaders, like me, share a positive view of Floridas economic potential. On behalf of the people of Florida, I want to thank all of you for your faith in Floridas future. I urge every member of the Legislature to join me in making job recruitment a daily task. I want to encourage each of you to become a Jobs Ambassador and direct new prospects to me, so we can work together to recruit potential job creators. Ask Florida business owners, What can we do to help you expand your business? Ask business leaders around the world, Why not move to Florida? Last July I submitted a detailed plan to the people of Florida to create 700,000 jobs over seven years. They reviewed the plan and voted to enact it. Last month, I delivered to you a budget that puts that plan into action and cuts taxes by $2 billion. These tax cuts put money back in the hands of families and business owners who will grow private sector jobs. An important priority in our jobs budget is to consolidate governments economic development efforts into a single, highly focused agency. Working with our public-private partner, we will have the resources to be effective, and the flexibility to adapt to particularly promising opportunities. This agency will be headquartered two doors down from my office, and its work will never be far from my mind. I come to the job of Governor after a 35-year career in the private sector. I want to use that business experience on behalf of the people of Florida. Im asking this legislature and the people of Florida to give me the tools and hold me accountable for results. Our jobs budget makes sure government is held accountable for every spending decision. And by focusing on the core missions of government and only the core missions this budget will give Florida a competitive edge in attracting jobs. I know the members of this body have thoughtful, constructive modifications to our jobs budget. But we must not lose our focus or blunt our momentum. Business people in Florida and around the world are watching what we do in the weeks ahead. They can locate anywhere. They will be deciding whether to invest in Florida, based, in part, on our ability to work together to remove the obstacles to business success. I am convinced that putting this plan into action will put our state on the road to prosperity. On behalf of the millions of Floridians who are desperate for new jobs, I ask you to pass our jobs budget promptly. We also

To summarize, the goal of this section 2.3 was to provide a use case of topic modeling for studying political leadership. We have shown that the U.S governors' commitment to their economic agenda, as measure by the consistency with which they address economic issues is strongly and positively associated with high business expansion in their states. Moreover, by identifying economically high achieving governors, we got a glimpse at policy actions they undertook. Thus,

instead of studying the associations between leaders characteristics and economic growth as in Jones and Olken (2005), and Besley et al. (2011), topic modeling approach to studying leadership offers a window for identifying the actions taken by leaders to achieve economic success, as the excerpts in Fig. 5 and 6 show.

3 LDA topics as proxies for leaders' priorities

After presenting a use case of topic modeling for leadership studies in Section 2.3, we now demonstrate why the topics identified by topic modeling algorithms are good proxies for leaders' professed agendas, or priorities. To do so, we study the association between topics discovered through LDA and state expenditures.

The states' expenditures data were collected from the Census Bureau website, the State Government Finances page. The selected spending variables are: expenditures on education, health care, public welfare, and spending on highways. Because the spending scales are different across states, we use the *z-scores* computed by state. Moreover, to remove the time trend of state spending, the *z-scores* were linearly detrended to get the fluctuations around the linear trend. Formally, the education expenditure variable (*z-scores*) was constructed as follows:

$$Educ\_expend_{i,t} = \frac{X_{i,t} - \bar{X}_i}{s_{X_i}}$$

where $X_{i,t}$ is the state i spending in education in year t, $\bar{X}_i$ is the state i average spending on education for the study period. s_{X_i} is the state i standard deviation of the variable X. The so constructed variables were linearly detrended to remove the time trend.

All the expenditure variables were constructed similarly. Fig. 7 shows how the spending on education (detrended values of z-scores) has evolved over time in a selected set of states. The colors indicate changes in terms (four year term), or governors. It can be noted from the graph that Florida's spending on education has increased steadily under governor Jeb Bush, then fluctuated downward after his tenure. In New Mexico, education experienced a budget increase during the first term of Governor Bill Richardson, then decreased since the beginning of his second term. It appears that the governor interest in education has diminished after his first term. In fact, while education issues occupied a disproportionate high share (on average) of his SoSAs during his first term, education occupied a disproportionate low share of his SoSAs during his second term. New York education system experienced a budget increase under governor George Pataki two terms then became volatile before decreasing persistently after his tenure. Oklahoma has had a steady increase of its education budget during governor Brad Henry first term and the first half of his second term, before falling steadily since 2009. Virginia experienced a persistent increase of its

education budget from 2003 to 2008, then fell drastically in 2009, and has remained low since then.

For the analysis that follows, we use the entire sample from 2001 to 2013, the longest time span for which we have a full range of SoSAs. Note that the speech is given at the beginning of each year, whereas the expenditures refer to fiscal year accounting of state spending, which start generaly in July first. The speech is the statement of priorities and the expenditures reflect policy actions.

3.1 Methodology

The proper choice of the number of topics has been the subject of intense research in the statistical literature. Several methods for selecting K have been suggested (Airoldi et al., 2010; Taddy, 2012; Grbovic et al., 2014; Cheng et al., 2015). The most popular method selects K based on the value for which the LDA model yields the highest likelihood (Griffiths and Steyvers, 2004). Often, this method leads to large values of K (with possible duplicate topics) and topics that are difficult to interpret. In this paper we adopt the principle that the optimal number of topics depends on the particular problem at hand. In our case, we wish to investigate the relationship between topics and a number of state expenditure variables. For this reason, we choose the number of topics based on the canonical correlations between the set of topics and the set of expenditure variables.

CCA aims at identifying the structural relationship between a set of X variables and a set of Y variables. It does so by projecting the data into a lower dimensional space in which the correlation between the variables for each set are eliminated, while the correlation between the newly constructed $\tilde{X}$ and $\tilde{Y}$ canonical variates are maximized (Hardoon et al. (2004); Alissa and K. (2005); Johnson and Wichern (2007, Chapter 10)). It is an effective method for summarizing the linear association between two sets of variables. CCA involves constructing indexes of sets of variables. The indexes are constructed in such a way that the correlation between the first index derived from the Y variables ($\tilde{Y}_1$) and the first index derived from the X variable ($\tilde{X}_1$) is maximized, and the indexes from the same set of variables are orthogonal. Each index is interpreted by identifying the raw variables that contribute the most to its construction. Specifically, the correlation between the index (canonical variate) and the raw variables shows the variables that contribute the most to the construction of the index (see Appendix A.2 for further exposition of the CCA method). A main benefit of CCA over regular regression (MANCOVA for example) is its ability to reveal the fundamental relationship between two set of variables (X and Y), especially in a situation where there is a large number of variables. In fact, the CCA method has been used in economics as an effective dimension reduction method (Jacobs and Otter, 2008; Breitung and Pigorsch, 2013).

We iteratively fit a canonical correlation model, by increasing K until the marginal benefit for

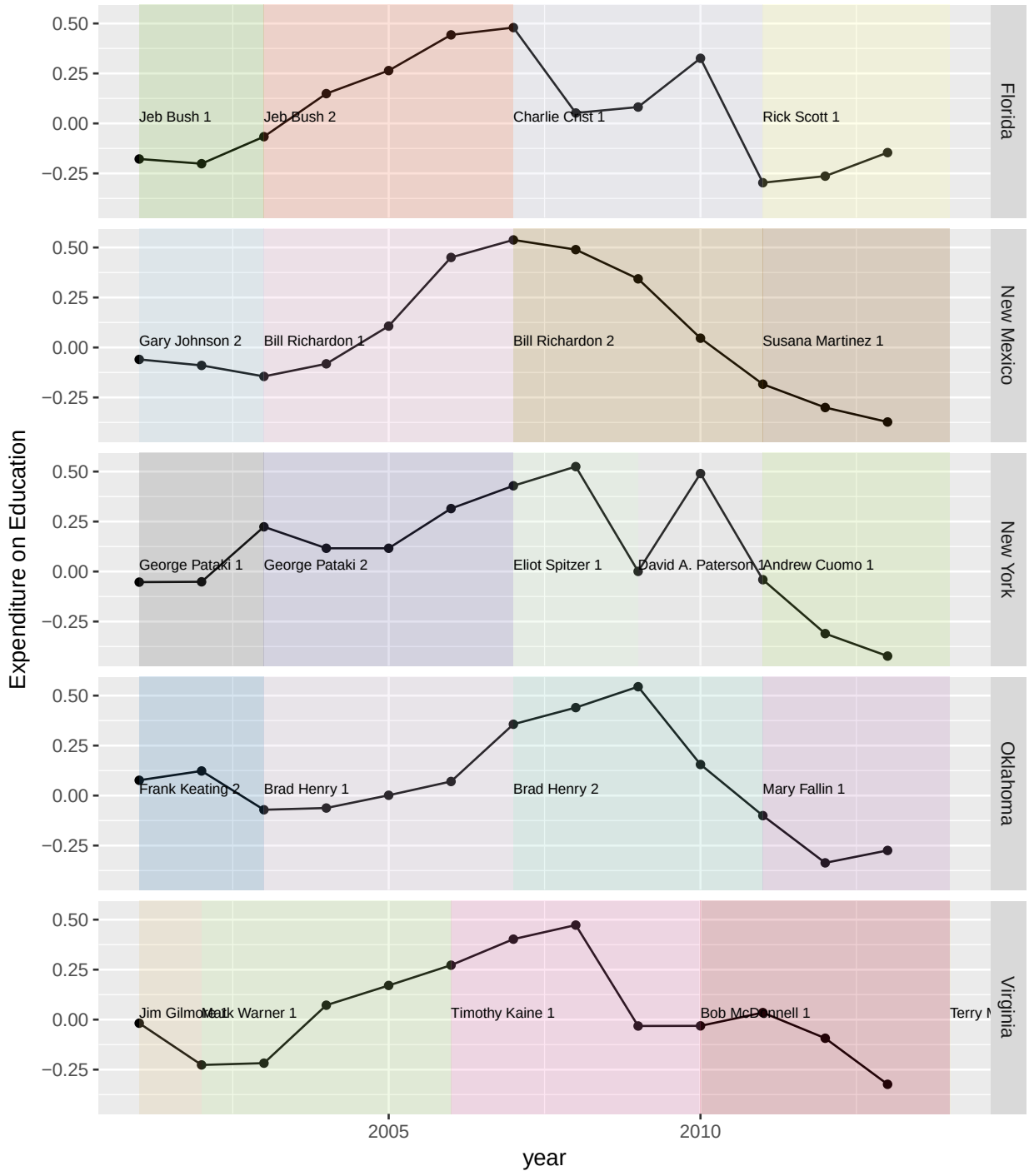


Figure 7: Time series plot of five states' spending on education (detrended z-score values)

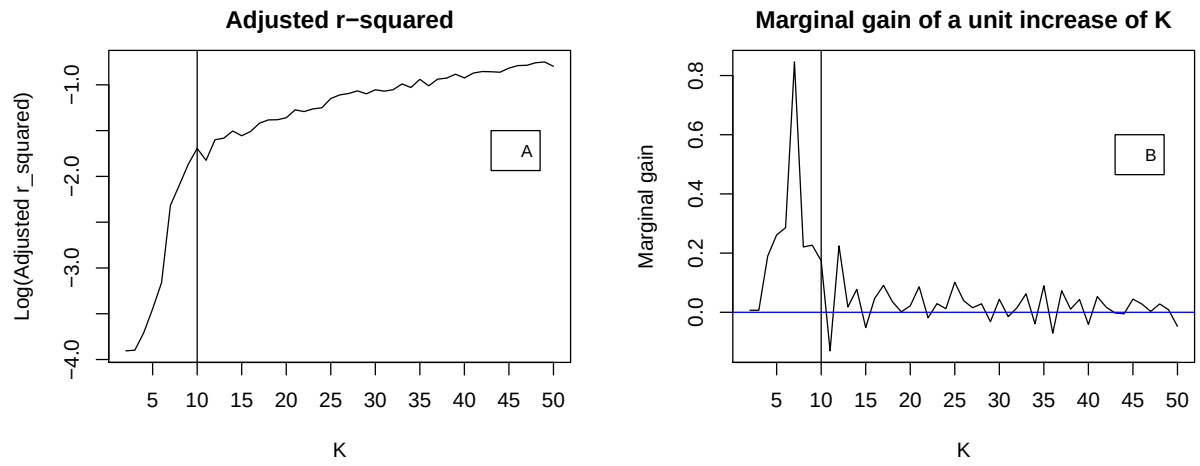


Figure 8: Panel A: Changes in the log values of the adjusted r-squared as K, the number of topics, increases. Panel B: Changes of the marginal gain of a unit increase in K.

adding an additional unit of K is close to zero. Figure 8, panel A shows the log of the adjusted r-squared as K goes from 2 to 50. There is a diminishing return to adding K, and the marginal gain fluctuates closely around zero when K > 10 (Figure 8 , panel B). Thus, K = 10 appears to be a reasonable number of topics to consider and we conduct the remainder of the analysis in this paper using this choice for K.

To summarize, our methodological approach can be termed LDA-CCA, and goes through three steps:

1. After pre-processing the 596 SoSAs (which yields 1034 unique words), we use the matrix of words counts to estimate the topics distributions in each document via LDA. The topics distributions are estimated for a varying number of topics K (Tables 1 and 3 are examples of topics distributions within documents for $K = 2$ and 3 respectively).

2. Then, we iteratively fit a CCA of the estimated topics and the expenditure variables, changing the number of topics K from 2 to 50 to decide on a reasonable K. $K = 10$ seems reasonable for the analysis (Figure 8).

3. Last, we perform a CCA of a set of 10 topics data matrix and a set of 4 expenditure variables. The goal is to uncover the fundamental linear association between the two datasets.

The approach taken in this paper (LDA-CCA) were independently proposed in Rasiwasia et al. (2010), in which the goal was to improve documents or picture retrieval algorithms using CCA methods to link text data and picture data.

3.2 Results

Associating priorities with the topics identified in the speeches is critical for the analysis of this paper. Table D.1 in appendix D provides a list of words for each topic, and these lists can be useful for associating priorities with topics. That table suggests that Topic 1 concerns energy production, and the economy; Topic 2 is about reforms, mostly about tax reform. Topic 4 seems to be about general provision of public services. Topic 7 concerns balancing budget, Topic 9 is about health care; and Topic 10 concerns education, technology and innovation.

The conclusion that a governor's speech is heavily weighted with a particular topic, does not necessarily imply that the topic is a budgetary priority. A speech devoted to denigrating the value of higher education and proposing detailed cuts would be heavily weighted to Topic 10, the higher education topic. It is revealing, therefore, to examine the words in the context of speeches with high concentration of the topic. Figures E.1 and E.2 in appendix E show excerpts of the SoSAs in Maryland 2001 and Nebraska 2011. These two excerpts contain a high share of the higher education theme (about 30% and 22% respectively from a 10 topics set). The excerpts are constructed by highlighting a few of the topic key words in the speeches, then identifying the section of the speeches where we observe most of the highlight. The two excerpts reveal a key feature of topic modeling. Beyond merely identifying the degree to which a topic is covered in a speech, we can also examine how that topic is covered in a particular speech.

From Figure E.1, it can be seen that the governor of Maryland of 2001, not only has a vision for his education policy, prior to becoming a governor, he taught for 27 years at the University of Maryland, College Park. In his speech, he mentions that his administration increased education spending by 70% in a 6 years period. Details about the foundations of his vision and how to achieve that vision are stated in the speech. On Figure E.2, the governor of Nebraska of 2011 clearly states that education, particularly higher education, research and innovation is a high priority for Nebraska. Consequently, despite budget shortfall, the education budget was increased, while budget cuts are implemented elsewhere. Further, education initiatives are delineated in the speech.

Figure 9 shows the relationship between the importance of Topic 10 and state spending on education in a selected number of states. It appears that, in general, the education priority expressed in the speeches matches the education priority expressed in the states' budgets.

3.2.1 Main results

Table 5 presents the statistical significance test statistics of no correlation between the state expenditures and the professed priorities of the governors.

The first row of Table 5 shows the statistical significance of the full model. The test statistics (of

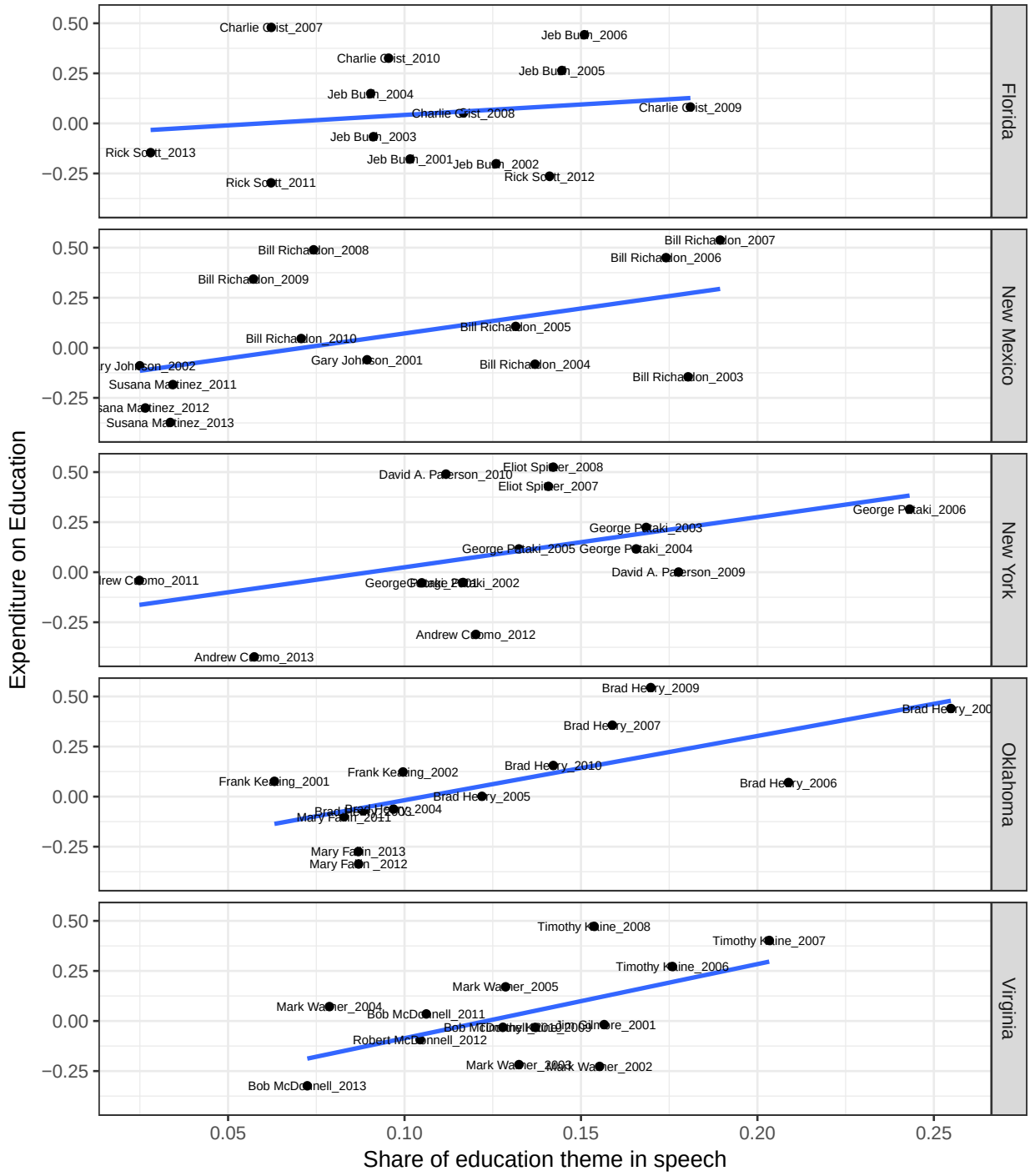


Figure 9: Relationship between the state expenditure on education (detrended values of the z-scores) and the educational content of the governors' SoSA.

Table 5: Statistical significance test statistics of no correlation between the two sets of variables.

	Wilks test stat (λ)	p-value
1	0.80	0
2	0.92	0.004
3	0.96	0.12
4	0.98	0.15

no correlation between the two sets of variables) is $\lambda = 0.8$ with a $p-value \simeq 0$, which suggests that the thematic contents of the speeches (X variables) significantly correlate with the spending variables (Y variables).

Table 6: Canonical correlations between the expenditure variables and the themes in the speeches

Pair	CanCorr (R)	RSquared
1	0.35	0.13
2	0.22	0.05
3	0.14	0.02
4	0.13	0.02

Next, from Table 6, the first canonical correlation is about 0.35, which implies that about 13% of the variation of the first canonical variate of the economic variables is explained by the first canonical variate of the topics. In sum, there does seem to be a relationship between the topics covered in the governors' SoSA and the structure of their states' expenditures. That relationship can be understood from the first pair of canonical variates. A canonical variate is a linear index variable, and is interpreted by identifying the raw (or initial) variables that contribute most to its construction. Identifying the contributing variables informs on the relationship between the X and Y variables.

From Table 7, the R column, it can be noted that the education expenditure variable, followed by the health care expenditure variable contribute mostly to the first canonical variate of the expenditure variables. The shared variance between the first canonical variate of the expenditure variables and the education spending variable is about 96%; and that with the health care spending variable is 15%. Likewise, The R column of the table suggests that the topics 4, 7, 9 and 10 contribute the most to the variation observed on the first canonical variate derived from the topics. Their shared variances are about 15%, 15%, 46% and 8%, respectively.

Examining the third column of Table 7 reveals that Topic 9 concerns to health care and Topic 10 concerns a combination of education, innovation, and Technology. An extended words table

Table 7: Correlation (R) between the first canonical variates and the raw variables

	R	RSquared	Topics words
welfare_expend	-0.09	0.01	
highways_expend	-0.26	0.07	
health_expend	-0.39	0.15	
educ_expend	-0.98	0.96	
Can Corr (Pair 1)	0.35	0.13	
Topic.4	0.39	0.15	work, depart, legisl, develop, issu, effort, servic
Topic.7	0.39	0.15	budget, govern, econom, make, educ, work, busi
Topic.2	0.25	0.06	reform, govern, system, spend, chang, school, incom
Topic.5	0.08	0.01	work, depart, legisl, develop, issu, effort, servic
Topic.6	0.08	0.01	peopl, know, just, come, make, like, work
Topic.3	-0.03	0	nation, famili, work, live, help, serv, world
Topic.8	-0.09	0.01	fund, million, budget, increas, propos, program, provid
Topic.1	-0.13	0.02	busi, energi, creat, compani, work, peopl, econom
Topic.10	-0.29	0.08	econom, futur, educ, opportun, communiti, innov, resourc
Topic.9	-0.68	0.46	health, care, famili, make, work, help, cost

can be found in Appendix D. Topic 10 can be clearly identified by looking at Figures E.1 and E.2, which present excerpts of speeches that contain a high proportion of Topic 10. Note that Topic 4 and 7 are strong variables for the construction of the first topics canonical variate. Topic 4 seems to be about general provision of public services, and Topic 7 is clearly about balancing budget as illustrated by the presence of words such as budget, fiscal, challenge, balance, tough, difficult, decision etc. In sum, the first topics canonical variate can be interpreted as a contrast between providing education and health care policies versus balancing budget policies.

The above identification of the first pair of canonical variates leads us to conclude that higher contents of education at higher level, innovation and health care in governors' speeches is associated with greater subsequent spending on education and health care. The positive relationship is captured by the similarity of the signs of the correlations between these variables and their canonical variates (see Tables C.3 and C.4 in appendix C). However, a higher content in budget related vocabulary in the governor speech is associated with a lower spending in education and health care as suggested by the opposite sign of the correlation between the budget topic (Topic 7) and the first topics canonical variate.

These findings confirm that the SoSAs are not mere words. The underlying priorities laid out in the speeches can be measured objectively and the priorities are highly correlated with measured policy actions. There is a close match between the priorities expressed in the SoSAs and the priorities revealed in the structure of the states' budgets.

It should be noted that these results are plausibly biased downward for two main reasons. First, the analysis assumes the governor favorable attitude when expressing his or her priorities in the speech (for example, talking more about health care suggests that the governor intends to spend more on health care). That is not always the case. The governor may be talking about reforms for more efficiencies. Second, the analysis assumes that the governor has dominant power in budget planning. Though that is true in most states, it is not true in all states. Consequently, the correlation between the governor expressed priorities and the structure of the state budget is probably biased downward. These observations suggest that the application of the current method to less democratic societies is likely to show stronger results. In these societies, leaders have more flexibility to talk more about things they care about, instead of managing constituents' sensibilities when cutting budgets in some sectors of the economy.

3.2.2 Robustness of the results

The choice of the number of topics, K, though not entirely arbitrary, was not based on a rigorous statistical criterion. According to the adjusted r-squared criterion in Figure 8, the choice of $K = 12$ or $K = 50$ cannot be ruled out. It is therefore necessary to examine the sensitivity of the results to the choice of K. From Appendix C, Tables C.1 , C.5 and C.9 , the entries in the first two rows of each table show statistically significant results. We will focus our attention on the first pairs of canonical variates of each of the Ks for simplicity and relevance. Table 8 summarizes the test statistics of no correlations between the two sets of variables (expenditures and themes). The $p-values$ indicate highly significant results; therefore, we reject the null hypothesis and conclude that there is a correlation between at least one pair of expenditure and themes canonical variates.

Table 9 summarizes the canonical correlations for each assumed K. For example, the first row shows the four canonical correlations under the assumption of $K = 10$ topics model. Likewise, the second and third rows shows the four canonical correlations under the assumption of $K = 12$, and $K = 50$, respectively.

Table 8: Statistical significance test statistics of no correlation between any two sets of canonical variates for different values of K, the number of topics.

	Wilks test stat (λ)	p-value
K = 10	0.80	0
K = 12	0.78	0
K = 50	0.50	0

The canonical correlations increase with the number of topics. Referring to the first column of Table 9 (the column of the canonical correlations of the first pairs of canonical variates) we

Table 9: Canonical correlations between the expenditure variables and the themes in the speeches, for different values of K, the number of topics.

	CanCorr_1	CanCorr_2	CanCorr_3	CanCorr_4
K = 10	0.35	0.22	0.14	0.13
K = 12	0.37	0.23	0.15	0.14
K = 50	0.53	0.38	0.36	0.26

note that the increase going from $K = 10$ to 12 is minimal (0.02) and the increase from $K = 10$ to $K = 50$ (i.e. an approximately 400% increase in the number of topics K) is modest (about 0.18, i.e. 51% increase).

Though the canonical correlation increases with K, the qualitative result does not change with K. In fact, the main finding of the robustness check is that the conclusion remains the same whether the analysis is performed using a 10, 12, or 50 topics model. That is, the focus on health care and higher education is positively associated with states' spending in education and health care, whereas a focus on balancing the budget and the provision of other public services is negatively associated with states' spending in education and health care.

To see the similarity between the three results, recall that for $K = 10$, the highest contributing variables for the first expenditures canonical variate are education and health care expenditures (Table 7). This observation remains true when $K = 12$ or 50 (see the correlations between the first canonical variate and the raw expenditure variables in Tables C.7 and C.11 in appendix C). Table 10 compares the main contributing topics to the first topics' canonical variate, for each assumed K (full results can be found in Tables C.8 and C.12 in Appendix C). First, the first three columns of Table 10 list the words distributions of what we identified as higher education, innovation and technology topic under $K = 10$ (noted Top.10_10), $K = 12$ (noted Top.12_12), and $K = 50$ (noted Top.49_50), respectively. For $K = 50$, Topic.49 (Top.49_50) does not show education in the words list. However, a closer look at the words list reveals the prominence of innovation, science, research and technology for this topic (see Figure 10). Additionally, we note that for $K = 50$, Topic 9 is also important for the first topics canonical variate. Topic 9 is clearly about college level education (see Figure 10). Topic 9 and Topic 49 for $K = 50$ seem to be a split of Topic 10 for $K = 10$. Moreover, Topic.12 for $K = 12$ shows clearly that the topic refers to college level education as words such as "college", "university", "innovation", and "higher" are prominent in the words list. Second, the next three columns compare the health care topic for $K = 10, 12$, and 50; the health care theme is fairly easy to identify, since it shows "health" and "care" as the top words in the words lists. Third, the last two columns are all referring to the balancing of budget topic for $K = 10$, and for $K = 12$. The balanced budget topic is negatively associated with spending on education and health care in

all three assumed K.

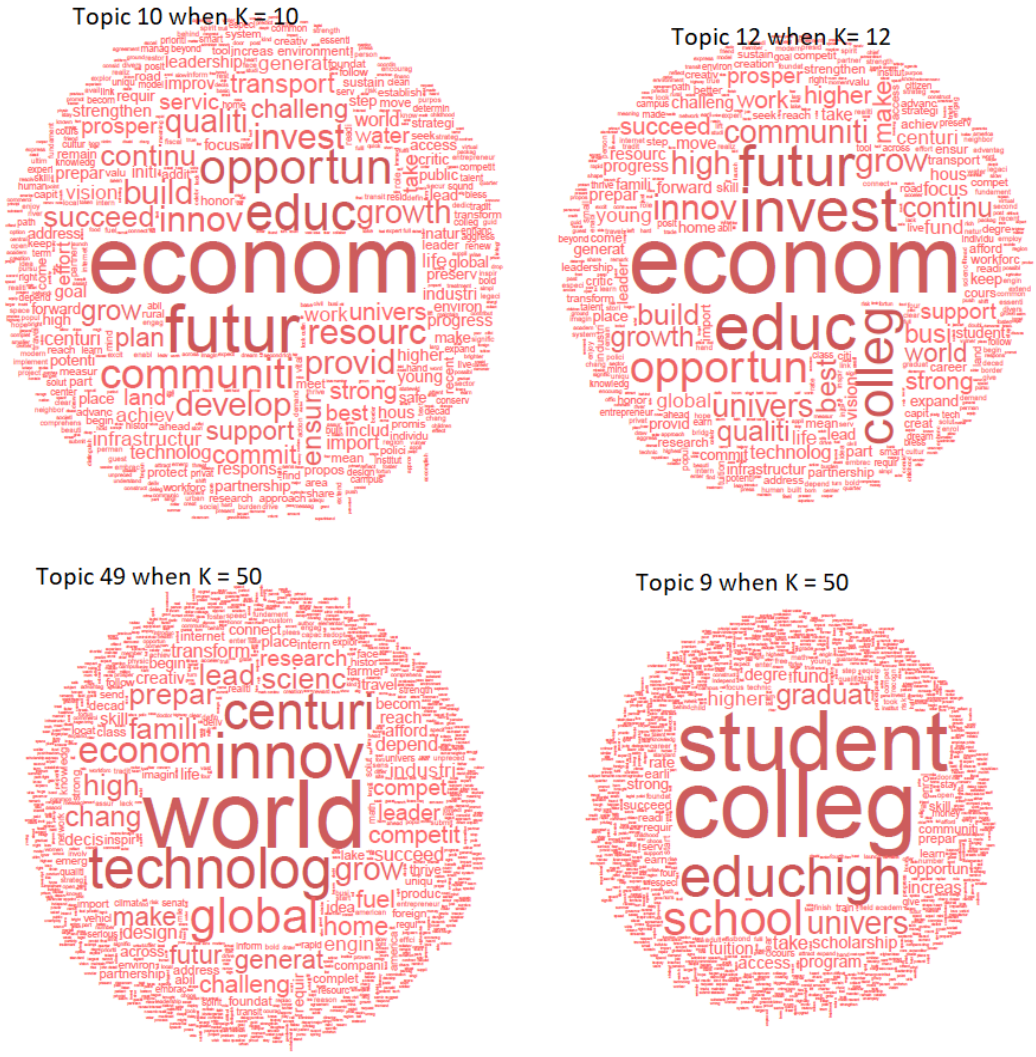


Figure 10: Comparison of Topic.10 when K = 10, Topic.12 when K = 12, and Topic.49 when K = 50

For completeness, Figures 11 and 12 show the correlations between the relevant topics for $K = 10$ and the relevant topics for $K = 12$; and the correlations between the relevant topics for $K = 10$ and the relevant topics for $K = 50$, respectively. By relevant topics, we refer to topics used for the interpretation of the results; one of the themes with the highest positive correlation with the canonical variate (i.e. the balanced budget topic), and the two themes with the highest negative correlation with the canonical variate (i.e. the education and health care topics). We present the complete figures of the correlations in Figures F.1 and F.2 in appendix F. The darkest cells of Figure 11 refer to the equivalent topics. By equivalence, we mean for example that the Topic 9 proportions in the speeches when $K = 10$ is very similar to the Topic 11 proportions of the $K = 12$ topics model (their correlation is 0.99). Put differently, the Topic 9 variable refers to the relative

Table 10: Comparison of the topics' words for different topic model (K = 10, 12 and 50)

	Top.10_10	Top.12_12	Top.49_50	Top.9_10	Top.11_12	Top.43_50	Top.7_10	Top.8_12
1	econom	econom	world	health	health	health	budget	budget
2	futur	educ	innov	care	care	care	govern	econom
3	educ	invest	technolog	famili	famili	insur	econom	public
4	opportun	futur	centuri	make	cost	cost	make	face
5	communiti	colleg	global	work	help	afford	educ	govern
6	innov	opportun	econom	help	make	famili	work	challeng
7	resourc	innov	grow	cost	insur	coverag	busi	futur
8	invest	grow	scienc	children	work	medic	servic	make
9	provid	univers	prepar	insur	children	access	face	fiscal
10	develop	high	lead	afford	afford	reduc	challeng	difficult
11	build	communiti	high	protect	provid	expand	balanc	crisi
12	growth	build	famili	provid	citizen	provid	dollar	balanc
13	qualiti	continu	chang	communiti	access	help	public	respons
14	continu	best	make	invest	school	home	fiscal	come
15	succeed	growth	futur	access	worker	hospit	decis	problem
16	ensur	qualiti	generat	peopl	invest	doctor	citizen	revenu
17	grow	world	home	citizen	protect	plan	difficult	california
18	support	busi	challeng	medic	peopl	premium	revenu	recess
19	challeng	make	competit	drug	plan	cover	tough	decis
20	plan	higher	fuel	expand	creat	uninsur	effici	tough
21	transport	work	leader	support	busi	protect	creat	children
22	commit	succeed	compet	home	medic	pass	save	protect
23	univers	strong	research	program	support	prevent	spend	choic
24	strong	support	industri	school	expand	increas	reduc	governor
25	water	prosper	engin	plan	nation	direct	fund	plan

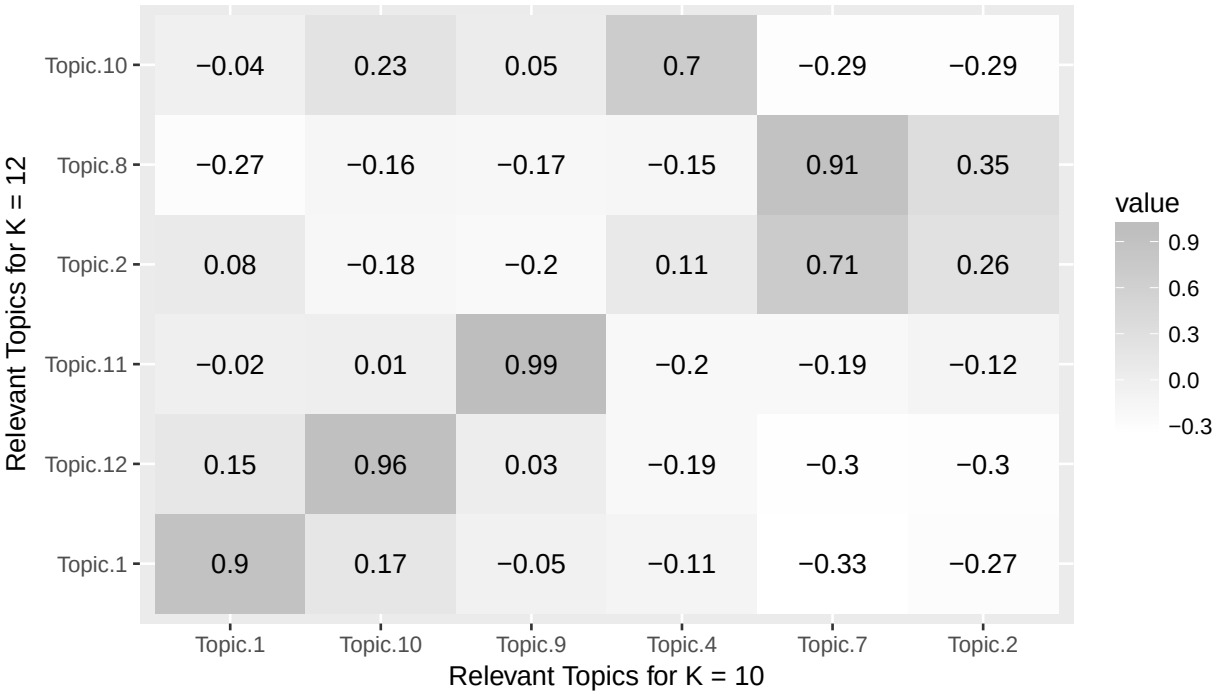


Figure 11: Correlations between the relevant topics when K = 10 and when K = 12

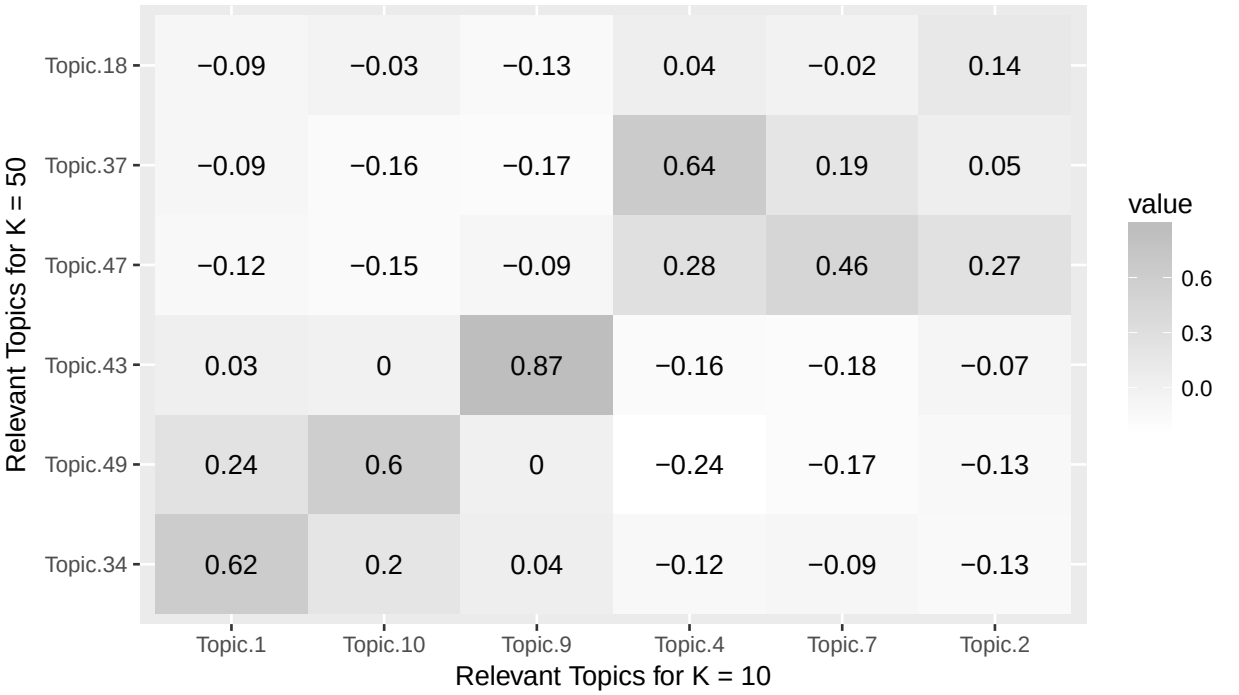


Figure 12: Correlations between the relevant topics when K = 10 and when K = 50

importance of the health care theme in each speech when we consider a $K = 10$ topics model; The Topic 11 variable refers to the relative importance of the health care theme in each speech when we consider a $K = 12$ topics model. Both of these topics have the same interpretation, based on their top words; and the correlation between them is 0.99, suggesting that they are almost identical variables.

To summarize, the robustness check shows that the qualitative interpretation of our results does not change with the number of topics, K, chosen; reinforcing the idea that the thematic contents of the governors' SoSAs contain substantial information with regard to the governors' priorities as materialized in the structure of the states' budgets. That further confirms the hypothesis that the thematic contents of the governors' SoSA can be used as a proxy for their policy agendas and actions.

The major tradeoff in selecting the number of topics involves the increase in explanatory power against the increase in complexity that reduces our ability to interpret the topics. The interpretation of the topics is not always straightforward; and the higher the number of topics to interpret, the harder the task of interpreting them. In the present case, we believe the gain is entirely offset by the complications of dealing with the larger number of topics. In any case, the basic results are robust to the choice of the number of topics, $K = 10$, 12, or 50.

4 Discussion

The present paper provided a rigorous test of the relationship between the thematic contents of the U.S governors' SoSA and budgetary outcomes. We studied the correlation between the themes covered in their speeches and the structure of the state's spending on education, health care, highways, and public welfare. To avoid arbitrarily choosing the number of themes (or topics) in the speeches, we iteratively ran several canonical correlations, changing the value of K. Figure 8, panel B shows that past 10 topics, the marginal gain for an additional topic is negligible. Consequently, the analysis proceeded with the assumption that every speech contains 10 topics with varying importance.

4.1 Main findings

The thematic contents of the U.S governors' SoSAs and subsequent changes in state expenditures are correlated. When governors devote more content to education, innovation, technology, and health care, state expenditure rises in education and health care. When they focus on balancing budgets and general provision of public services, state expenditures on education and health care fall. Thus, the LDA method, combined with the CCA method were successful in uncovering this expected relationship. We conclude that the thematic contents of the governors' speeches can be measured, and these measures are valid proxies for governors' priorities.

4.2 Implications

Topic modeling was initially developed as a document retrieval algorithm. For instance, a law firm may have a case that requires it to browse thousands of documents in search of possible useful documents for the case. A topic modeling algorithm can be used to quantify the thematic contents of each document; the firm can then quickly identify the most important documents based on their thematic contents. The usefulness of the algorithm for social sciences research is just beginning.

The findings in this paper provide an avenue for studying political leadership and economic development. As emphasized by Brady and Spence (2010), political leaders do many things. Some may be successful in improving the education, health care, security, and institutions of their states, or countries. By being able to measure their professed priorities as stated in their public statements, it is possible to study, quantitatively, the link between their priorities and their achievements. We can analyze more than economic growth, which, though desirable, may follow after other important factors such as infrastructures, human capital, and institutions have been created. CCA eases the study of several outcome variable simultaneously.

The main contribution of this paper is to combine LDA with CCA to show that the LDA technique can successfully reveal leaders' priorities from their political statements. The case of the U.S. governors' SoSA provides a particularly clean test case. Differences in language, traditions and customs are minimized. Further, subsequent changes in state expenditures provide a direct measure of actual priorities. If the technique is successful in revealing leader's priorities, we can then apply the technique to more difficult questions. For example, do political leaders who persistently talk about economic development achieve higher economic growth? Do political leaders who persistently talk about education improve the education attainment of their school systems? Do political leaders whose professed priority is business promotion succeed in attracting new businesses? Topic modeling does not only provide an approach to quantitatively study these associations, it also provides a path to understanding the mechanisms of these associations, as Figure E.1 and E.2 illustrate. These two figures show the how and the why two leaders are promoting their higher education agendas. Consequently, we may be able to address the opacity of the "exact mechanism at work in explaining how leadership matter" (Besley et al., 2011, p.F219).

The importance of politics for economic development is increasingly being recognized in the mainstream economic literature. However, the analysis is mainly theoretical. It has "sought to develop positive models of how policy actually gets chosen" (Acemoglu and Robinson, 2013). Evidence remains largely anecdotal or consist of individual case studies; often with a focus on the role of the state for economic development (Chang, 2002; Reinert, 2007; Evans, 2012; Mazzucato, 2015). The leadership approach focuses on the instrumental role of the lead actor of the state (Brady and Spence, 2010). Though the economics literature is still embryonic (Jones and Olken, 2005; Besley et al., 2011; Blinder and Watson, 2016; Easterly and Pennings, 2016), a well formed theory of political leader as the most significant actor of state policies can be found in political science (Hermann et al., 2001; Hermann, 2008).

Using dimension reduction methods akin to principal component analysis is well known in the economics literature (Ram, 1982; Temple and Johnson, 1998; Bernanke and Boivin, 2003; Bérenger and Verdier-Chouchane, 2007; Tabellini, 2010; Decancq and Lugo, 2013). Topic modeling is a hierarchical Bayesian approach of PCA applied to count data. This paper brought together machine learning techniques (LDA), traditional statistical methods (CCA), political science ideas and economics case studies to provide a pathway to systematically quantify political leaders' priorities and explore the association between these priorities and economic outcomes.

4.3 Limitations

Though the conclusions are reasonably robust to the choice of the number of topics, the conclusions must be tempered somewhat by the recognition that the choice of the number of topics is not

based on an agreed-upon statistical criterion. Like regular factor models, choosing the appropriate number K of topics (or components) remains an art rather than algorithmic. Second, associating a topic with a priority is complicated by the fact that the LDA algorithm may fail to reveal the level of positive or negative *"sentiment"* associated with a given topic. In principle, a speech that is heavily focused on education could reflect a governor's priority to reduce spending on education. Currently, there is no satisfactory substitute for direct examination of speeches to uncover this problem. Finally, the LDA algorithm is based on Markov Chain Monte Carlo or Variational Bayesian estimation methods. The estimated parameters are not unique (local maxima concerns for the case of Variational Bayes).

4.4 Challenges

There are several challenges to applying topic modeling algorithms to study the association between political leaders' priorities and economic outcomes. First, the diversity of languages used by leaders to communicate make it challenging for cross-country analysis. However, computer translation can still be useful in converting speeches to a single language. Moreover, the choice of the types of speeches matters for the type of questions the researcher seeks to answer (Hermann, 2008). Second, the unstructured nature of text data makes it computationally demanding (Einav and Levin, 2014; Varian, 2014); for example, retrieving and pre-processing the data is not always trivial. Guides exist to ease these processes (Munzert, 2014; Silge and Robinson, 2017). Finally, the interpretation of CCA generated factors are not always obvious.

5 Conclusion

The role of political leaders for economic growth remains a challenging question to study, due in part to the lack of agreed-upon measures of leadership. This paper argued in favor of using the leaders' public statements to identify their professed priorities and use these professed priorities as proxies for what they proclaim to be doing. Using the U.S governors' SoSA as a test case, we identified the thematic contents of these speeches and showed that the higher the U.S governors talk about education at higher level and health care, the more their states spend on education and health care. The more they talk about balancing budgets, or about general provision of public services, the lower they spend on education and health care, in general.

The U.S governors' SoSAs is a felicitous data for matching governors' professed priorities and actualized priorities. Indeed, the U.S. governors use the SoSAs to layout their priorities, and to mobilize constituents to support their proposed budgets. Consequently, the themes covered in the SoSAs reflect the priorities spelled out in the governors' proposed budgets. Our findings suggest

that the LDA algorithm can successfully quantify the relative importance of each theme in each speech; and that these thematic measures strongly correlate with the states' budgets structure, as shown by the use of CCA. The paper makes the case that topic modeling provides a pathway to quantifying political leaders professed agendas, and their commitments to their agendas, as measured by the consistency with which some themes are present in successive speeches. We illustrated the importance of this approach to studying leadership and economic growth by showing that U.S governors' commitment to their economic agenda is strongly associated with business expansion. Moreover, our analysis reveals that beyond quantifying leaders' priorities as stated in their speeches, we can get a glimpse at the motivations of these priorities and actions taken to promote them. The paper suggests that further analysis of these and other political speeches may yield insights into the influence of political leadership on economic development.

References

Acemoglu, D. and Robinson, J. A. (2013). Economics versus politics: Pitfalls of policy advice. *The Journal of Economic Perspectives*, 27(2):173–192.

Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T., and Shringarpure, S. (2010). Reconceptualizing the classification of pnas articles. *Proceedings of the National Academy of Sciences*, 107(49):20899–20904.

Alexopoulos, M. and Cohen, J. (2015). The power of print: Uncertainty shocks, markets, and the economy. *International Review of Economics & Finance*, (0):–.

Alissa, S. and K., H. R. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment*, 84(1):34–48.

Bai, J. and Ng, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(6):1577–1606.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8:53–80.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Bérenger, V. and Verdier-Chouchane, A. (2007). Multidimensional measures of well-being: standard of living and quality of life across countries. *World Development*, 35(7):1259–1276.

Bernanke, B. S. and Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, 50(3):525–546.

Besley, T., Montalvo, J. G., and Reynal-Querol, M. (2011). Do educated leaders matter? *The Economic Journal*, 121(554).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Blinder, A. S. and Watson, M. W. (2016). Presidents and the us economy: An econometric exploration. *American Economic Review*, 106(4):1015–45.

Boukus, E. and Rosenberg, J. V. (2006). The information content of fomc minutes. *Federal Reserve Bank of New York*.

Brady, D. and Spence, M. (2010). *Leadership and Growth : Commission on Growth and Development.* World Bank., https://openknowledge.worldbank.org/handle/10986/2404.

Breitung, J. and Pigorsch, U. (2013). A canonical correlation approach for selecting the number of dynamic factors. *Oxford Bulletin of Economics and Statistics*, 75(1):23–36.

Chang, H.-J. (2002). *Kicking away the ladder: development strategy in historical perspective*. Anthem Press.

Cheng, D., He, X., and Liu, Y. (2015). Model selection for topic models via spectral decomposition. In *Artificial Intelligence and Statistics*, pages 183–191.

Decancq, K. and Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews*, 32(1):7–34.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Easterly, W. and Pennings, S. (2016). Shrinking dictators: how much economic growth can we attribute to national leaders. *Development Research Institute Working Paper*, (94).

Einav, L. and Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210).

Evans, P. B. (2012). *Embedded autonomy: States and industrial transformation*. Princeton University Press.

Ferguson, M. (2006). *The Executive Branch of State Government: People, Process, and Politics*. ABC-CLIO's about state government. ABC-CLIO.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as data. Technical report, National Bureau of Economic Research.

Grbovic, M., Halawi, G., Karnin, Z., and Maarek, Y. (2014). How many folders do you really need?: Classifying email into a handful of categories. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 869–878, New York, NY, USA. ACM.

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

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Grun, B. and Hornik, K. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software, Articles*, 40(13):1–30.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133.

Hardoon, D. R., Szedmak, S. R., and Shawe-taylor, J. R. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664.

Heidbreder, B. (2012). Agenda setting in the states: How politics and policy needs shape gubernatorial agendas. *Politics & Policy*, 40(2):296–319.

Hermann, M. G. (2008). *Using Content Analysis to Study Public Figures.* Palgrave Macmillan.

Hermann, M. G., Preston, T., Korany, B., and Shaw, T. M. (2001). Who leads matters: The effects of powerful individuals. *International Studies Review*, 3(2):83–131.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196.

Jacobs, J. P. and Otter, P. W. (2008). Determining the number of factors and lag order in dynamic factor models: A minimum entropy approach. *Econometric Reviews*, 27(4-6):385–397.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis (6th Edition)*. Pearson, 6 edition.

Jones, B. F. and Olken, B. A. (2005). Do leaders matter? national leadership and growth since world war II. *The Quarterly Journal of Economics*, 120(3):835–864.

Kapetanios, G. and Marcellino, M. (2010). Factor-gmm estimation with large sets of possibly weak instruments. *Computational Statistics & Data Analysis*, 54(11):2655–2675.

Landauer, T. K., Mcnamara, D. S., Dennis, S., and Kintsch, W., editors (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

Mazzucato, M. (2015). *The entrepreneurial state: Debunking public vs. private sector myths*, volume 1. Anthem Press.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

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

Munzert, S. (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Okonjo-Iweala, N. (2012). *Reforming the Unreformable: Lessons from Nigeria*. MIT Press.

Ram, R. (1982). Composite indices of physical quality of life, basic needs fulfilment, and income: A 'principal component' representation. *Journal of Development Economics*, 11(2):227–247.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM.

Reinert, E. S. (2007). *How rich countries got rich and why poor countries stay poor*. Constable.

Romer, C. D. and Romer, D. H. (2015). New Evidence on the Impact of Financial Crises in Advanced Countries. NBER Working Papers 21021, National Bureau of Economic Research, Inc.

Schneider, B. R., Estarellas, P. C., and Bruns, B. (2018). The politics of transforming education in ecuador: Confrontation and continuity, 2006-17. "Online; accessed 15-June-2018: `https://www.riseprogramme.org/sites/www.riseprogramme.org/files/publications/RISE_WP-021_Schneider%2C%20Cevallos%20Estarellas%2C%20Bruns.pdf`".

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2017). Measuring news sentiment. Federal Reserve Bank of San Francisco.

Shlens, J. (2014). A tutorial on principal component analysis. *CoRR*, abs/1404.1100.

Silge, J. and Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Incorporated.

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

Stock, J. H. and Watson, M. (2011). Dynamic factor models. *Oxford Handbook on Economic Forecasting*.

Tabellini, G. (2010). Culture and institutions: economic development in the regions of europe. *Journal of the European Economic Association*, 8(4):677–716.

Taddy, M. (2012). On estimation and selection for topic models. In Lawrence, N. D. and Girolami, M. A., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 1184–1193.

Temple, J. and Johnson, P. A. (1998). Social capability and economic growth. *The Quarterly Journal of Economics*, 113(3):965–990.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Winter, D. G. (2005). Things i've learned about personality from studying political leaders at a distance1. *Journal of Personality*, 73(3):557–584.

A LDA-CCA

A.1 Latent Dirichlet Allocation: Bayesian Variational Derivation of the Posterior Distribution

A topic ϕ_k is a distribution over V unique words, each having a proportion $\phi_{k,v}$; i.e $\phi_{k,v}$ is the relative importance of the word v for the definition (or interpretation) of the topic k. It is assumed that:

$$\phi_k \sim Dirichlet_V(\beta)$$

That is:

$$p(\phi_k|\beta) = \frac{1}{B(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta_i - 1}$$

Where $B(\beta) = \frac{\prod_{v=1}^{V} \Gamma(\beta_i)}{\Gamma(\sum_{v=1}^{V} \beta_v)}$, and $\beta = (\beta_1, ..., \beta_V)$. Since we have K independent topics (by assumption),

$$p(\phi|\beta) = \prod_{k=1}^{K} \frac{1}{B(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1}$$

A document d is a distribution over K topics, each having a proportion $\theta_{d,k}$, i.e. $\theta_{d,k}$ is the relative importance of the topic k, in the document d. We assume:

$$\theta_d \sim Dirichlet_K(\alpha)$$

That is:

$$p(\theta_d|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1}$$

And since we have D independent documents (by assumption),

$$p(\theta|\alpha) = \prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1}$$

It is further assumed that $\beta_v = \beta$, and $\alpha_k = \alpha$

Let z be the latent topic assignment variable, i.e. the random variable $z_{d,n}$ assigns the word $w_{d,n}$ to the topic k in document d. $z_{d,n}$ is a vector of zeros and 1 at the k^{th} position ($z_{d,n} = [0,0,...1,0,..]$). Define $z_{d,n,k} = I(z_{d,n} = k)$ where I is an indicator function that assigns 1 to the random variable $z_{d,n}$ when $z_{d,n}$ is the topic k, and 0 otherwise.We assume:

$$z_{d,n} \sim Multinomial(\theta_d)$$

That is:

$$\begin{aligned} p(z_{d,n,k}|\theta_d) &= \theta_{d,k} \\ &= \prod_{k=1}^{K} \theta_{d,k}^{z_{d,n,k}} \end{aligned}$$

A document is assumed to have N_d independent words, and since we assume D independent documents, we have:

$$\begin{aligned} p(z|\theta) &= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \theta_{d,k}^{z_{d,n,k}} \\ &= \prod_{d=1}^{D} \prod_{k=1}^{K} \prod_{n=1}^{N_d} \theta_{d,k}^{z_{d,n,k}} \\ &= \prod_{d=1}^{D} \prod_{k=1}^{K} \prod_{v=1}^{V} \theta_{d,k}^{n_{d,v} * z_{d,v,k}} \\ &= \prod_{d=1}^{D} \prod_{v=1}^{V} \prod_{k=1}^{K} \theta_{d,k}^{n_{d,v} * z_{d,v,k}} \end{aligned}$$

$n_{d,v}$ is the count of the word v in document d.

The word $w_{d,n}$ is drawn from the topic's words distribution ϕ_k:

$$w_{d,n}|\phi_{k=z_{d,n,k}} \sim Multinomial(\phi_{k=z_{d,n}})$$

$$\begin{aligned} p(w_{d,n} = v|\phi_{k=z_{d,n}}) &= \phi_{k,v} \\ &= \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{k,v}^{w_{d,n,v} * z_{d,n,k}} \end{aligned}$$

$w_{d,n}$ is a vector of zeros and 1 at the v^{th} position. Define $w_{d,n,v} = I(w_{d,n} = v)$ where I is an indicator function that assigns 1 to the random variable $w_{d,n}$ when $w_{d,n}$ is the word v, and 0 otherwise.

There are D independent documents, each having N_d independent words, so:

$$p(w|\phi) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{k,v}^{w_{d,n,v} * z_{d,n,k}}$$

$$p(w|\phi) = \prod_{d=1}^{D} \prod_{v=1}^{V} \prod_{k=1}^{K} \phi_{k,v}^{n_{d,v} * z_{d,v,k}}$$

The joint distribution of the observed words w and unobserved (or hidden variables) θ, z, and ϕ is given by:

$$P(w,z,\theta,\phi|\alpha,\beta) = p(\theta|\alpha)p(z|\theta)p(w|\phi,z)p(\phi|\beta)$$

The goal is to get the posterior distribution of the unobserved variables:

$$p(z,\theta,\phi|w,\alpha,\beta) = \frac{P(w,z,\theta,\phi|\alpha,\beta)}{\int\int \sum_z P(w,z,\theta,\phi|\alpha,\beta) d\theta d\phi}$$

$\int\int \sum_z P(w,z,\theta,\phi|\alpha,\beta) d\theta d\phi$ is intractable, so approximation methods are used to approximate the posterior distribution. The seminal paper of LDA (Blei et al., 2003) uses the Mean Field Variational Bayes (an optimization method) to approximate the posteriors distribution (See Bishop (2006, p. 462) or Blei et al. (2017) for an exposition of the theory of the variational method). The mean field variational inference uses the following approximation:

$$p(z,\theta,\phi|w,\alpha,\beta) \simeq q(z,\theta,\phi) = q(z)q(\theta)q(\phi)$$

From Bishop (2006, p.466), we have:

$$q^*(z) \propto exp\left\{E_{\theta,\phi}\left[log(p(z|\theta)) + log(p(w|\phi,z))\right]\right\}$$

$$q^*(\theta) \propto exp\left\{E_{z,\phi}\left[log(p(\theta|\alpha)) + log(p(z|\theta))\right]\right\}$$

$$q^*(\phi) \propto exp\left\{E_{\theta,z}\left[log(p(\phi|\beta)) + log(p(w|\phi,z))\right]\right\}$$

Using the expressions above, we have:

$$log(q^*(z)) \propto E_{\theta,\phi}\left[\sum_{d=1}^{D}\sum_{v=1}^{V}\sum_{k=1}^{K} n_{d,v} * z_{d,v,k}\left(log(\theta_{d,k}) + log(\phi_{k,v})\right)\right]$$
$$\propto \sum_{d=1}^{D}\sum_{v=1}^{V}\sum_{k=1}^{K} n_{d,v} * z_{d,v,k}\left(E(log(\theta_{d,k})) + E(log(\phi_{k,v}))\right)$$

Note that $x|p \sim Multinomial_K(p) \iff log\left(p(x|p)\right) = \sum_{k=1}^{K} x_k log(p_k)$, and let's define $log(p_k) = E(log(\theta_{d,k}) + E(log(\phi_{k,v}))$, so $p_k = exp(E(log(\theta_{d,k})) + E(log(\phi_{k,v})))$. Thus,

$$q^*(z) \propto \prod_{d=1}^{D}\prod_{v=1}^{V}\prod_{k=1}^{K}\left[exp(E(log(\theta_{d,k})) + E(log(\phi_{k,v})))\right]^{n_{d,v} * z_{d,v,k}}$$

That is,

$$z_{d,v}|w_d, \theta_d, \phi_k \sim Multinomial_K(p_k)$$

and by the multinomial properties,

$$E(z_{d,v,k}) = p_k = exp(E(log(\theta_{d,k})) + E(log(\phi_{k,v})))$$

$$q^*(\theta) \propto exp\left\{E_z\left[\sum_d\sum_k(\alpha-1)log(\theta_{d,k}) + \sum_d\sum_k\sum_v n_{d,v} * z_{d,v,k} log(\theta_{d,k})\right]\right\}$$
$$= \prod_d^D\prod_{k=1}^{K} exp\left\{(\alpha + \sum_{v=1}^{V} n_{d,v}E(z_{d,v,k}) - 1)log(\theta_{d,k})\right\}$$
$$= \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{\alpha+\sum_{v=1}^{V} n_{d,v}E(z_{d,v,k})-1}$$

Thus, the approximate posterior distribution of the topics distribution in a document d is:

$$\theta_d|w_d, \alpha = Dirichlet_K(\tilde{\alpha}_d)$$

where $\tilde{\alpha}_d = \alpha + \sum_{v=1}^{V} n_{d,v}E(z_{d,v,.})$. Note that $\tilde{\alpha}_d$ is a vector of K dimension.

By the properties of the Dirichlet distribution, the expected value of $\theta_d|\tilde{\alpha}_d$ is given by:

$$E(\theta_d|\tilde{\alpha}_d) = \frac{\alpha + \sum_{v=1}^{V} n_{d,v}E(z_{d,v,.})}{\sum_{k=1}^{K}[\alpha + \sum_{v=1}^{V} E(z_{d,v,k})]} \tag{A.1}$$

The numerical estimation of $E(\theta_d|\tilde{\alpha}_d)$ gives the estimates of the topics within each document d, (θ_d). It is worth noting that $E(z_{d,v,k})$ can be interpreted as the *responsibility* that topic k takes for explaining the observation of the word v in document d. Ignoring for a moment the denominator of equation A.1, $E(\theta_{d,k}|\tilde{\alpha}_{d,k})$ is similar to a regression equation where $n_{d,v}$ are the observed counts of words in document d, and $E(z_{d,v,k})$ are the parameter estimates (or weight) of the words. That illustrates that the importance of a topic in a document is due to the high presence of words $(n_{d,v})$ referring to that topic, and the weight of these words $(E(z_{d,v,k}))$.

Similarly,

$$
\begin{aligned}
q^*(\phi) &\propto exp\left\{E_z\left[\sum_{k=1}^{K}\sum_{v=1}^{V}(\beta-1)log(\phi_{k,v})+\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V}n_{d,v}*z_{d,v,k}log(\phi_{k,v})\right]\right\} \\
&=\prod_{k=1}^{K}\prod_{v=1}^{V}exp\left\{(\beta+\sum_{d=1}^{D}n_{d,v}*E(z_{d,v,k})-1)log(\phi_{k,v})\right\} \\
&=\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{\beta+\sum_{d=1}^{D}n_{d,v}*E(z_{d,v,k})}
\end{aligned}
$$

Thus, the approximate posterior distribution of the words distribution in a topic ϕ_k is:

$$\phi_k|w,\beta \sim Dirichlet_V(\tilde{\beta}_k)$$

where $\tilde{\beta}_k = \beta+\sum_{d=1}^{D}n_{d,v}*E(z_{d,.,k})$. Note that$\tilde{\beta}_k$ is a vector of V dimension.

And the expected value of $\phi_k|\tilde{\beta}_k$ is given by:

$$E(\phi_k|\tilde{\beta}_k)=\frac{\beta+\sum_{d=1}^{D}n_{d,v}*E(z_{d,.,k})}{\sum_{v=1}^{V}(\beta+\sum_{d=1}^{D}n_{d,v}*E(z_{d,v,k}))} \tag{A.2}$$

The numerical estimation of $E(\phi_k|\tilde{\beta}_k)$ gives the estimates of the words relative importance for each topic k, (ϕ_k). Ignoring the denominator in equation A.2, $E(\phi_{k,v}|\tilde{\beta_{k,v}})$ is the weighted sum of the the frequencies of the word v in each of the documents $(n_{d,v})$, the weights being the responsibility topic k takes for explaining the observation of the word v in document d $(E(z_{d,v,k}))$.

Here, we have derived the posteriors expected values of the θs and ϕs using the words counts $n_{d,v}$, which is different from Blei et al. (2003). Posterior formulae similar to the current derived solution can be found in Murphy (2012, p. 962).

In sum, the rows of $\phi_{K,V} = \left[E(\phi_k|\tilde{\beta}_k)\right]_{K,V}$ are useful for interpreting (or identifying) the themes, which relative importance in each document are represented by the columns of $\theta_{D,K} = [E(\theta_d|\tilde{\alpha}_d)]_{D,K}$.

A.2 Canonical Correlation Analysis

Next, $Y_{q,1}$, the column vector of expenditure variables, and $\theta_{K,1}$, the column vector of the thematic content vaiable of the speeches, are used to perform Canonical Correlation Analysis (CCA). Here, we give a short explanation of CCA. To do so, assume $Cov(\theta) = \Sigma_\theta$; $Cov(Y) = \Sigma_Y$ are both positive definite and $Cov(\theta, Y) = \Sigma_{\theta,Y} = \Sigma_{Y,\theta}^T$, where T in $\Sigma_{Y,\theta}^T$ stands for transpose, and $Cov(X)$ stands for covariance of X.

CCA seeks to identify and quantify the linear associations between two sets of variables. Its usefulness stems from the fact that $\Sigma_{\theta,Y}$ may be large such that simultaneously making sense of the elements of $\Sigma_{\theta,Y}$ is unwieldy. Thus CCA summarizes the fundamental relationships contained in $\Sigma_{\theta,Y}$ into a digestible and informative manner.

Let's consider two linear combinations of θ and Y as follows:

$$U = a^T Y$$

and

$$V = b^T \theta$$

Where $a \in \mathbb{R}^{q\times 1}$ and $b \in \mathbb{R}^{K\times 1}$. Then, we have $Var(U) = a^T \Sigma_y a$; $Var(V) = b^T \Sigma_\theta b$; and $Cov(U,V) = a^T \Sigma_{y,\theta} b$. CCA finds a and b to maximize $Corr(U,V)$.

$$\underset{a,b}{Max}\, Corr(U,V) = \frac{a^T \Sigma_{y,\theta} b}{\sqrt{([a^T \Sigma_y a][b^T \Sigma_\theta b])}} \tag{A.3}$$

Observe that A.3 is invariant to the re-scaling of a and/or b; i.e.

$$\frac{\alpha a^T \Sigma_{y,\theta} b}{\sqrt{([\alpha^2 a^T \Sigma_y a][b^T \Sigma_\theta b])}} = \frac{a^T \Sigma_{y,\theta} b}{\sqrt{([a^T \Sigma_y a][b^T \Sigma_\theta b])}}$$

Since the choice of r-scaling does not affect the correlation between U and V, choosing a and b to maximize A.3 is equivalent to

$$\underset{a,b}{Max}\, a^T \Sigma_{y,\theta} b$$

subject to

$$a^T \Sigma_y a = 1$$

,

$$b^T \Sigma_\theta b = 1$$

The corresponding Lagrangian is:

$$L(\lambda, a, b) = a^T \Sigma_{y,\theta} b - \frac{\lambda_y}{2}[a^T \Sigma_y a - 1] - \frac{\lambda_\theta}{2}[b^T \Sigma_\theta b - 1]$$

The first order condition (FOC) yields:

$$\frac{\partial L}{\partial a} = \Sigma_{Y,\theta} b - \lambda_y \Sigma_y a = 0 \tag{A.4}$$

$$\frac{\partial L}{\partial b} = \Sigma_{\theta,Y} a - \lambda_\theta \Sigma_\theta b = 0 \tag{A.5}$$

Subtracting $b^T \frac{\partial L}{\partial b}$ from $a^T \frac{\partial L}{\partial a}$ gives:

$$\begin{aligned} a^T \Sigma_{Y,\theta} b - \lambda_y a^T \Sigma_y a - b^T \Sigma_{\theta,Y} a + \lambda_\theta b^T \Sigma_\theta b = 0 \\ \Longrightarrow \lambda_y a^T \Sigma_y a = \lambda_\theta b^T \Sigma_\theta b \\ \Longrightarrow \lambda_y = \lambda_\theta = \lambda \end{aligned}$$

Assuming Σ_y is nonsingular,

$$b = \frac{1}{\lambda} \Sigma_\theta^{-1} \Sigma_{\theta,Y} a \qquad (by\, equation\, A.5) \tag{A.6}$$

Substituting A.6 into A.4 gives:

$$\begin{aligned} \frac{1}{\lambda} \Sigma_{y,\theta} \Sigma_\theta^{-1} \Sigma_{\theta,Y} a - \lambda \Sigma_y a = 0 \\ \Longrightarrow \left[\Sigma_y^{-1} \Sigma_{y,\theta} \Sigma_\theta^{-1} \Sigma_{\theta,Y} - \lambda^2 I \right] a = 0 \end{aligned} \tag{A.7}$$

Thus, we have an eigen decomposition problem, where λ^2 and a are respectively the eigenvalue and eigenvector of the matrix $\Sigma_y^{-1} \Sigma_{y,\theta} \Sigma_\theta^{-1} \Sigma_{\theta,Y}$.

By replacing the solution for a in A.7 into A.6, we solve for b.

To summarize, CCA seeks a and b to maximize the correlation between U_l and V_l where $U_l = a_l^T Y$, $V_l = b_l^T \theta$. a_l is the l^{th} eigenvector of $\Sigma_y^{-1} \Sigma_{y,\theta} \Sigma_\theta^{-1} \Sigma_{\theta,Y}$; b_l is the l^{th} eigenvector of $\Sigma_\theta^{-1} \Sigma_{\theta,Y} \Sigma_Y^{-1} \Sigma_{Y,\theta}$. $l = [1, 2, ...L]$, with $L = min(q, K)$ by the property $Cov(U_i, V_j) = 0$ for $i \neq j$. Hopefully the first few U's and V's tells the main stories nested within the two datasets. Note, λ_l is the canonical correlation between U_l, and V_l. The correlations between U_l and Y give the relative importance of each variable y_i of Y in the construction of U_l. Likewise for the correlations between V_l and θ. For further exposition of CCA, see Johnson and Wichern (2007, Chapter 10)

To relate this development to the main text, referring to the R column of 7, the first four rows of the table can be understood as $Corr(U_1, Y)$, the fifth row of the table gives the canonical correlation λ, and the remaining rows of the table gives $Corr(V_1, \theta)$.

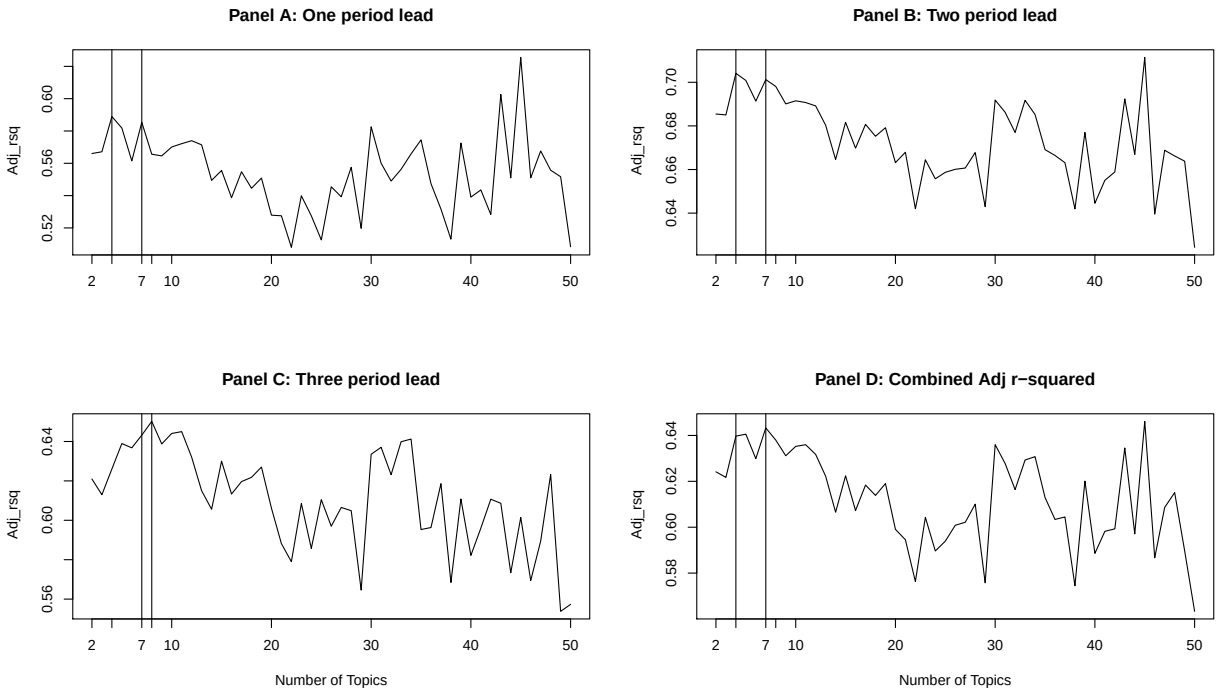


Figure 13: Values of Adjusted R-Squared as K changes. K = 4, 7, and 8 seem reasonable

B Selection of K, and parameters posterior distributions

B.1 Selecting K for the Bayesian regression model

We use a traditional model selection approach to select K. We relied on the adjusted r-squared method to select the K, the number of topics for which the regression model yields the highest adjusted r-squared. In practice, we iterate the regression model through different values for K ($K = 2,3,...50$), and for different dependent variable (one, two, and three period leads). Note that for our regression model, we use future business entry rate as our dependent variable. Figure 13 shows the values of the adjusted r-squared in the y-axis, and the K values in the x-axis. Respectively, panel A, B, C, and D represent the changes in the values of the adjusted r-squared as K changes when the dependent variables are the one, two, three period leads, and the combined values of the adjusted r-squared of the three models.

In light of Figure 13, $K = \{4,7,8\}$ seem to be all reasonable choices for K. Indeed, though there are values of K for which the adjusted r-squared are bigger than the one given by $K = \{4,7,8\}$, those K are too large and would be difficult to interpret without further variables selection methods (such as LASSO, or subset selection). Consequently, we will consider $K = 4$ as an optimum number of topics to work with. We also use $K = \{7,8\}$ as a robustness check, and found that, qualitatively, the main results hold.

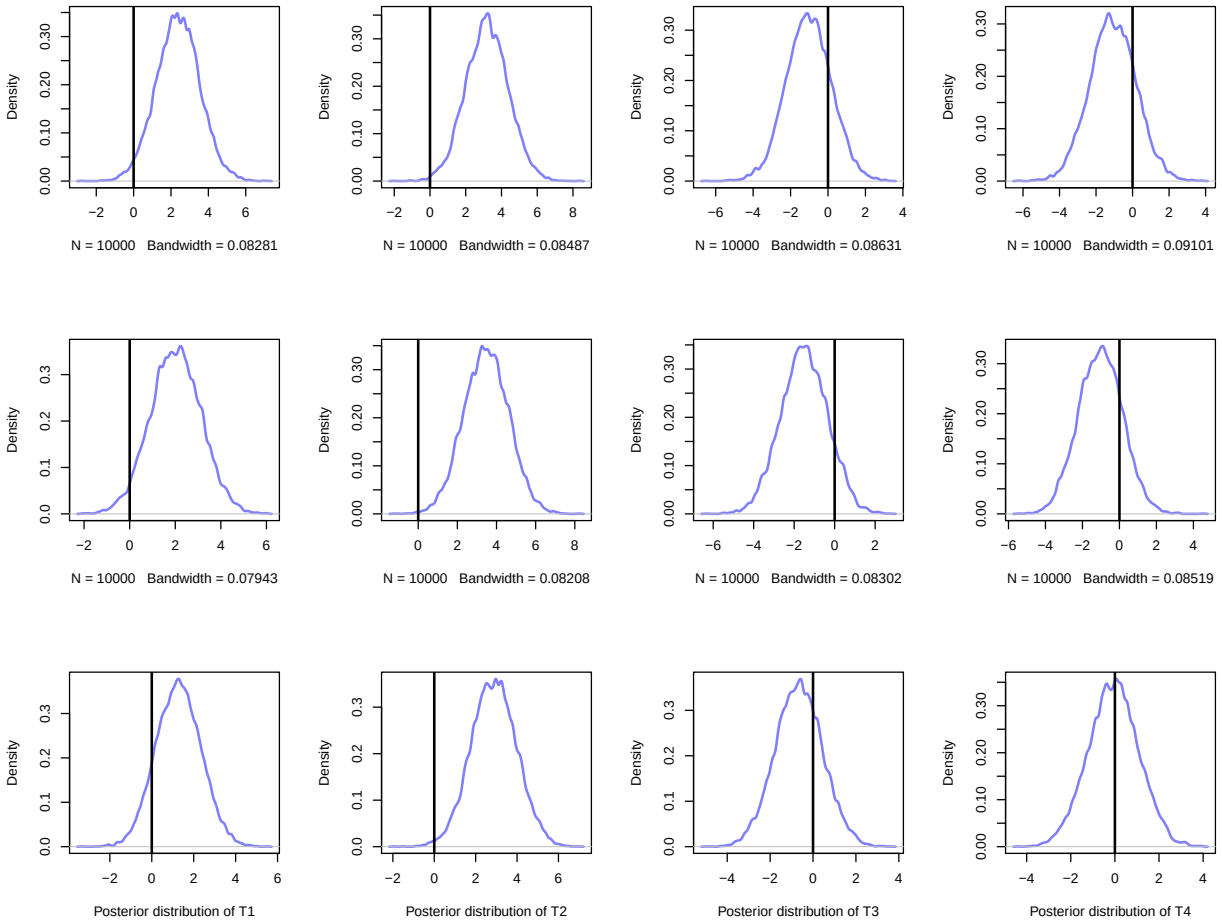


Figure 14: Posterior distributions of the topics' consistency parameter estimates. From row 1 to row 3 are regressions with one, two, and three period leads, respectively

B.2 Posterior distributions of the parameters estimates

From Fig. 14, row 1 to row 2, the vertical line represents the zero value of the parameter estimate. It appears that the posterior distribution of topic 2 is always almost greater than 0, suggesting that only positive values of topic 2 are consistent with the observed data.

Table B.1: One period lead regression results

	Mean	StdDev	2.5%	97.5%
b0	-6.90	3.00	-13.00	-1.20
b1	1.00	0.08	0.87	1.20
t1	2.40	1.20	0.08	4.70
t2	3.20	1.20	0.84	5.50
t3	-1.00	1.20	-3.40	1.40
t4	-1.10	1.30	-3.60	1.40
sigma	7.40	0.47	6.40	8.30

Table B.2: Two period lead regression results

	Mean	StdDev	2.5%	97.5%
b0	-6.60	2.80	-12.00	-1.10
b1	1.10	0.06	0.95	1.20
t1	2.00	1.10	-0.18	4.20
t2	3.50	1.10	1.30	5.70
t3	-1.50	1.20	-3.80	0.78
t4	-1.00	1.20	-3.40	1.30
sigma	7.00	0.44	6.10	7.90

Table B.3: Three period lead regression results

	Mean	StdDev	2.5%	97.5%
b0	-7.30	2.70	-13.00	-2.00
b1	1.00	0.07	0.87	1.10
t1	1.30	1.10	-0.82	3.40
t2	2.90	1.10	0.73	5.00
t3	-0.65	1.10	-2.80	1.50
t4	-0.06	1.10	-2.30	2.20
sigma	6.70	0.42	5.90	7.50

C Canonical correlation results

C.1 For K = 10

Table C.1: Statistical significance test statistics of the canonical correlations.

	id	stat	approx	df1	df2	p.value
1	Wilks	0.80	3.30	40	2,209.00	0
2	Wilks	0.92	1.90	27	1,703.00	0.004
3	Wilks	0.96	1.40	16	1,168	0.12
4	Wilks	0.98	1.50	7	585	0.15

Table C.2: Canonical correlations between the expenditure variables and the themes in the speeches

	Pair	Can_cor	r_squared
1	1	0.35	0.13
2	2	0.22	0.05
3	3	0.14	0.02
4	4	0.13	0.02

Table C.3: Correlation between the expenditures raw data and the expenditures canonical variates

	Can_var.1	Can_var.2	Can_var.3	Can_var.4
public_welfare_expend	-0.09	-0.52	0.30	0.80
highways_expend	-0.26	0.33	0.90	-0.07
health_expend	-0.39	-0.64	0.10	-0.66
educ_expend	-0.98	0.13	-0.06	0.13

Table C.4: Correlation between the raw topics and the topics canonical variates

	Can_var.1	Can_var.2	Can_var.3	Can_var.4
Topic.4	0.39	-0.09	0.06	0.07
Topic.7	0.39	0.34	0.14	-0.01
Topic.2	0.25	0.08	-0.36	-0.06
Topic.5	0.08	-0.29	0.19	-0.05
Topic.6	0.08	0.06	0.40	-0.46
Topic.3	-0.03	-0.52	0.11	0.11
Topic.8	-0.09	-0.04	0.17	-0.10
Topic.1	-0.13	0.03	-0.32	0.52
Topic.10	-0.29	0.02	-0.51	-0.36
Topic.9	-0.68	0.45	0.02	0.34

C.2 For K = 12

Table C.5: Statistical significance test statistics of the canonical correlations.

	id	stat	approx	df1	df2	p.value
1	Wilks	0.78	3.10	48	2,236.00	0
2	Wilks	0.91	1.80	33	1,712.00	0.01
3	Wilks	0.96	1.30	20	1,164	0.18
4	Wilks	0.98	1.30	9	583	0.25

Table C.6: Canonical correlations between the expenditure variables and the themes in the speeches

	Pair	Can_cor	r_squared
1	1	0.37	0.14
2	2	0.23	0.05
3	3	0.15	0.02
4	4	0.14	0.02

Table C.7: Correlation between the expenditures raw data and the expenditures canonical variates

	Can_var.1	Can_var.2	Can_var.3	Can_var.4
public_welfare_expend	-0.07	-0.53	-0.82	-0.19
highways_expend	-0.21	0.35	-0.47	0.79
health_expend	-0.41	-0.61	0.48	0.48
educ_expend	-0.98	0.14	-0.12	-0.09

Table C.8: Correlation between the raw topics and the topics canonical variates

	Can_var.1	Can_var.2	Can_var.3	Can_var.4
Topic.2	0.62	0.49	-0.25	-0.26
Topic.8	0.27	0.28	-0.21	0.19
Topic.10	0.23	-0.05	-0.05	-0.13
Topic.5	0.16	-0.26	0.07	0.20
Topic.7	0.12	0.11	0.05	0.61
Topic.6	0.11	-0.31	-0.08	0.21
Topic.3	0.11	-0.05	0.37	-0.29
Topic.4	0.004	-0.49	-0.16	0.03
Topic.9	-0.07	0.05	0.05	0.19
Topic.1	-0.31	-0.20	-0.23	-0.38
Topic.12	-0.32	0.07	0.54	-0.20
Topic.11	-0.67	0.44	-0.34	-0.15

C.3 For K = 50

Table C.9: Statistical significance test statistics of the canonical correlations.

	id	stat	approx	df1	df2	p.value
1	Wilks	0.50	2.00	200	2,166.00	0
2	Wilks	0.70	1.40	147	1,628.00	0.001
3	Wilks	0.81	1.20	96	1,088	0.07
4	Wilks	0.93	0.83	47	545	0.79

Table C.10: Canonical correlations between the expenditure variables and the themes in the speeches

	Pair	Can_cor	r_squared
1	1	0.53	0.28
2	2	0.38	0.14
3	3	0.36	0.13
4	4	0.26	0.07

Table C.11: Correlation between the expenditures raw data and the expenditures canonical variates

	Can_var.1	Can_var.2	Can_var.3	Can_var.4
public_welfare_expend	-0.20	-0.89	0.33	-0.25
highways_expend	-0.27	-0.28	-0.03	0.92
health_expend	-0.40	0.49	0.77	0.10
educ_expend	-0.97	0.01	-0.24	-0.03

Table C.12: Correlation between the raw topics and the topics canonical variates

	Can_var.1	Can_var.2	Can_var.3	Can_var.4
Topic.47	0.38	-0.11	-0.23	0.16
Topic.37	0.26	-0.29	-0.32	0.08
Topic.18	0.26	0.09	0.05	-0.02
Topic.42	0.25	0.12	0.03	-0.14
Topic.11	0.24	-0.06	-0.31	-0.05
Topic.16	0.21	0.01	0.05	0.41
Topic.8	0.20	-0.25	-0.02	0.18
Topic.33	0.20	-0.06	-0.03	0.06
Topic.24	0.20	-0.08	0.24	0.05
Topic.45	0.19	0.25	-0.15	-0.03
Topic.26	0.15	0.04	0.12	-0.19
Topic.39	0.15	0.20	-0.05	-0.03
Topic.14	0.13	0.03	0.05	0.04
Topic.50	0.09	-0.02	-0.02	-0.13
Topic.48	0.09	-0.14	0.02	-0.05
Topic.10	0.09	0.21	-0.13	-0.26
Topic.2	0.08	0.05	0.01	0.09
Topic.12	0.07	0.19	-0.05	0.24
Topic.40	0.07	0.02	0.14	0.21
Topic.20	0.06	0.16	-0.10	-0.09
Topic.27	0.04	0.08	0.03	0.05
Topic.15	0.04	0.14	-0.25	0.11
Topic.44	0.04	-0.11	-0.37	-0.07
Topic.35	0.03	0.05	0.12	0.03
Topic.31	0.03	0.02	-0.05	0.19
Topic.30	0.02	-0.23	0.28	-0.11
Topic.23	0.02	0.09	0.38	0.05
Topic.41	0.01	-0.11	0.04	0.13
Topic.1	-0.004	-0.0001	0.01	-0.38
Topic.38	-0.004	0.20	0.04	0.13
Topic.13	-0.02	0.08	-0.12	-0.002
Topic.32	-0.03	-0.01	0.09	0.04
Topic.46	-0.05	-0.05	0.23	-0.09
Topic.17	-0.06	-0.18	0.29	0.05
Topic.19	-0.06	0.23	-0.04	-0.13
Topic.3	-0.07	-0.01	-0.14	0.17
Topic.7	-0.10	-0.04	-0.08	-0.09
Topic.5	-0.11	-0.06	-0.07	0.11
Topic.21	-0.11	0.02	-0.02	-0.03
Topic.29	-0.14	0.14	-0.05	-0.07
Topic.6	-0.17	-0.08	-0.09	-0.08
Topic.4	-0.17	-0.07	0.16	-0.24
Topic.36	-0.17	-0.12	-0.19	-0.08
Topic.22	-0.18	0.03	-0.04	-0.36
Topic.28	-0.19	-0.17	0.03	-0.002
Topic.25	-0.19	-0.20	-0.11	-0.03
Topic.9	-0.24	0.14	0.01	-0.02
Topic.34	-0.26	-0.01	-0.15	-0.11
Topic.49	-0.33	0.12	0.02	0.08
Topic.43	-0.45	0.01	-0.30	0.24

D Topics Words Tables

Only the relevant topics are shown in the tables below.

D.1 For K = 10

Table D.1: Descending, ordered list of the important words for each topic

Topic 1	Topic 2	Topic 4	Topic 7	Topic 9	Topic 10
busi	reform	work	budget	health	econom
energi	govern	depart	govern	care	futur
creat	system	legisl	econom	famili	educ
compani	spend	develop	make	make	opportun
work	chang	issu	educ	work	communiti
peopl	school	effort	work	help	innov
econom	incom	servic	busi	cost	resourc
produc	rate	govern	servic	children	invest
help	billion	continu	face	insur	provid
make	busi	governor	challeng	afford	develop
develop	taxpay	public	balanc	protect	build
build	cost	communiti	dollar	provid	growth
invest	make	import	public	communiti	qualiti
nation	propos	administr	fiscal	invest	continu
train	public	program	decis	access	succeed
like	pass	citizen	citizen	peopl	ensur
industri	properti	process	difficult	citizen	grow
power	choic	member	revenu	medic	support
home	growth	offic	tough	drug	challeng
high	administr	area	effici	expand	plan
million	competit	look	creat	support	transport
technolog	local	resourc	save	home	commit
worker	four	improv	spend	program	univers
manufactur	better	address	reduc	school	strong
just	good	encourag	fund	plan	water
countri	financ	employe	recess	safe	servic
workforc	lower	system	futur	live	prosper
plant	rais	educ	governor	legislatur	best
grow	benefit	respons	agenc	colleg	land
good	econom	general	respons	continu	vision

D.2 For K = 12

Table D.2: Descending, ordered list of the important words for each topic

Topic 1	Topic 2	Topic 8	Topic 10	Topic 11	Topic 12
energi	busi	budget	develop	health	econom
compani	work	econom	communiti	care	educ
creat	govern	public	work	famili	invest
produc	make	face	public	cost	futur
develop	servic	govern	provid	help	colleg
busi	peopl	challeng	effort	make	opportun
peopl	budget	futur	program	insur	innov
nation	econom	make	resourc	work	grow
build	better	fiscal	improv	children	univers
help	employe	difficult	servic	afford	high
work	help	crisi	commit	provid	communiti
million	creat	balanc	ensur	citizen	build
industri	depart	respons	initi	access	continu
power	save	come	addit	school	best
econom	agenc	problem	protect	worker	growth
home	educ	revenu	includ	invest	qualiti
make	offic	california	safe	protect	world
train	effici	recess	succeed	peopl	busi
invest	general	decis	propos	plan	make
manufactur	continu	tough	respons	creat	higher
plant	focus	children	plan	busi	work
like	best	protect	support	medic	succeed
technolog	cost	choic	system	support	strong
renew	keep	governor	administr	expand	support
plan	move	plan	continu	nation	prosper
project	worker	recoveri	water	join	fund
just	hard	take	drug	coverag	global
research	good	peopl	center	child	life
dollar	employ	ahead	local	safe	centuri
market	look	turn	goal	prescript	young

D.3 For K = 50

Table D.3: Descending, ordered list of the(Einav and Levin, 2014; Varian, 2014) important words for each topic

Topic 18	Topic 34	Topic 37	Topic 43	Topic 47	Topic 49
nation	energi	depart	health	govern	world
rate	power	believ	care	servic	innov
educ	renew	focus	insur	employe	technolog
incom	produc	make	cost	cost	centuri
rank	wind	board	afford	work	global
improv	electr	process	famili	effici	econom
growth	effici	director	coverag	save	grow
averag	clean	govern	medic	better	scienc
past	like	execut	access	make	prepar
busi	compani	team	reduc	agenc	lead
continu	econom	agenc	expand	improv	high
better	make	manag	provid	taxpay	famili
higher	build	budget	help	health	chang
best	just	servic	home	reduc	make
made	keep	final	hospit	citizen	futur
succeed	reduc	educ	doctor	effect	generat
number	generat	includ	plan	like	home
import	natur	posit	premium	spend	challeng
good	nation	exampl	cover	look	competit
decad	industri	local	uninsur	respons	fuel
reduc	lead	grant	protect	move	leader
citizen	sourc	structur	pass	money	compet
unemploy	fuel	differ	prevent	offic	research
system	plant	commiss	increas	result	industri
econom	take	better	direct	elimin	engin
lowest	resourc	number	offer	growth	depend
program	save	offic	child	administr	design
lower	price	order	join	chang	transform
place	conserv	approach	nurs	creat	succeed
just	exampl	attract	system	consolid	afford

E Highlight of Key Words in Texts

We constructed a software which highlight the key words of a given topic in a chosen text. Here, we show the highlight of the higher education topic (Topic 10) key words in two speeches in which the topic is highly represented.


no precautions against the storm as it gathers, we risk becoming a bitterly divided society. If we do not make a meaningful commitment to higher education, we will see generations of young people lose hope, as they face a doubly harsh dilemma: Higher education becoming more and more essential for success in the new economy . . . yet increasingly available only to those fortunate enough to be born into a family of wealth. If we do not take the next steps in Smart Growth, we shut out thousands of families as parts of our State fall into decay, while at the same time we will see others living not just insulated--but isolated--behind gated communities. And if we do not truly embrace justice, fairness and inclusion for all, we will see the same divisiveness that hurt our nation for so long continue to plague us . . . making our dreams of justice, fairness and inclusion a mere illusion. I speak for all Marylanders when I say we must--and we will--reject these possibilities. We will come together--just as we have year after year--and embrace policies that free our individual and collective potential. As always, we must start with education . . . especially higher education. We can all be proud of the renewed emphasis we have given to higher education in the past six years: Together, we have increased significantly our support for higher education since the Lt. Governor and I took office: We more than doubled financial aid with new merit scholarships and a major increase in "need based" aid. Senator Hoffman, I thank you for leading the fight on this effort. We strengthened support for our Historically Black Institutions. Senator Blount, you have been a true champion on this issue. We also embarked upon a $1.2 billion campus construction program to build new science and technology facilities on campuses across Maryland. President Miller, I thank you for your leadership in the area. We must be mindful, however, that much of our work together was merely to "catch up" to where Maryland should have been. As we all know, higher education bore a disproportionate burden of budget cuts in the recession of the early 1990s. We are just now fully recovering. Having brought stability to our colleges and universities, we must look towards the future with a determination to go beyond the status quo with dramatic, bold steps. Last year in my State of the State address, I shared my vision for higher education. I said then and I repeat today: "I want the word 'tuition' to be seen as an anachronism. All children will move into college just as they now move from junior high to high school. Maryland's institutions of higher education will be among the best in the country . . . and they will be free". This vision is just as critical today. I taught at the University of Maryland, College Park for 27 years. It is my strong belief that as individuals--and as a society--we will only reach our full potential if we encourage and enable our citizens to pursue knowledge for its own sake. At the same time, I also understand the responsibility we have to produce men and women ready to excel in the new economy. Access to higher education is essential in achieving both a civil society and a thriving economy. This view is supported by a recent Wall Street Journal article highlighting factors high-tech industry leaders consider when making location decisions. The most important factor was access to a skilled and educated workforce. The second most important factor was proximity to world-class research institutions, including colleges and universities. In contrast, financial incentives came in last . . . tenth out of ten. There is widespread agreement that higher education is the engine that will propel our society into a brighter, more prosperous future. Unfortunately, too many of our citizens are priced out of the college classroom and--unacceptably--out of promising careers and successful lives. Today, a sound Kindergarten through twelfth grade education is not sufficient. We never say to a tenth grader "sorry, you cannot afford the next two years--you have to quit school". To do so would be outrageous and unacceptable. Yet we do the exact same thing two years later. That is morally wrong. Higher education for all is a necessity. The fact is, we are moving aggressively towards our vision of making higher education a universal experience. We capped tuition increases and introduced the HOPE Scholarship Program to make college more affordable and more accessible. We owe it to future generations to do even more. We must work together towards the goal of universal access to higher education. I hold no illusion that this will happen in the next few years or that it will be easy to accomplish. If, however, our State and our Nation are to continue to prosper, we must make progress towards the goal of opening our colleges and universities to everyone. We also want excellence at our centers of higher education. We are moving towards that goal. St. Mary's College is recognized as the nation's best public liberal arts college. Morgan State is well known for educating many of our nation's African-American leaders. University of Maryland, College Park and University of Maryland, Baltimore County are home to some of the nation's leading research centers. And Prince George's Community College was recently named a national model for undergraduate education. With the financial commitment we will make to higher education this year, we continue our progress towards State-wide excellence. Our budget contains a $1.3 billion investment in higher education . . . an investment in our future. This means an increase of almost 70% since the Lt. Governor and I took office! Money to make our colleges and universities the best in the nation. Money to help more Marylanders pursue their dream of a prosperous future through higher education. We also must find ways to take bold steps in extending the reach and broadening the impact of our Smart Growth program. We should all be proud that Smart

Figure E.1: Excerpt of the SoSA of the governor of Maryland in 2001 (Education, Science, Innovation and Technology)

momentum and sets the stage for a prosperous decade. We will invest in Nebraskas future by focusing on economic growth and jobs. We will invest in Nebraskas young people by prioritizing education and by focusing on education accountability. Economic success and education success are linked together. We need both. We are focused on creating higher paying jobs and developing a more highly educated work force. We want our graduates and young professionals to be prepared for high-quality, high-skill jobs with dynamic companies doing business right here in Nebraska. In preparing for our future, the Nebraska Department of Economic Development conducted a review of Nebraskas economic standing. The resulting Battelle study was a thorough assessment of Nebraskas competitive advantages that suggested strategies for growing new and innovative jobs, industries and talent. The study revealed that Nebraska has succeeded in developing an unusually diverse economy and a number of industries are ideally positioned for new growth. Twelve industries have a strong presence in Nebraska with additional potential to grow, including: agriculture and food processing, financial services, biosciences, computer and software services, renewable energy, transportation, warehousing and logistics, research and development and engineering services, health services, business management and administrative services, hospitality and tourism, precision metals manufacturing and agricultural machinery. The Battelle study provided recommendations on how to strengthen support for the companies that make up our fastest growing sectors and the people they employ. Another report prepared by your Innovation and Entrepreneurial Task Force, chaired by Senator Conrad, outlined the need to improve Nebraskas entrepreneurial environment. Many of the recommendations in your legislative report are similar to the Battelle study. It is critical that we invest in economic growth and jobs. Thats why I am pleased to announce today the Talent and Innovation Initiative, a four-part plan designed to enhance our economic momentum. First, I am proposing a Nebraska Internship Program to increase the number of college and university students interning with Nebraska businesses. This $1.5 million training program will be funded by redirecting resources from the Nebraska Job Training Cash Fund and matched by funds from the private sector. Second, I am proposing the creation of the Business Innovation Act to leverage entrepreneurship, to increase private sector research and innovation, and to expand small business outreach efforts. This $7 million program would be funded by redirecting resources within the Department of Economic Development and new general funds. Third, I am proposing the creation of a new Site and Building Development Fund to increase the number of sites and buildings available for business development projects. This fund is needed now in order to continue Nebraskas economic growth. This $3 million fund would be created by redirecting resources from the Nebraska Affordable Housing Trust Fund. Fourth, I am proposing a $5 million Angel Investment Tax Credit Program to foster high-tech startups in Nebraska. The Angel Investment Tax Credit Program is key to increasing the number of higher paying jobs in our state. These investments in economic growth would be combined with two new education initiatives. First, my budget recommendations support the Department of Education, the University of Nebraska and Nebraskas P-16 Initiative in their joint efforts to develop a virtual high school. A rigorous online high school curriculum offers important opportunities to rural Nebraska and urban areas alike. The $8.5 million initiative will be funded from lottery funds. A virtual high school would allow Nebraska high school students to take courses ranging from basic Spanish classes to advanced placement courses. In rural Nebraska, it can be difficult to hire foreign language, math and science teachers. A virtual high school would allow rural schools and rural communities the opportunity to survive. Online courses allow students to complete course work on their timetable in the evenings or on weekends. A virtual high school is a way to expand learning beyond the traditional school day and school year. My second education proposal is a one-time $25 million investment in the University of Nebraskas Innovation Campus. This proposal would jump start and accelerate the development of Innovation Campus. The University of Nebraska is a critical component to our states economic future. With its pending move to the Big Ten, the University of Nebraska has an outstanding opportunity to significantly increase student enrollment, expand its rapidly growing research base and develop public-private partnerships at Innovation Campus that will increase job opportunities for Nebraskans. This bold investment is needed now, not five years from now. Additionally, I am very supportive of Senator Ashfords efforts to reduce truancy. Last year, 22,000 Nebraska students missed more than 20 days of school, and students cant learn if they are not in school. For example, Commissioner of Education Roger Breed has informed me that students who miss more than 20 days of school score approximately 30 points less on the reading assessment. Many schools would see a significant increase in reading scores if truancy were reduced. Even though Nebraska has a nearly $1 billion projected shortfall, our two-year budget prioritizes education. State funded state aid to education in FY12 remains at $810 million and increases by $50 million to $860 million in FY13. I am not proposing any reduction in higher education funding for the University of Nebraska, our state colleges and Nebraskas community colleges. In order to prioritize education and economic growth, my budget proposal significantly reduces funding for many agencies and eliminates several programs. Many of the proposals in your LR542 report have been included in my budget recommendations. The decisions were difficult but

Figure E.2: Excerpt of the SoSA of the governor of Nebraska in 2011 (Education, Science, Innovation and Technology)

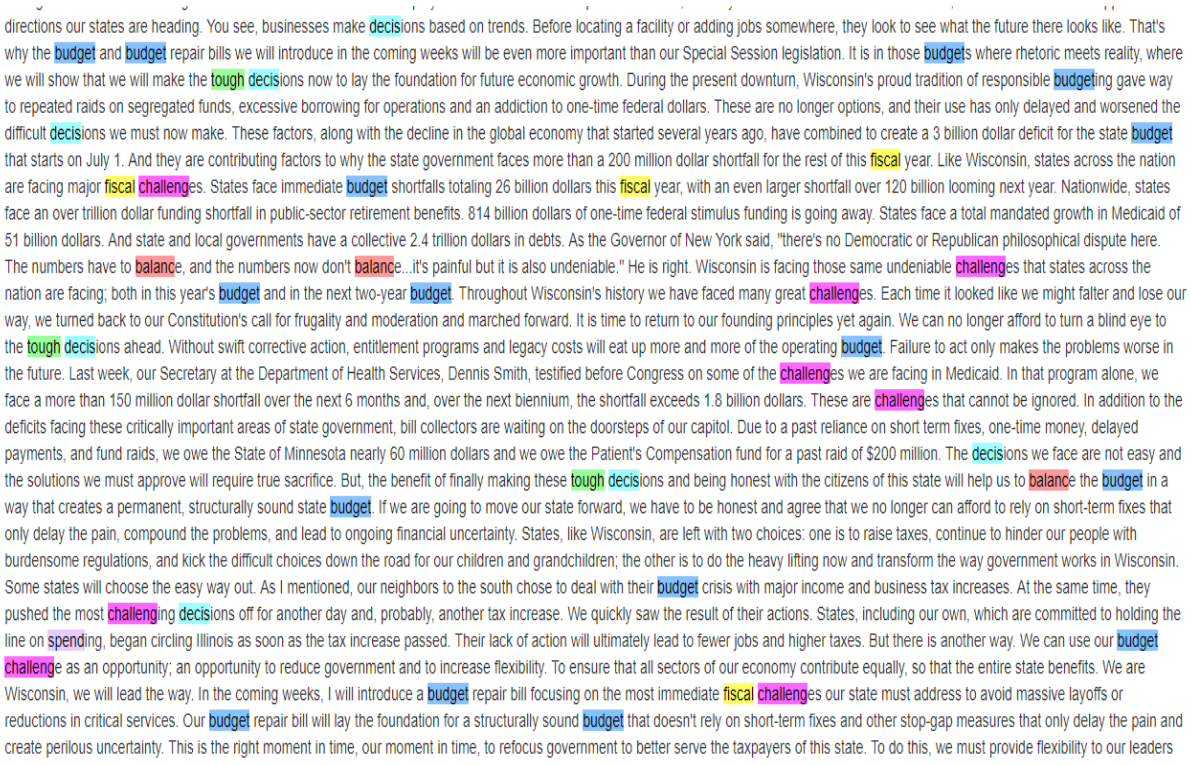

directions our states are heading. You see, businesses make decisions based on trends. Before locating a facility or adding jobs somewhere, they look to see what the future there looks like. That's why the budget and budget repair bills we will introduce in the coming weeks will be even more important than our Special Session legislation. It is in those budgets where rhetoric meets reality, where we will show that we will make the tough decisions now to lay the foundation for future economic growth. During the present downturn, Wisconsin's proud tradition of responsible budgeting gave way to repeated raids on segregated funds, excessive borrowing for operations and an addiction to one-time federal dollars. These are no longer options, and their use has only delayed and worsened the difficult decisions we must now make. These factors, along with the decline in the global economy that started several years ago, have combined to create a 3 billion dollar deficit for the state budget that starts on July 1. And they are contributing factors to why the state government faces more than a 200 million dollar shortfall for the rest of this fiscal year. Like Wisconsin, states across the nation are facing major fiscal challenges. States face immediate budget shortfalls totaling 26 billion dollars this fiscal year, with an even larger shortfall over 120 billion looming next year. Nationwide, states face an over trillion dollar funding shortfall in public-sector retirement benefits. 814 billion dollars of one-time federal stimulus funding is going away. States face a total mandated growth in Medicaid of 51 billion dollars. And state and local governments have a collective 2.4 trillion dollars in debts. As the Governor of New York said, "there's no Democratic or Republican philosophical dispute here. The numbers have to balance, and the numbers now don't balance...it's painful but it is also undeniable." He is right. Wisconsin is facing those same undeniable challenges that states across the nation are facing; both in this year's budget and in the next two-year budget. Throughout Wisconsin's history we have faced many great challenges. Each time it looked like we might falter and lose our way, we turned back to our Constitution's call for frugality and moderation and marched forward. It is time to return to our founding principles yet again. We can no longer afford to turn a blind eye to the tough decisions ahead. Without swift corrective action, entitlement programs and legacy costs will eat up more and more of the operating budget. Failure to act only makes the problems worse in the future. Last week, our Secretary at the Department of Health Services, Dennis Smith, testified before Congress on some of the challenges we are facing in Medicaid. In that program alone, we face a more than 150 million dollar shortfall over the next 6 months and, over the next biennium, the shortfall exceeds 1.8 billion dollars. These are challenges that cannot be ignored. In addition to the deficits facing these critically important areas of state government, bill collectors are waiting on the doorsteps of our capitol. Due to a past reliance on short term fixes, one-time money, delayed payments, and fund raids, we owe the State of Minnesota nearly 60 million dollars and we owe the Patient's Compensation fund for a past raid of $200 million. The decisions we face are not easy and the solutions we must approve will require true sacrifice. But, the benefit of finally making these tough decisions and being honest with the citizens of this state will help us to balance the budget in a way that creates a permanent, structurally sound state budget. If we are going to move our state forward, we have to be honest and agree that we no longer can afford to rely on short-term fixes that only delay the pain, compound the problems, and lead to ongoing financial uncertainty. States, like Wisconsin, are left with two choices: one is to raise taxes, continue to hinder our people with burdensome regulations, and kick the difficult choices down the road for our children and grandchildren; the other is to do the heavy lifting now and transform the way government works in Wisconsin. Some states will choose the easy way out. As I mentioned, our neighbors to the south chose to deal with their budget crisis with major income and business tax increases. At the same time, they pushed the most challenging decisions off for another day and, probably, another tax increase. We quickly saw the result of their actions. States, including our own, which are committed to holding the line on spending, began circling Illinois as soon as the tax increase passed. Their lack of action will ultimately lead to fewer jobs and higher taxes. But there is another way. We can use our budget challenge as an opportunity; an opportunity to reduce government and to increase flexibility. To ensure that all sectors of our economy contribute equally, so that the entire state benefits. We are Wisconsin, we will lead the way. In the coming weeks, I will introduce a budget repair bill focusing on the most immediate fiscal challenges our state must address to avoid massive layoffs or reductions in critical services. Our budget repair bill will lay the foundation for a structurally sound budget that doesn't rely on short-term fixes and other stop-gap measures that only delay the pain and create perilous uncertainty. This is the right moment in time, our moment in time, to refocus government to better serve the taxpayers of this state. To do this, we must provide flexibility to our leaders

Figure E.3: Excerpt of the SoSA of the governor of Iowa 2011 (Topic 7, Balanced Budget)

F Correlations Plots

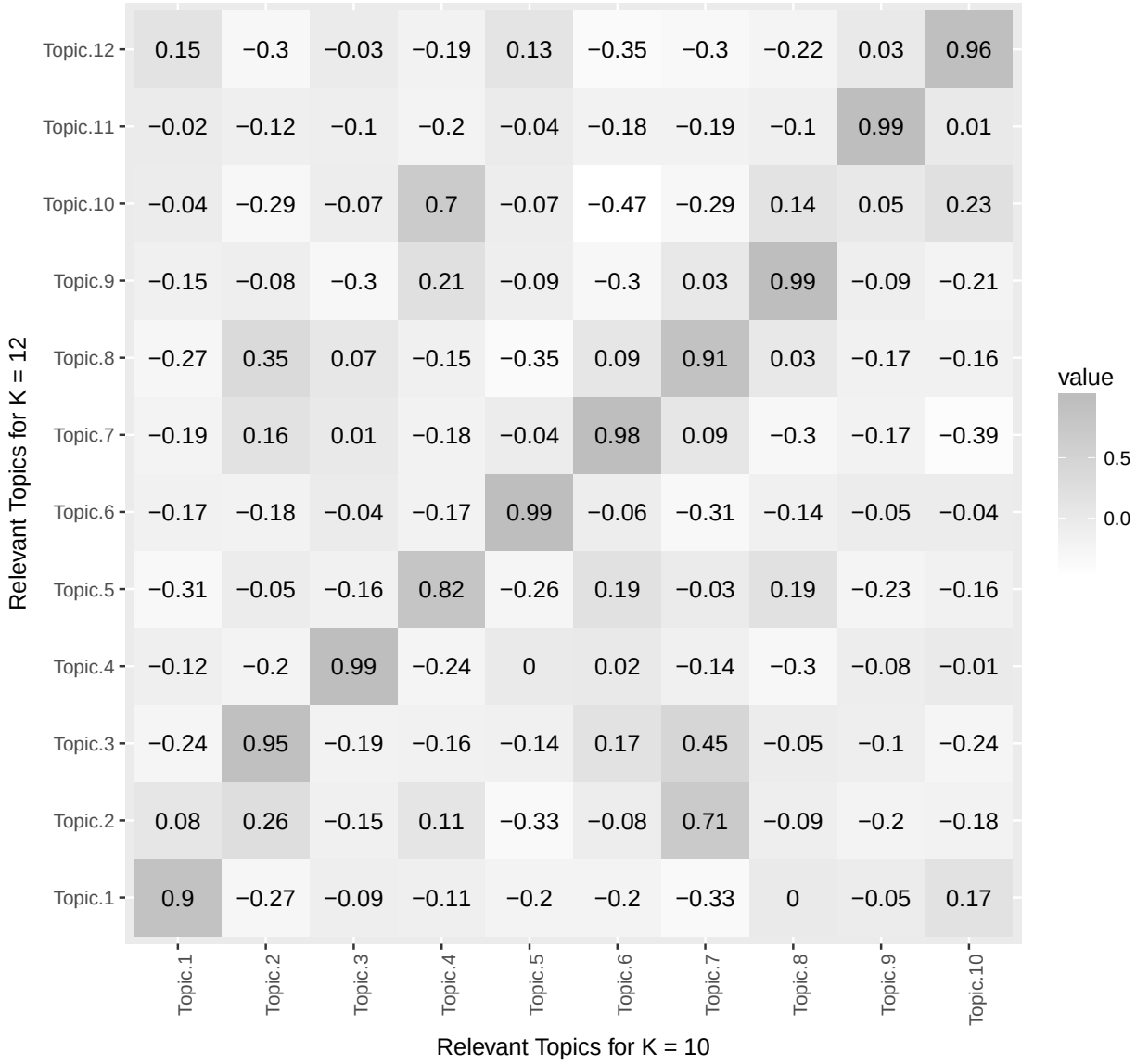


Figure F.1: Correlations between the topics when K = 10, and the topics when K = 12.

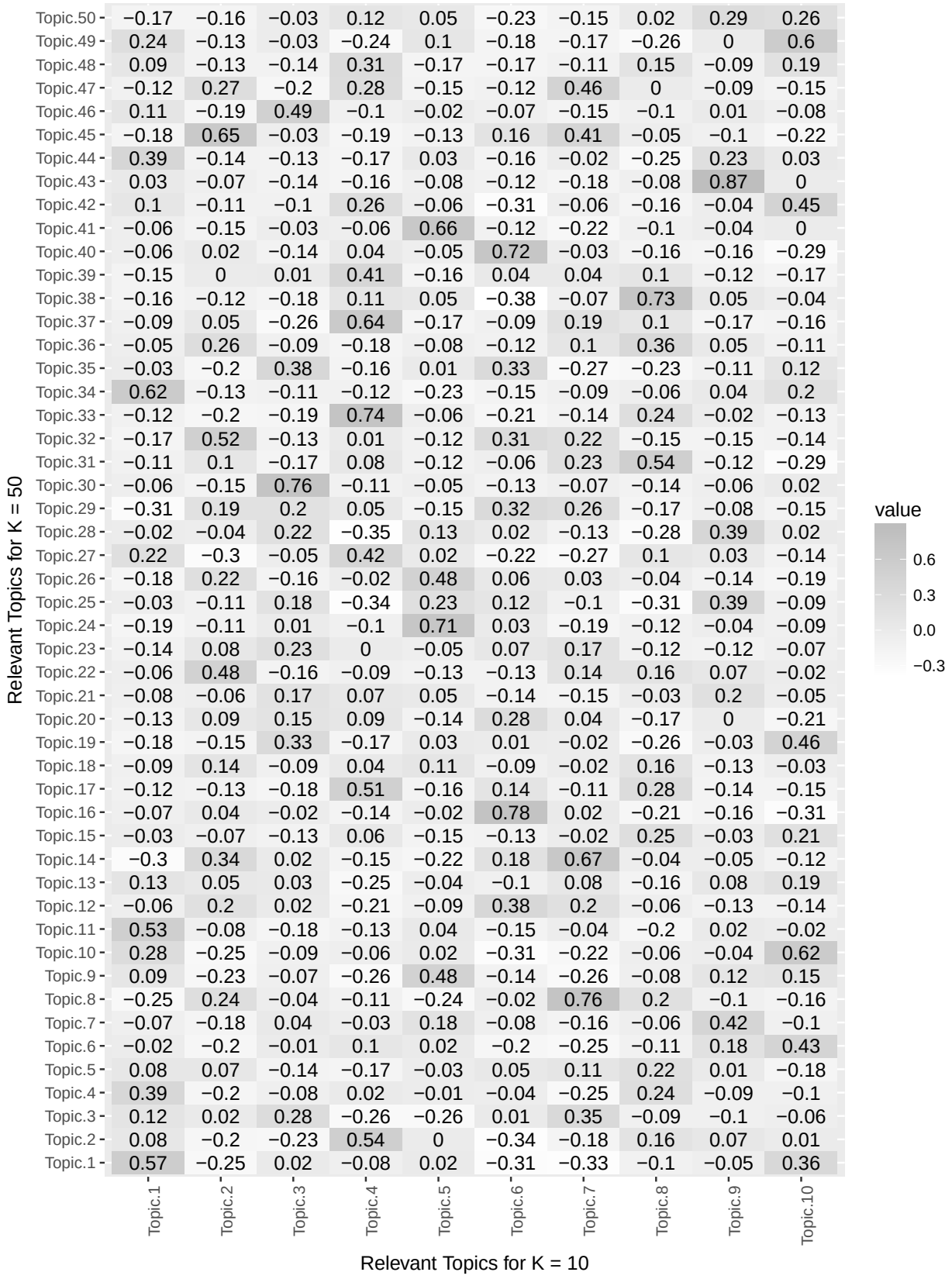


Figure F.2: Correlations between the topics when K = 10, and the topics when K = 50.