1 Appendix for "What do Economic Education Scholars Study? Insights from Machine Learning"

Topic Model

In this section, we use a topic model developed in the computer science literature to estimate the number of latent topics across a dataset of economic education scholarly papers. Specifically, we use the Latent Dirichlet Allocation topic model presented in the seminal paper by Blei, Ng, and Jordan (2003). The LDA model is the most commonly used unsupervised probabilistic machine learning method for a document corpus. The LDA model assumes that each document is a collection of latent topic groups. These topic groups are assumed to share a common Dirichlet prior. The topics themselves are assumed to be represented by a probability distribution of words. The probability distribution of the words within a topic are also assumed to follow a Dirichlet prior.

Given a document corpus D with I documents, where each document i has N_i words, the LDA models D as a mixture of T topics where there are a total of number of terms $V = \sum_{i=1}^{I} N_i$. Let each topic $t \in 1, \ldots, T$ be generated by

$$\phi_t \sim Dir(\beta),$$

where ϕ_t is of length V and each document $i \in 1, \ldots, I$ be generated by

$$\theta_i \sim Dir(\alpha).$$

where θ_i is of length T for each document. Next, we select topic

$$z_{in}|\theta_i \sim Multinomial(\theta_i),$$

where z_{in} is the topic for the *n-th* word in document i. Finally, each word w_{in} is selected such that

$$w_{in}|z_{in} \sim Multinomial(\phi_{z_{in}}),$$

$n \in 1, \ldots, N_i$ is an index for all the words in document i. The hyperparameters α and β affect the documents distributions over topics and the topics distribution over words, respectively.

The LDA models is constructed as two mixtures. First, for a given topic there is a probability distribution of words. The parameter ϕ_t contains the mixture weights across all words. Second, for a given document there is a probability distribution of topics. The parameter θ_i contains the mixture weights across all topics.

The probability of observing the data contain in document corpus D is given by the likelihood function

$$p(D|\alpha, \beta) = \prod_{i=1}^{I} \int p(\theta_i|\alpha) \sum_{n=1}^{N_i} p(z_{in}|\theta_i) p(w_{in}|z_{in}, \phi) p(\phi|\beta) d\theta_i$$

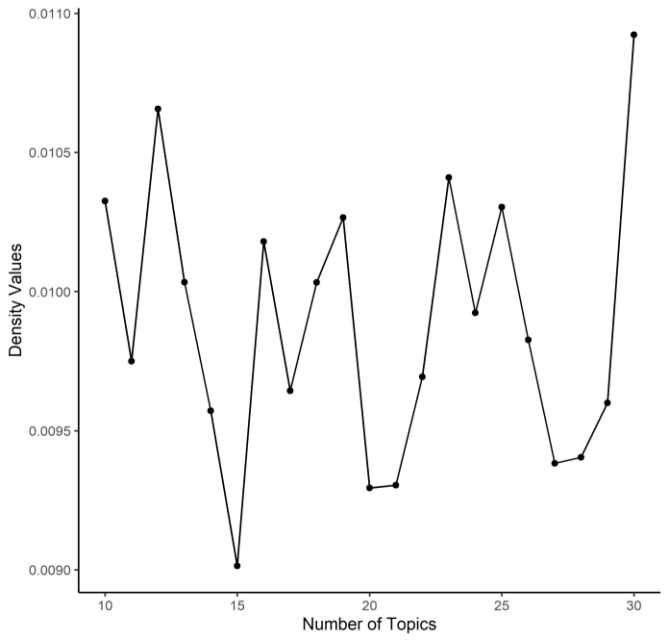
The model parameters are estimated by selecting α and β to maximize the likelihood function.

There are a few challenges to estimate the parameters of the likelihood function. First, integration of the mixture model can be computationally intensive. To account for this complication, we adopt the Gibbs Sampling method developed by Phan, Nguyen, and Horiguchi (2008). Second, the number of topics is fixed within the estimation routine. Several methods exists to choose the appropriate number of topics. These methods include running the LDA model over several values of T and the selecting the value that maximizes the Log-Likelihood function. Other variants use a penalty function for increasing number of topics and select the optimal number of topics using an Akaike information criterion (AIC) or Bayesian information criterion (BIC) measure. We follow Cao et al. (2009), which adaptively selects the best LDA model as determined by the density

function of topics. We estimate the LDA model for $T \in 10,\ldots,30$ and map the density values in figure below.[i] Using this method, we find 15 topics is the optimal value.[ii] As a robustness check, we repeat the estimation randomly choosing different seed values for the Gibbs Sampler. We find the optimal number of topics varies between 12 and 18 using the Cao et al. (2009) density measure (see Figure 1).

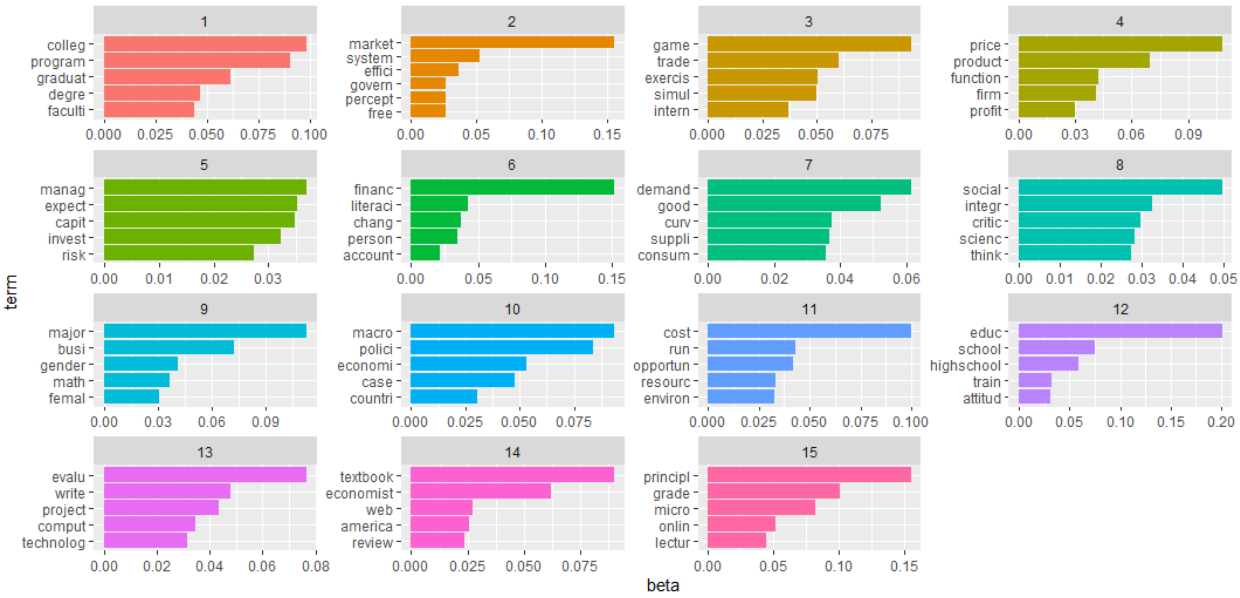
Words are ranked within topic by probability of association with the topic. In Figure 2, we illustrate the relative weight each word has on the topic assignment. We utilize words found in the top ten per topic to construct the thematic names whenever possible. For example, Topic 10 is labelled Macro Policy because the two words with the highest probability of being associated with this latent topic is *macro* and *policy*.

Figure 1: Density Value by Number of Topics



Note: The optimal number of topics is 15 since it provides us with the lowest density score.

Figure 2: The 15 topics and top 5 word probabilities


1
colleg
program
graduat
degre
faculti
0.000 0.025 0.050 0.075 0.100
2
market
system
effici
govern
percept
free
0.00 0.05 0.10 0.15
3
game
trade
exercis
simul
intern
0.000 0.025 0.050 0.075
4
price
product
function
firm
profit
0.00 0.03 0.06 0.09
5
manag
expect
capit
invest
risk
0.00 0.01 0.02 0.03
6
financ
literaci
chang
person
account
0.00 0.05 0.10 0.15
7
demand
good
curv
suppli
consum
0.00 0.02 0.04 0.06
8
social
integr
critic
scienc
think
0.00 0.01 0.02 0.03 0.04 0.05
9
major
busi
gender
math
femal
0.00 0.03 0.06 0.09
10
macro
polici
economi
case
countri
0.000 0.025 0.050 0.075
11
cost
run
opportun
resourc
environ
0.000 0.025 0.050 0.075 0.100
12
educ
school
highschool
train
attitud
0.00 0.05 0.10 0.15 0.20
13
evalu
write
project
comput
technolog
0.00 0.02 0.04 0.06 0.08
14
textbook
economist
web
america
review
0.000 0.025 0.050 0.075
15
principl
grade
micro
onlin
lectur
0.00 0.05 0.10 0.15
term
beta

[i] The Gibbs Sampler Estimator of the LDA model can be sensitive to the initial seed value. For this reason, we use 200 different seed values to obtain a better estimate of the Cao et. al. (2009) density function.

[ii] We use to R packages to estimate the LDA model. First, we find the optimal number of topics using the *ldatuneing* package, which contains the Cao et al. (2009) topic density metric (see more at http://rpubs.com/siri/ldatuning). Second, we use the R package *topicmodeling* to estimate the LDA model. https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf