CS 281 Final Project

Applying Correlated Topic Models and Latent Dirichlet Allocation for Hypothesis-Free Discovery in RNA-seq Experimental Data

Ankit Gupta

December 2015

1 Abstract

Next-generation sequencing, including RNA-sequencing (RNA-seq), has introduced a wealth of data into the hands of biologists, offering them a novel means of making biological inferences. However, there exist few techniques to try to understand genomic expression patterns from raw RNA-seq data in a hypothesis-free manner. This is further complicated by the sparsity introduced through single cell RNA-seq, in which individual cells are sequenced, and only a small fraction of their expressed genes can be expected to be detected. To address these issues, this paper implements and applies Correlated Topic Models (CTM) to discover structure in RNA-seq data, and determine the relationships between sets of genes. A fast variational EM algorithm was implemented and trained on various types of RNA-seq datasets. Ultimately, it was found that CTM can perform at least as well as other methods of dimensionality reduction, such as Principal Component Analysis or Latent Dirichlet Allocation, while offering more expressive and explanatory power. This technique can allow researchers to investigate relationships between genes without hypotheses apriori. Implementations can be found at `https://github.com/ankitvgupta/rnaseqtopicmodeling`.

2 Introduction

When working with biological data, there are many ways to test specific hypotheses, such as whether certain genes are up or down regulated when comparing samples. There are many software packages that can do this differential expression analysis, and this has led to many biological discoveries. However, these tools are generally more useful for testing particular hypotheses from data, rather than trying to discover structure without as much previous knowledge.

This paper attempts to apply topic modeling strategies to biological data in order to do exactly this form of analysis. Specifically, the data being analyzed is RNA-seq expression data. RNA-seq is a technique for measuring the expression of genes in a sample by sequencing the RNA molecules in the cells, and then aligning those sequences to known reference genomes. The result of these process is "feature counts", which effectively can be thought of as the number of RNA molecules expressed at each gene locus in a reference genome, and thus they provide a measure of how much each gene is expressed.

This data representation is similar, in many ways, to the word-count vector data representation broadly used in the analysis of textual data. In particular, in that field, "documents" are generally characterized by the number of times each word in a known fixed-size vocabulary occurs. In the case of the biological data, a "document" is a sample that was sequenced, and a "word" is a particular gene. Using this characterization, the analysis done on textual data can be applied to RNA-seq data, and inferences can be made about the biological data set.

One of the techniques that can be used to do this type of unstructured analysis is Latent Dirichlet Allocation (LDA) [2]. LDA is a technique to discover latent topics that underlie textual data. In the case of

biological data, a topic can be thought of as a group of genes that perhaps have some sort of coexpression. This means that LDA, in theory, can be used on raw RNA-seq data to determine whether it has structure in terms of groups of genes that are coexpressed.

However, there are a number of limiting aspects of this model. For one, one of the assumptions that LDA makes is that the topics are independent. In essence, this independence of the topics suggests that one set of coexpressed genes can be thought of as largely independent of other sets of genes. However, we know in biology that this is likely not the case. Instead, there are intuitively sets of genes that are more related to some sets than others, as we might expect stomach-related genes to be more related to those of other digestive organs than neurons, or for genes from a muscle tissue to be more related to other muscle tissues than skin tissue. Thus, it seems prudent that in order to more thoroughly understand how biological phenomena are modeled through topics, that we be able to make inferences about the correlations between topics as well, which is exactly what a Correlated Topic Model (CTM) [1] offers.

It is important to note that the way that these correlations are being discovered is through the cooccurance of genes in the same sample. Thus, they are not directly arising from, say, an observation that a gene "downregulates" another, which would be powerful. However, there is still value in being able to see that some topics are negatively correlated with others, as this suggests that certain groups of genes may be expressed when others are not, and scientists can use this to conduct further experiments.

Furthermore, there is an important application of this project in single cell RNA-seq. Whereas in bulk RNA-seq, a sample to be sequenced is taken from a tissue (meaning many cells combined together), and the sample size allows for a large fraction of the expressed genes to successfully be sampled, in single cell RNA-seq, many fewer genes are sampled. In this technique, one single cell is sequenced as thoroughly as possible, and it is often difficult to get more than a few thousand genes from any particular cell, even if it's expressing many more. Thus, it would be useful to analyze this potentially sparse data. When topic modeling is successfully applied to text corpora, such as news articles, it is expected that each article only contains a very small subset of the words in the vocabulary. In other words, topic modeling seems to still perform well in settings with sparse data.

Thus, there are a number of motivations for this project. The mostly important is to be able to use CTM to find groups of related genes, and to infer the relationships between them to make biological inferences. However, to do this, we must be able to assess whether CTM is adequately modeling the original samples. CTM, just like LDA, can be used as a dimensionality reduction technique, as the per-document distribution over the topics can be used to reduce the original high-dimensional data (usually in the tens of thousands of genes) into a lower-dimensional space. Then, classifiers can be trained on a subset of this low-dimensional data with the original samples' labels (such as which organ they came from), and verified against a held-out test set. If CTM performs well on this task relative to other dimensionality reduction methods or to classification using the raw data, that would support the claim that CTM is modeling the meaningful parts of the data effectively, and thus we can confidently analyze the topics that it discovers.

In general, the experiments in this paper are used to show that as a dimensionality reduction technique, CTM performs well compared to various baselines. However, the primary use-case of this method will likely not be as a dimensionality reduction technique. Instead, the success of the dimensionality reduction shows that CTM captures the interesting variation in the original data, which lends credibility to true purpose: being able to find topics in the genes and show the relationships between them.

3 Technical Background

In this section, the two main machine learning models being investigated, Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM), will be discussed. First, LDA will briefly be covered, and then CTM will be explained in greater depth, as it was implemented from scratch in this paper. LDA was effectively used as a baseline, and as such its implementation was drawn from the python library "lda".

LDA is a generative model (Figure 1) in which a document can be viewed as mixture of topics, and a topic can be viewed as a distribution over a fixed-size vocabulary. Furthermore, the topic distribution in

LDA has a Dirichlet prior, hence the naming. LDA is widely used and is well-described in [2]. The reader should refer to [2] for a review if necessary.

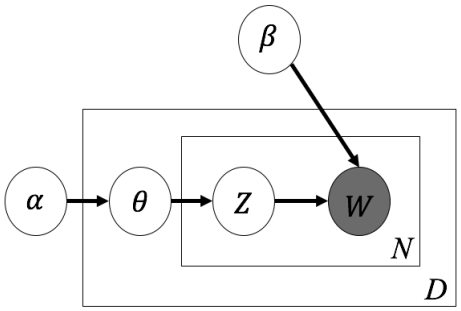

Figure 1: **LDA Graphical Model**
Recreated from [7].

CTM is another generative model (Figure 2) that is, in many ways, similar to LDA. However, rather than using Dirichlet distributions and priors, CTM uses the logistic normal distribution. In particular, the generative process first requires sampling an η from a multlvariate Gaussian, and then mapping it to the simplex, getting a sample from the logistic normal distribution. Let this mapping be $f(\eta)$. The Gaussian gives the covariance matrix we will eventually be interested in. Then, the rest follows a sampling technique similar to LDA, where we draw a topic assignment $Z_n \sim Mult(f(\eta))$ and word $w_n \sim Mult(\beta_{z_n})$ where β is the distribution over the vocabulary for each topic. In other words, as stated in the original paper [1], this process is much like LDA, except that the topic proportions $f(\eta)$ are the result of a draw from a logistic normal distribution, rather than from a dirichlet.

Unfortunately, due to the use of a logistic normal distribution, we lose the conjugacy between the Dirichlet and multinomial that LDA is able to exploit to have relatively simple collapsed Gibbs sampling updates. Thus, the derivation of mean-field variation inference algorithm used to learn the variational parameters in this model and infer the optimal model parameters is explained in the next section.

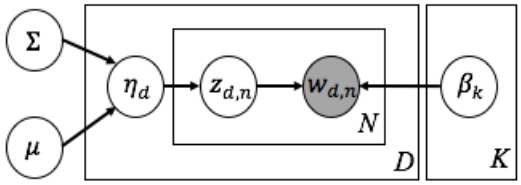

Figure 2: **CTM Graphical Model**
Recreated based on image from [1].

4 Model

In CTM with K topics, we are fundamentally trying to learn three parameters: $\beta_{1:K}$, which is the distribution over the vocabulary for each topic, as well as μ and Σ, which are the parameters of the logistic normal distribution. Each of the documents has a η_d sampled from the μ, Σ. We can do this inference problem using variational EM with mean-field variational inference. The variational distribution of the η_d is K univariate Gaussians $\{\lambda_i, \nu_i\}$. The variational distributions of the topic assignments of each word in a document $z_{1:N}$

are specified by N K-dimensional multinomial parameters $\phi_{1:N}$. Using the mean-field method, we get

$$q(\eta_{1:K}, z_{1:N}|\lambda_{1:K}, \nu^2_{1:K}, \phi_{1:N}) = \prod_{i=1}^{K} q(\eta_i|\lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n|\phi_n)$$

Now, we can use Jensen's inequality to bound the log likelihood of a particular document:

$$\log p(w_{1:N}|\mu, \Sigma, \beta) \geq E_q[\log p(\eta|\mu, \Sigma)] + \sum_{n=1}^{N} E_q[\log p(z_n|\eta) + \sum_{n=1}^{N} E_q[\log p(w_n|z_n, \beta)] + H(q)$$

This is just as in equation 21.10 of [6]. Variational inference will optimize this equation with respect to the variational parameters, thereby tightening the bound, and thus finding the variational distribution q that minimizes the KL divergence between it and the true posterior.

The actual calculations needed to do this optimization amount to taking the gradients with respect to the various parameters. These calculations are detailed in [1], but they are briefly summarized here for completeness. Much of these were re-derived in implementing the CTM.

The first term of the above bound is

$$E_q[\log p(\eta|\mu, \Sigma) = \frac{1}{2}\log|\Sigma^{-1}| - \frac{K}{2}\log 2\pi - \frac{1}{2}\mathrm{Tr}(\mathrm{diag}(\nu^2)\sigma^{-1}) + (\lambda - \mu)^T\Sigma^{-1}(\lambda - \mu)$$

To simplify the second term, we will have to introduce another variational parameter. To start, the second term is

$$E_q[\log p(z_n|\eta) = E_q[\eta^T z_n] - E_q[\log(\sum_{i=1}^{K} \exp(\eta_i))]$$

Now, we upper bound the negative term, and get that

$$E_q[\log(\sum_{i=1}^{K} \exp(\eta_i))] \leq \zeta^{-1}(\sum_{i=1}^{K} E_q[\exp(\eta_i)]) - 1 + \log\zeta$$

where there is a new variational parameter ζ that is needed because the upper bound is being calculated using a Taylor expansion. The expectation term is just the expectation of the a log normal distribution, which is just $\exp(\lambda_i + \nu_i^2/2)$ Thus, the second term is

$$E_q[\log p(z_n|\eta) = \sum_{i=1}^{K} \lambda_i\phi_{n,i} - \zeta^{-1}(\sum_{i=1}^{K} \exp(\lambda_i + \nu_i^2/2)) + 1 - \log\zeta$$

Then, more simply, the third term of the equation is

$$E_q[\log p(w_n|z_n, \beta) = \sum_{i=1}^{K} \phi_{n,i}\log\beta_{i,w_n}$$

and the fourth is the entropy, which is

$$\sum_{i=1}^{K} \frac{1}{2}(\log\nu_i^2 + \log 2\pi + 1) - \sum_{n=1}^{N}\sum_{i=1}^{K} \phi_{n,i}\log\phi_{n,i}$$

Now, and finally, we can calculate the updates by calculating the gradients, and we get that

$$\hat{\zeta} = \sum_{i=1}^{K} \exp(\lambda_i + \nu_i^2/2)$$

$$\hat{\phi}_{n,i} \propto \exp(\lambda_i)\beta_{i,w_n}$$

Unfortunately, for the other parameters, a gradient-based optimizer is needed since the solutions cannot be calculated directly, and so for these last two, the gradients were determined and used in the implementation using `scipy.minimize` with the appropriate minimizer.

$$\frac{dL}{d\lambda} = -\Sigma^{-1}(\lambda - \mu) + \sum_{n=1}^{N} \phi_{n,1:K} - (N/\zeta)\exp(\lambda + \nu^2/2)$$

$$\frac{dL}{d\nu_i^2} = \Sigma_{ii}^{-1}/2 - N/(2\zeta)\exp(\lambda_i + \nu_i^2/2) + \frac{1}{2\nu_i^2}$$

5 Inference

To carry out the parameter estimation, a variational EM algorithm had to be implemented. In the E-step, the likelihood bound with respect to the variational parameters above was maximized. This meant that fast implementations of each of those updates were written. In the M-step, the model parameters were updated using the learned variational parameters from the E-step, according to the following update equations from [1].

$$\hat{\beta}_i \propto \sum_d \phi_{d,i} n_d$$

$$\hat{\mu} = \frac{1}{D}\sum_d \lambda_d$$

$$\hat{\Sigma} = \frac{1}{D}\sum_d I\nu_d^2 + (\lambda_d - \hat{\mu})(\lambda_d - \hat{\mu})^T$$

The variational EM algorithm was implemented in Python, and the only outside packages used were numpy and scipy for doing the linear algebraic operations and gradient-based optimization. Furthermore, since the variational inference occurs on each document separately for a given iteration of the M-step, that stage could be parallelized. Thus, the Python Multiprocessing module was used to efficiently use a multicore machine, and ultimately the majority of the CTM-based inference experiments were run on Harvard's Odyssey computing cluster on 64-core machines. This dramatically increased the performance of the algorithm, and this potential for parallelizing the M-step is a major benefit of this algorithm. Note that to determine baselines and other parts of the project, libraries such as the "lda" python library and scikit-learn for classification were used. However, the core machine learning model that was being investigated, CTM, was written from scratch with numpy.

6 Experiments

There were many stages of this projects, which meant that a number of experiments were conducted. Centrally, there were two major factors that were being measured. The first was whether correlated topic models could function as a viable dimensionality reduction strategy. The purpose of this approach was to determine whether CTM could be used to embed a very sparse high-dimensional document (a sample could have tens of thousands of genes, many of which will be barely expressed), in a lower dimensional space, while still maintaining the salient properties needed to determine the identity of the sample (such as the organ it was from). For example, in [2], the authors showed that distribution over topics for each document in LDA could be used as an effective dimensionality reduction technique, which did quite well on classification tasks relative to various baselines.

Thus, for the correlated topic model, it was important to understand whether the CTM could function as an effective dimensionality reduction technique, and the variational parameters for the topic distributions

were used (as suggested in [1]) to do this reduction. Then, the dimensionality-reduced documents were broken into training and test sets, and a classifier was trained (out of a standard machine learning package) using the given sample identities, and the test set accuracy was determined with 5-fold cross-validation. The same process was performed on dimensionality reduced data from LDA, and the same using Principal Components Analysis (PCA), another common technique that can be applied to dimensionality reduction, with the same number of components as topics. Lastly, these results were compared to what was found by using the raw features, and it was determined whether the dimensionality reduction affected the performance of the classification significantly.

That was the majority of the quantitative work. The next stage was more qualitative, and thus is more open-ended and will be worked on more in the future. However, this part of the technique may be most promising in terms of generating biologically meaningful results. In particular, the various topics that LDA and CTM output were analyzed by looking at the genes corresponding to the highest parameter values for each topic. Those genes were used to identify whether the topic appeared to correspond to a biologically meaningful phenomenon (as it often did), and the correlation values between topics were considered. This is definitely a portion of the project that requires additional quantitative, rather than qualitative, analysis. In the future, additional methods such as gene set enrichment analysis will be used to quantitatively determine whether the genes associated with a given topic are associated with biological phenomena. At the moment, a basic version of gene set enrichment was used from `geneontology.org`, which is detailed in the Results and Discussion section.

7 Results and Discussion

7.1 Experiment 1: Proof of Concept

The first experiment done was a simple proof of concept. The data was the gene expression for several dozen sex-related genes from 30 female and 40 male samples. The goal was to see whether both LDA and CTM could effectively find "female" topics and "male topics". Indeed, when each of these algorithms were run, they quickly converged to solutions where there were topics with only female genes, and topics with only male genes. Furthermore, the correlated topic model's covariance matrix indicated a positive correlation between male topics, and a negative correlation between male and female topics, consistent with the biology we would expect.

7.2 Experiment 2: Rat Transcriptome

Once Experiment 1 justified that there was some validity in pursuing the CTM and LDA as methods of discovering gene expression structure, this technique was applied to a more varied and interesting dataset. In particular, the dataset was a recently published rat transcriptome [8], with samples from 11 different organs, at 4 time points, with both sexes, and many biological replicates. The classification being tested was the identity of the organ that the samples were from. The total number of unique genes expressed across all of the samples was around 30000. For this part, the 2000 genes with the highest variance amongst samples were kept (the other genes quickly dropped in variance), and their expression levels were used as the input to the various topic models. Then, a Logistic Regression classifier was trained on the inferred topic distributions, found using CTM, for each sample, and the same was done for the topic distributions from LDA, dimensionality reduction done through PCA, and on the raw gene counts. In Figure 3 are the results attained from classification on the test set (5-fold cross-validated). Note that in each of the columns, the same number of topics/dimensions are used.

There are a few interesting takeaways from here. For one, all of the dimensionality reduction techniques lead to a reduction in classification accuracy relative to the raw feature counts when there were few dimensions, which is reasonable, since 5 dimensions may not be enough to capture all of the variation. However, CTM appeared to perform better on this dataset than did PCA or LDA effectively across the board, except for in cases with lots of principal components, which makes sense since having 40+ principal components

Features	Classification Accuracy by #Topics/Components				
	5 Topics	15 Topics	25 Topics	35 Topics	45 Topics
CTM	.88	.99	.993	1.0	.996
PCA	.63	.975	.99	.996	1.0
LDA	.59	.89	.969	.972	.956
Raw	.98				

Figure 3: **Rat Dataset Results**

is likely capturing almost all of the patterns in the data. Most importantly, it can be seen that CTM is performing better than LDA on this dataset, which is promising, as it suggests that the independence assumption between topics in LDA may be leading to suboptimal topic assignments. Furthermore, with as few as 5 topics, CTM was able to far outshine PCA as a dimensionality reduction method.

Then, the topics that this generated were analyzed. For the time being, relatively manual means to determine the structure in the topics were employed by essentially looking up sets of genes with the basic "enrichment analysis" tool on `geneontology.org`, based on a gene classification system called PANTHER [4] [5], to determine if a set of genes is highly associated with any biological processes or molecular functions. This led to some interesting results. In particular, in the 5-topic situation, fairly clear topics were generated. One was associated with skeletal and cardiac muscle assembly (muscular organs), one with progesterone and testosterone processes (sex organs and hormone-secretion organs), one with brain development, one with cell-cell adhesion genes (such as platelet aggregation), and one related to fat/lipid processing. Interestingly, the sex-related topic seemed correlated to the brain-development topic, which can perhaps be attributed to the hormonal control of many sexual organs, where the hormones are secreted from the brain. Another interesting correlation is the negative correlation between the fat-metabolism and brain-development topics, which makes sense given the lack of fat metabolism happening in the brain.

That being said, the fact that the topics this found were fairly broad led to the consideration of whether CTM would perform as well in biological contexts where the difference in expression between samples was less varied.

7.3 Experiment 3: Neuron subtypes

Thus, another dataset was used, in which RNA-seq data from various subtypes of neurons in the mouse [9] was acquired. While these were different classes of neurons and had different gene expression, the difference in expression between two neurons is much less than between two completely different organs. Thus, it was hypothesized that CTM would not perform as well in this context.

As predicted, the rate of success was a bit lower. However, largely speaking, CTM was still able to capture most of the differences between the reduced samples, and perform around as well, or better than, PCA in the tests with enough topics. The results are in Figure 4.

Features	Classification Accuracy by #Topics/Components			
	5 Topics	15 Topics	25 Topics	35 Topics
CTM	.78	.918	.908	.922
PCA	.91	.902	.916	.914
LDA	.75	.91	.90	.89
Raw	.93			

Figure 4: **Neuron Dataset Results**

This shows very promising results. As expected, having higher numbers of topics appears to improve the classification success, as there is more explanatory power in the model. However, it was found that with

higher numbers of topics, often more iterations were needed to reach convergence. Furthermore, as expected, having more topics substantially slows down the speed of the algorithm.

In terms of the topic-gene distributions themselves, here the results were much more specific. When the gene set analysis was conducted on topics from the 5-topic CTM, no meaningful gene sets could be found, other than ones about broad neuron function, which are not particularly useful or insightful, given that all of the samples were from various neurons. This makes sense, however, since one can see in the data table that the 5-topic CTM performed pretty poorly compared to other methods, suggesting it did not model the data well, and thus we should not expect it to have found meaningful topics. This may mean that 5 topics is simply not enough to model the meaningful variation in the data.

Then, the topics generated from the 15-topic CTM, which performed well on the classification task, were analyzed. Rather than finding broad gene sets for biological categories like "skeletal muscles" as in Experiment 2, the gene sets here seemed to be for very specific cellular functions that make sense for neurons. For example, one topic had genes that positively regulate the post-synaptic membrane, and ones involved in tau-protein kinase activity. Two involved ATP synthesis and energy coupled proton transport, and had a high correlation with one another, as expected. A third was specifically about the translation of genes into proteins, suggesting that neurons highly associated with that factor may be actively expressing genes. Another involved calcium ion transport and synaptic transmission, both of which are related to neurotransmitter release. Moreover, there are dozens of correlations between topics worth investigating. For example, the neurotransmitter release topic is negatively correlated with the ATP synthesis and energy coupled proton transport topic, which may suggest that samples expressing ATP synthesis genes do not release neurotransmitters as much, or perhaps that the limitations of single cell RNA sequencing made it difficult to detect both the synaptic and mitochondrial areas of a neuron. In any case, this approach shows that one can quickly begin to draw biological inferences about the data, and start to explore, perhaps through other methods that are more hypothesis-driven, these various inferences.

There are many important takeaways here, but in general one can see that CTM appears to perform well at modeling these data, and at least in looking at these preliminary results, the topics that it generates appear to have coherence in terms of their biological meanings. That being said, being able to do this analysis is more difficult for samples that are relatively similar, like the neuron data, but still quite good relative to other techniques. Moreover, this approach can be used to generate hypotheses that can be then be tested by other means.

8 Conclusions and Related Work

Ultimately, this work suggests that correlated topic models present an improvement over LDA and PCA as means of performing dimensionality reduction on RNA-seq data. Moreover, the success of classification algorithms suggests that CTM effectively models RNA-seq data, and that the topics can be used to infer biological meaning. One of the advantages of this method is that it can be used for apriori hypothesis-free analysis of the data. While in the majority of current settings, researchers use statistical techniques to answer specific biological questions about gene function, this technique can allow researchers to gain a high-level understanding of the genome, which can then be used to make inferences and generate new hypotheses.

There are several steps where future work can improve these outcomes. Currently, speed is one of the limiting factors in the implementation of this work. Thus, using faster inference algorithms and more-efficient cluster implementations (such as through MPI) can improve performance. Furthermore, the success of this approach in finding interesting and biologically-relevant topics suggests that this could be one of the early steps in an experiment to understand the role of genes in a sample. So, this work has been discussed with a few individuals in various biology groups at Harvard University, and it will soon be applied to try to improve their data analysis. Also, there are related topic models, such as Pachinko Allocation [3], that similarly model the relationship between topics and would perhaps allow for even more expressive power than CTM, which are worth exploring further. These are worth investigating because they may be able to express relationships more strongly than just the pairwise relationships that CTM gives, and thus this may be even more useful for understanding the nature of complex biological phenomena.

References

[1] David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM, 2006.

[4] Huaiyu Mi, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. Large-scale gene function analysis with the panther classification system. *Nature protocols*, 8(8):1551–1566, 2013.

[5] Huaiyu Mi, Anushya Muruganujan, and Paul D Thomas. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, 41(D1):D377–D386, 2013.

[6] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[7] Wikipedia. Latent dirichlet allocation — wikipedia, the free encyclopedia, 2015. [Online; accessed 1-December-2015].

[8] Ying Yu, James C Fuscoe, Chen Zhao, Chao Guo, Meiwen Jia, Tao Qing, Desmond I Bannon, Lee Lancashire, Wenjun Bao, Tingting Du, et al. A rat rna-seq transcriptomic bodymap across 11 organs and 4 developmental stages. *Nature communications*, 5, 2014.

[9] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

Code

Implementations can be found at `https://github.com/ankitvgupta/rnaseqtopicmodeling`. Many of the computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.