Online Technical Appendix
Hansen, McMahon, and Prat (2017)

This appendix details various technical aspects of the estimation of latent Dirichlet allocation for the paper "Transparency and Deliberation within the FOMC: a Computational Linguistics Approach". It gives background on the Dirichlet distribution, defines LDA and derives Gibbs sampling equations,[1] and explains how we form aggregate topic distributions. All source code is available on `https://github.com/sekhansen/text-mining-tutorial`, and an example of implementing the analysis is worked through on `http://nbviewer.ipython.org/github/sekhansen/text-mining-tutorial/blob/master/tutorial_notebook.ipynb`.

1 Properties of the Dirichlet distribution

A Dirichlet random variable is characterized in terms of a parameter vector $\alpha = (\alpha_1, \ldots, \alpha_K)$ and has a probability density function given by

$$\text{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

where

$$B(\alpha) \equiv \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \text{ and } \alpha_0 \equiv \sum_{k=1}^{K} \alpha_k.$$

Now suppose we have count data of the form $X = (x_1, \ldots, x_n)$ where each x_i is a single draw from a categorical distribution with parameter $\theta = (\theta_1, \ldots, \theta_K)$; that is, $\Pr[\, x_i = k \mid \theta \,] = \theta_k$. The probability of observing a particular X is then

$$\Pr[\, X \mid \theta \,] = \prod_{k=1}^{K} \theta_k^{N_k}$$

where N_k is the number of observations drawn in category k. If we place a Dirichlet prior on the θ parameters, then the posterior satisfies

$$\Pr[\, \theta \mid X, \alpha \,] \propto \Pr[\, X \mid \theta \,] \Pr[\, \theta \,] \propto \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

[1] Heinrich (2009) contains additional material.

which after normalization gives

$$\Pr[\,\theta \mid X, \alpha\,] = \frac{1}{B(\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)} \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1},$$

a Dirichlet with parameters $(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$. In other words, the Dirichlet is conjugate to the categorical distribution.

Given conjugacy, we can derive a closed-form expression for the model evidence $\Pr[\,X \mid \alpha\,]$, which must satisfy

$$\Pr[\,\theta \mid X, \alpha\,] = \frac{\Pr[\,\theta \mid \alpha\,] \Pr[\,X \mid \theta, \alpha\,]}{\Pr[\,X \mid \alpha\,]}.$$

We have given expressions for all these probabilities except $\Pr[\,X \mid \alpha\,]$ above—note that $\Pr[\,X \mid \theta, \alpha\,] = \Pr[\,X \mid \theta\,]$, and simple substitution gives

$$\Pr[\,X \mid \alpha\,] = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{1}$$

where $N = \sum_k N_k$.

2 LDA: Model and Estimation

The data generating process of Latent Dirichlet Allocation is the following:

1. Draw β_k independently for $k = 1, \ldots, K$ from Dirichlet(η).
2. Draw θ_d independently for $d = 1, \ldots, D$ from Dirichlet(α).
3. Each word $w_{d,n}$ in document d is generated from a two-step process:
   (a) Draw topic assignment $z_{d,n}$ from θ_d.
   (b) Draw $w_{d,n}$ from $\beta_{z_{d,n}}$.

The observed data is $W = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$, while the unobserved data is $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_D)$. Given the statistical structure of LDA, the joint likelihood factors as follows:

$$\Pr[\,W, Z \mid \alpha, \eta\,] = \Pr[\,W \mid Z, \eta\,] \Pr[\,Z \mid \alpha\,].$$

Expressions for each factor are easy to derive from results in the previous section. Let m_k^d be the number of words from topic k in document d. By (1), the probability of observing this in one document is

$$\frac{\Gamma(K\alpha)}{\Gamma(N_d + K\alpha)} \prod_k \frac{\Gamma(m_k^d + \alpha)}{\Gamma(\alpha)}.$$

Where $N_d = \sum_k m_k^d$ is the total number of words in document d. So, the probability of observing the topic assignments across all documents is

$$\prod_d \frac{\Gamma(K\alpha)}{\Gamma(N_d + K\alpha)} \prod_k \frac{\Gamma(m_k^d + \alpha)}{\Gamma(\alpha)} = \left[\frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)}\right]^D \prod_d \frac{\prod_k \Gamma(m_k^d + \alpha)}{\Gamma(N_d + K\alpha)}.$$

By similar calculations we obtain

$$\Pr[W \mid Z, \eta] = \left[\frac{\Gamma(V\eta)}{\Gamma^V(\eta)}\right]^K \prod_k \frac{\prod_v \Gamma(m_v^k + \eta)}{\Gamma(\sum_v m_v^k + V\eta)},$$

where m_v^k be the number of times token v is assigned to topic k.

In this way, we have written the probability of the data (W, Z) in terms of word and topic assignment counts, and eliminated the θ_d and β_k parameters. This expression facilitates the derivation of a collapsed Gibbs sampler.

2.1 Full conditional distribution

To construct a collapsed Gibbs sampler, we need to compute the probability that $z_{d,n} = k$ given the other topic assignments $Z_{-(d,n)}$ and words W. By Bayes' rule, we have

$$\Pr\left[\, z_{d,n} = k \;\middle|\; Z_{-(d,n)}, W, \alpha, \eta \,\right] = \frac{\Pr\left[\, z_{d,n} = k, Z_{-(d,n)}, W \;\middle|\; \alpha, \eta \,\right]}{\Pr\left[\, Z_{-(d,n)}, W \;\middle|\; \alpha, \eta \,\right]} =$$
$$\frac{\Pr\left[\, W \;\middle|\; z_{d,n} = k, Z_{-(d,n)}, \eta \,\right] \Pr\left[\, z_{d,n} = k, Z_{-(d,n)} \;\middle|\; \alpha \,\right]}{\Pr\left[\, W \;\middle|\; Z_{-(d,n)}, \eta \,\right] \Pr\left[\, Z_{-(d,n)} \;\middle|\; \alpha \,\right]} \propto$$
$$\frac{\Pr\left[\, W \;\middle|\; z_{d,n} = k, Z_{-(d,n)}, \eta \,\right] \Pr\left[\, z_{d,n} = k, Z_{-(d,n)} \;\middle|\; \alpha \,\right]}{\Pr\left[\, W_{-(d,n)} \;\middle|\; Z_{-(d,n)}, \eta \,\right] \Pr\left[\, Z_{-(d,n)} \;\middle|\; \alpha \,\right]}.$$

The final step follows first from $\Pr\left[\, W \;\middle|\; Z_{-(d,n)}, \eta \,\right] = \Pr[\, w_{d,n} \,] \Pr\left[\, W_{-(d,n)} \;\middle|\; Z_{-(d,n)}, \eta \,\right]$ given that $z_{d,n}$—which generates $w_{d,n}$—is drawn independently of $Z_{-(d,n)}$.

From this expression, one can directly compute terms from the equations above. Let's rewrite $\Pr\left[\, z_{d,n} = k, Z_{-(d,n)} \;\middle|\; \alpha \,\right]$ in the following way, equivalent to the expression above

$$\left[\prod_{d' \neq d} \frac{\Gamma(K\alpha)}{\Gamma(n_{d'} + K\alpha)} \prod_K \frac{\Gamma(m_k^{d'}) + \alpha}{\alpha}\right] \left[\prod_{k' \neq k} \frac{\Gamma(m_{k'}^d + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(K\alpha)}{\Gamma(n_d + K\alpha)} \frac{\Gamma(m_k^d + \alpha)}{\Gamma(\alpha)} =$$
$$\left[\prod_{d' \neq d} \frac{\Gamma(K\alpha)}{\Gamma(n_{d'} + K\alpha)} \prod_K \frac{\Gamma(m_k^{d'}) + \alpha}{\alpha}\right] \left[\prod_{k' \neq k} \frac{\Gamma(m_{k'}^d + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(K\alpha)}{\Gamma(n_d + K\alpha)} \frac{\Gamma(m_{k,-n}^d + 1 + \alpha)}{\Gamma(\alpha)}$$

where $m_{k,-n}^d$ is the number of slots in document d assigned to topic k excluding the nth

slot. One can similarly write $\Pr\left[\, Z_{-(d,n)} \;\middle|\; \alpha \,\right]$ as

$$\left[\prod_{d' \neq d} \frac{\Gamma(K\alpha)}{\Gamma(n_{d'} + K\alpha)} \prod_k \frac{\Gamma(m_k^{d'}) + \alpha}{\alpha}\right] \left[\prod_{k' \neq k} \frac{\Gamma(m_{k'}^d + \alpha)}{\Gamma(\alpha)}\right] \frac{\Gamma(K\alpha)}{\Gamma(n_d - 1 + K\alpha)} \frac{\Gamma(m_{k,-n}^d + \alpha)}{\Gamma(\alpha)}$$

Using the fact that $\frac{\Gamma(x+1)}{\Gamma(x)} = x$, we obtain that

$$\frac{\Pr\left[\, z_{d,n} = k, Z_{-(d,n)} \;\middle|\; \alpha \,\right]}{\Pr\left[\, Z_{-(d,n)} \;\middle|\; \alpha \,\right]} = \frac{m_{k,-n}^d + \alpha}{n_d - 1 + K\alpha}.$$

Similarly, one can write $\Pr\left[\, W \;\middle|\; z_{d,n} = k, Z_{-(d,n)}, \eta \,\right]$ as

$$\left[\prod_{k' \neq k} \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_v^{k'} + V\eta)} \prod_v \frac{\Gamma(m_v^{k'} + \eta)}{\Gamma(\eta)}\right] \left[\prod_{v' \neq v} \frac{\Gamma(m_{v'}^k + \eta)}{\Gamma(\eta)}\right] \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_v^k + V\eta)} \frac{\Gamma(m_v^k + \eta)}{\Gamma(\eta)} =$$
$$\left[\prod_{k' \neq k} \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_v^{k'} + V\eta)} \prod_v \frac{\Gamma(m_v^{k'} + \eta)}{\Gamma(\eta)}\right] \left[\prod_{v' \neq v} \frac{\Gamma(m_{v'}^k + \eta)}{\Gamma(\eta)}\right] \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_v^k + V\eta)} \frac{\Gamma(m_{v,-(d,n)}^k + 1 + \eta)}{\Gamma(\eta)}$$

where v is implicitly the token assignment of the nth slot of document d. Moreover $\Pr\left[\, W \;\middle|\; Z_{-(d,n)}, \eta \,\right]$ becomes

$$\left[\prod_{k' \neq k} \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_v^{k'} + V\eta)} \prod_v \frac{\Gamma(m_v^{k'} + \eta)}{\Gamma(\eta)}\right] \left[\prod_{v' \neq v} \frac{\Gamma(m_{v'}^k + \eta)}{\Gamma(\eta)}\right] \frac{\Gamma(V\eta)}{\Gamma(\sum_v m_{v,-(d,n)}^k + V\eta)} \frac{\Gamma(m_{v,-(d,n)}^k + \eta)}{\Gamma(\eta)}.$$

We know that $\sum_v m_{v,-(d,n)}^k + 1 = \sum_v m_v^{k'}$: because the nth slot of document d with token v has been assigned topic k, if we exclude it from the dataset, then the total number of times the token v is assigned to topic k is one less. So

$$\frac{\Pr\left[\, W \;\middle|\; z_{d,n} = k, Z_{-(d,n)}, \eta \,\right]}{\Pr\left[\, W_{-(d,n)} \;\middle|\; Z_{-(d,n)}, \eta \,\right]} = \frac{m_{v,-(d,n)}^k + \eta}{\sum_v m_{v,-(d,n)}^k + V\eta}.$$

We conclude that

$$\Pr\left[\, z_{d,n} = k \;\middle|\; Z_{-(d,n)}, W, \alpha, \eta \,\right] \propto \frac{m_{v,-(d,n)}^k + \eta}{\sum_v m_{v,-(d,n)}^k + V\eta} \left(m_{k,-n}^d + \alpha\right).$$

Note that we can exclude the $n_d - 1 + K\alpha$ term since it doesn't vary with k.

2.2 Gibbs sampling algorithm

From the conditional distribution computed above, we implement the following algorithm:

1. Randomly allocate to each token in the corpus a topic assignment drawn uniformly

from $\{1, \ldots, K\}$.

2. For each token, sequentially draw a new topic assignment via multinomial sampling where

$$\Pr\left[\, z_{d,n} = k \;\middle|\; Z_{-(d,n)}, W, \alpha, \eta \,\right] \propto \frac{m^k_{v,-(d,n)} + \eta}{\sum_v m^k_{v,-(d,n)} + V\eta} \left(m^d_{k,-n} + \alpha\right).$$

3. Repeat step 2 4,000 times as a burn in phase.

4. Repeat step 2 4,000 more times, and store every 50th sample.

2.3 Predictive distributions

The collapsed Gibbs sampler gives an estimate of each word's topic assignment, but not the parameters θ_d and β_k since these are collapsed out of the likelihood function. In order to describe a document's topic distribution and the topics themselves, the following predictive distributions are used.

$$\widehat{\beta}^v_k = \frac{m^k_v + \eta}{\sum_{v=1}^V \left(m^k_v + \eta\right)} \tag{2}$$

and

$$\widehat{\theta}^k_d = \frac{m^d_k + \alpha}{\sum_{k=1}^K \left(m^d_k + \alpha\right)}. \tag{3}$$

The derivation of the predictive distribution for a Dirichlet is standard, see for example Murphy (2012) section 3.4.4.

2.4 Convergence

As with all Markov Chain Monte Carlo methods, the realized value of any one chain depends on the random starting values and determining if a chain has converged is important. To address these concerns, for each specification of the model we run five different chains starting from five different initial seeds. Along each chain we then compute the value of the model's *perplexity* at regular intervals. Perplexity is a common measure of fit in the natural language processing literature. The formula is

$$\exp\left[-\frac{\sum_{d=1}^D \sum_{v=1}^V n_{d,v} \log\left(\sum_{k=1}^K \widehat{\theta}^k_d \widehat{\beta}^v_k\right)}{\sum_{d=1}^D N_d}\right]$$

where $n_{d,v}$ is the number of times word v occurs in document d.

Table 1: Perplexity scores for five chains (40-topic model)

Iteration	Chain 1	Chain 2	Chain 3	Chain 4	Chain 5
4000	1002.015275	1000.84559	1001.237111	1002.678028	1000.940867
5000	1001.984839	1001.151012	1001.403589	1002.654002	1001.017522
6000	1002.106541	1001.08223	1001.862532	1002.640741	1000.594744
7000	1002.094333	1000.698283	1001.677683	1001.914573	1000.766348
8000	1002.242386	1000.935695	1001.559882	1002.37538	1000.462472
Mean	1002.093755	1000.939742	1001.581274	1002.331845	1000.711352
SD	0.207715083	0.217669923	0.193039164	0.250924429	0.282573422

Table 1 reports values of perplexity along five chains drawn for the 40-topic model at various iterations.[2] Various features are worth noting. First, from the 4,000th iteration onwards the perplexity values are quite stable, indicating that the chains have converged. Second, the differences in perplexity across chains are marginal, indicating that the estimates are not especially sensitive to starting values. Third, chain 5 performs marginally better in terms of average perplexity, so we use it in our baseline analysis. Fourth, chain 3 achieves the lowest standard deviation, so we use it for the robustness check described in section 7 of the main paper.

3 Estimating aggregate document distributions

As explained in the text, we are more interested in the topic distributions at the meeting-speaker-section level rather than at the individual statement level. Denote by $\theta_{i,t,s}$ the topic distribution of the aggregate document. Let $w_{i,t,s,n}$ be the nth word in the document, $z_{i,t,s,n}$ its topic assignment, $v_{i,t,s,n}$ its token index, $m_k^{i,t,s}$ the number of words in the document assigned to topic k, and $m_{k,-n}^{i,t,s}$ the number of words besides the nth word assigned to topic k. To re-sample the distribution $\theta_{i,t,s}$, for each iteration $j \in \{4050, 4100, \ldots, 8000\}$ of the Gibbs sampler:

1. Form $m_k^{i,t,s}$ from the topic assignments of all the words that compose the aggregate document (i, t, s) from the Gibbs sampling.

2. Drop $w_{i,t,s,n}$ from the sample and form the count $m_{k,-n}^{i,t,s}$.

3. Assign a new topic for word $w_{i,t,s,n}$ by sampling from

$$\Pr\left[\, z_{i,t,s,n} = k \;\middle|\; z_{-(i,t,s,n)}, \mathbf{w}_{i,t,s} \,\right] \propto \widehat{\beta}_k^{v_{i,t,s,n}} \left(m_{k,-n}^{i,t,s} + \alpha\right) \tag{4}$$

   where $z_{-(i,t,s,n)}$ is the vector of topic assignments in document (i, s, t) excluding word n and $\mathbf{w}_{i,t,s}$ is the vector of words in the document.

[2] We only report values at a limited number of iterations here for space. The same analysis done over finer intervals yields very similar results.

4. Proceed sequentially through all words.

5. Repeat 20 times.

We then obtain the aggregate document predictive distribution

$$\widehat{\theta}_{i,t,s}^{k} = \frac{m_{k}^{i,t,s} + \alpha}{\sum_{k=1}^{K} \left(m_{k}^{i,t,s} + \alpha \right)}. \tag{5}$$

This is identical to the regular Gibbs sampling procedure except for two differences. First, the topics are kept fixed at the values of their predictive distributions rather than estimated. Second, because topics do not need to be estimated, many fewer iterations are needed to sample topic assignments for each document.

References

Heinrich, G. (2009). Parameter estimation for text analysis. Technical report, vsonix GmbH and University of Leipzig.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. MIT Press.