1. Introduction

Tumors cell lines are composed of different sub-populations of cells which often exhibit shared patterns of gene expression. Biologists are interested in two key questions: given gene expressions values, (1) can we identify biologically meaningful gene modules, and (2) can we identify cell sub-populations and find gene modules that illuminate their differences? Our goal with this project was to address these questions using Bayesian methods.

To address the first question, we explored the use of Latent Dirichlet Allocation (LDA), two non-parametric topic models, and a dynamic-topic LDA. To address the second question, we explored the use of a finite mixture model as well as a clustering topic model.

1.1. Description of Data

To do our analysis, we used a single-cell RNA-sequencing (scRNA-seq) dataset obtained from human melanoma samples (Tirosh, 2016). The data consisted of the expression values of 22,712 genes for each of 4645 cells. In total, the dataset amounted to 0.86 GB, presenting significant computational challenges when attempting posterior inference. Apart from a few computational tricks (online inference, multi-core parallelization), for most methods, we pruned non-informative genes from the dataset, ranking based on the deviation in value across cells. The Seurat biological toolkit in R was also used for the purpose of low-variance feature selection (Satija, 2012). We eschewed dimensionality reduction techniques such as PCA because of loss of its direct feature interpretability. Prior researchers have labeled the cells in our dataset; there are a total of 9 cell categories (Tirosh, 2016). The complete dataset, as well our preprocessed and pruned versions, are openly available (Yang & Koppula).

1.2. Prior work

Prior research has explored the use of various computational techniques to analyze gene expression data. Most commonly, Spearman and Pearson correlation metrics are frequently used infer sets of genes that cluster together (Xie, 2015; Borensztein, 2017). Other techniques, including PCA followed by linear regression, has been used for expression-based cell clustering (Stegle, 2016). Yu et al. propose an unsupervised classifier ensemble as another approach to cell clustering (Yu, 2016).

More Bayesian approaches have also been tried in prior work from the Pe’er lab. (Prabhakaran, 2016) uses a Hierarchical Dirichlet Mixture Model to learn cell clusterings. (Azizi, 2017) builds on this to jointly learn optimal normalization pre-processing of the data. Bayesian networks have also been used in an attempt to learn gene dependencies from expression data (Pe’er, 2000). Our work uses different models to explore gene expression data, but where appropriate (e.g. LDA vs. non-parametric models), we compare results.

1.3. Structure of Report

We first discuss our experiments using Bayesian topic models to discover topics in scRNA-seq data: LDA in Section 2, non-parametric topic models in Section 3 and Dynamic Time Models in Section 4. Then, we discuss our experiments in clustering: Finite Mixtures in Section 5, and Integrated Topic-Clustering in Section 6. We conclude our paper with our observations from across all our studies.

For the purpose of reproducibility, all code can be found at https://github.com/skoppula/882.
distribution. The identity of the word is drawn from topic’s word distribution. Assuming the reader is familiar with LDA, we relegate further details and formalization of the model to (Blei, 2003).

In the context of analyzing gene expression data, we are interested in discovering 'topics' that comprise of a set of top-$N$ genes within the topic distribution that are biologically related. For example, together the genes may direct a specific chemical function in a cell. Biologists denote such sets of genes as 'gene modules' which can be cross-referenced with existing gene module databases.

2.2. Implementation

A first attempt using the built-in Python $\texttt{lda}$ package resulting in early memory overflows during what we suspect was pre-allocation of per-document variables. The source code was not available, so we had few clues.

We switched to two open-source implementations: an online mean-field variational Bayes for posterior estimation (Nothman, 2017), and a broken C++ Gibbs sampler for LDA (OpenDataGroup, 2015).

We fixed portions of the sampler to compile properly and extended the sampler to run across four cores. Details of our sampling procedure can be found in Appendix 9.2. We compared these two posterior estimation approaches using our entire dataset, using a 10% held-out testing partition. We experimented using $k = 5, 10, 25, 50$ topics. We did not require dataset pruning after these optimizations.

2.3. Experiments

We evaluated the interpretability of our topics (ranked lists of genes) using hypergeometric tests (Wagner, 2015). In brief, this determines the enrichment of our topics for any existing collections of genes catalogued by biologists in the gene module database $\texttt{MSigDB}$ (Broad-Institute, 2015). More details about the score can be found in the Appendix.

Figure 9 in the Appendix shows gene module matches in $\texttt{MSigDB}$ for which the $p$-value of the match is at least less than 0.3. Notice that models with more topics tended to have more matches with higher significance. The $p$-values in our tests do not factor for multiple hypothesis corrections, so at the moment we are only using them as relative measures of model quality.

It is validating to see that many of the similar modules, such as the $\texttt{BRCA1}$ and $\texttt{DOXORUBICIN}$, are cancer related, given that our expression assay was from tumor cells.

Extracting the parameters of each model, we calculated log-likelihood on our held-out test set. This is shown in Figure 1. We found that for this dataset increasing the topic count increases log-likelihood, roughly linear from 5 to 50. Further experiments will need to test at what point log-likelihood drops off; this is one way to optimize the number of topics.

As an interesting aside, Figure 2 shows the time until inference completion (collected during our experiments). Gibbs appears to scale poorly with the parameter dimensionality, in contrast to online variational Bayes $^2$.

Posterior predictive checks to test the mutual information between the each cell and its words' topics is something that we did not have time, but would be interesting to examine. It’s not clear that this inde-

\footnote{It is important to note that comparing the magnitude of the time to complete inference in Figure 2 is not particularly meaningful; as described in Appendix 9.2, we chose a reasonable guess for the number of iterations and optimization threshold after which to stop estimation.}
pendence assumption is true in the context of gene expression data, because of cross-talk and regulatory mechanisms between gene modules.

3. Non-Parametric Topic Models

3.1. Model Description

One limitation of LDA is the need to define the number of topics a priori. Non-parametric models overcome this limitation by assuming a countably infinite number of topics. We trained and compared two non-parametric topic models with LDA: the Hierarchical Dirichlet Process (HDP) and the Indian Buffet Process Compound Dirichlet Process (IDP-DP). On a high level, HDP is very similar to LDA, with the main difference being that the draws from a Dirichlet distribution in LDA are replaced with draws from a Dirichlet process in HDP (Blei, 2005). One drawback to the HDP is that there tends to be a correlation between how frequently a topic appears across all documents and how prevalent this topic is within documents that it appears in. This may be undesirable in circumstances where there are ‘rare’ topics with high prevalence in a small number of documents. IBP-DP tries to overcome this problem by separately modeling the frequency of documents that contain a topic and the prevalence of this topic within a document where it is present (Blei, 2005). Both generative models are described in greater detail in the Appendix (sections 9.3 and 9.4).

3.2. Experiment: LDA vs. HDP

We trained an HDP topic model and compared it to LDA models with similar numbers of topics, using held-out perplexity and gene set enrichment analysis as described above as our evaluation metrics. We used an existing C++ implementation of the Gibbs sampler provided by the authors of the original paper (Blei, 2005). Since the available implementation of the posterior inference algorithm for HDP was too slow to run on the entire dataset, we did feature selection using the Seurat toolkit in R (Satija, 2012) to reduce the number of genes in our dataset with low variance across cells, and trained all models on this subset of data. We tuned the concentration parameters in HDP to obtain a reasonable number of topics (i.e., 150); subsequently, we trained LDA models with similar numbers of topics for comparison. We found that the HDP model had higher perplexity on a held-out dataset (Figure 11). Moreover, the topics in the HDP model tended to have less significant enrichment for known gene sets from MSigDB, and we found that this difference to be significant (Figure 12, Mann-Whitney-U test, \( p < 0.0001 \)). Overall, the LDA model proved superior to the HDP model in this experiment.

3.3. Experiment: LDA vs. IBP-DP

One drawback to the HDP is that there tends to be a correlation between how frequently a topic appears across all documents and how prevalent this topic is within documents that it appears in. Williamson et al. (Blei, 2010) proposed the ‘focused topic model’ to overcome this drawback. We implemented their model to compare it with LDA on the Reuters-21578 dataset. Since the code from the original IBP-DP paper was not available, we implemented an inference algorithm using collapsed Gibbs sampling (Blei, 2010). Due to the non-conjugacy of the model, sampling each latent variable from its full conditional required using another sampling method. To sample the topics parameters, the number of which changes depending on how many topics are represented in the dataset, we used slice sampling based on the semi-ordered stick-breaking representation of the model (Teh, 2007). For more details about implementation, please refer to those papers.

We tested our code on a subset of the Reuters-21578 dataset, using several different values of the concentration hyper-parameter \( \alpha \), which influences the number of clusters. Although higher values of \( \alpha \) yielded better log-likelihood values, we found that it resulted in a large number of very small topics, which are not very useful (Figure 13). Qualitatively, we did not find the topics from IBP-DP (Figure 14) to be more coherent than topics than LDA (Figure 15). The most prevalent topics from the IBP-DP each corresponded to similar topics from LDA; less prevalent topics tended to consist of a few unrelated words. These results discouraged us from optimizing the code to train this model on our scRNA-seq dataset, as we do not think it would yield more coherent gene modules than LDA. We emphasize that these results are not completely unexpected, as the authors of the IBP-DP paper did not show any topics from their model, nor did they assess the quality of their model with metrics other than perplexity.

4. Dynamic Topic Model

4.1. Model Description

As tumor cells proliferate, they undergo a process called differentiation. The set of cells types during tumor emergence may differ from the set of types expressed days later. Correspondingly, patterns of gene
expression may also change over time. The set of gene modules expressed in the cells, or the composition of each gene module (topic), may also change.

To capture these evolving topics, we look to dynamic topic models (Blei & Lafferty, 2006). In brief, dynamic topic models establish a conditional distribution over the hyperparameters $\alpha$ and $\beta$, that govern the document’s topic distribution and topic’s word distribution, respectively. The distribution over each hyperparameter is conditioned on the prior in the previous time step, allowing changes to topic composition and assignment. This can be seen graphically in the plate model in Figure 3. For convenience, we provide a concise description of the generative process for the implemented hierarchical model for a time slice $t$ in Appendix 9.5.

![Figure 3](image)

**Figure 3.** Plate diagram for the Dynamic Topic Model. Re-produced from (Blei & Lafferty, 2006).

Note that in contrast to LDA, dynamic topic models use a logistic-normal to express proportion uncertainty. This is stated more explicitly in Appendix 9.5.

### 4.2. Implementation

Unfortunately, due to the non-conjugacy of the Gaussian/Multinomial in our logistic-normal setup, integrating out parameters for any sort of Gibbs sampling becomes hard. Instead, the original DTM paper uses a variational approximation based on a Kalman filter, that preserves time dependencies, unlike a mean-field approximation.

Re-implementing the variational approximation, while educational and interesting, would be a project of itself, so our first attempt in employing dynamic time models to our gene expression data re-used the Kalman Filter-based variational inference used by the original authors, published recently by the Blei Lab (Blei & Gerrish, 2015). The results we show in the subsequent section are using a wrapper we’ve implemented around the group’s inference code.

The authors stumbled upon a recent arXiv pre-print that proposed a set of sampler update rules to create a correct DTM Gibbs sampler (Bhadury, 2016). Using the probabilistic programming framework Edward, and modifying the in-built sampler to follow these updates, we were able to obtain a working Gibbs-sampler for sample time-sliced data for a dynamic topic model defined in Edward (Tran et al., 2016). While we don’t have time to conduct benchmarks or re-run our results right now, after verifying that results are consistent, the authors hope to submit a merge for this Edward sketch into the mainstream Edward examples library this coming summer.

### 4.3. Experiments

We subdivided our data into three separate time slices, and pruned each slice, retaining the top 20% of high-variance genes. We experimented using $k = 15$, 30, and 50 topics.

Similar to our analysis in Section 2, we used the hypergeometric test to evaluate interpretability.

Figure 10 in the Appendix shows gene module matches for which the $p$-value is less than 0.3, across the three time slices and the three topic count sizes. Interestingly, we notice a very different set of matched gene modules, with exception of the KLF1 module$^3$. Like before, we notice that some of the identified topics are relevant to tumor cells: ST_TUMOR_NECROSIS_FACTOR_PATHWAY and WEST_ADRENOCORTICAL_TUMOR_MARKERS_UP to name the two most prominent topics.

We do not see that much variation across time-slices, demonstrating that our time-slice partitions are largely uniform. For example, for model $k = 50$, we see that topic 20 consistently matches with MSigDB module PILON_KLF1_TARGETS_UP, suggesting that its gene composition is not varying.

Models with more topics tended to have more matches with higher significance. But again, the $p$-values in our tests do not factor for multiple hypothesis corrections, so at the moment we are only using them as relative measures of model quality.

After extracting the mode of learned parameters, we evaluated the log-likelihood of our held-out test set, also partitioned by time. This is shown in Figure 4. We find a leveling of log-likelihood after $k = 30$.

With more time, we would ideally repeat the experiments to obtain confidence intervals on these experiments, and examine the gene-module/cell/time-slice

---

$^3$This could very well be a result of our dataset pruning, which we did to keep inference runtimes manageable.
5. Mixture Model

5.1. Model Description

The second question – whether we could learn cell groupings – we tackled using a canonical Gaussian mixture model. In brief, the model assigns every data point to a cluster; every cluster is Gaussian whose parameters are learned through posterior inference. A plate diagram can be found at (Ben Wing, 2016). This model is different from the previously mentioned Pe’er papers which use Dirichlet Processes Mixtures to study expression data. By plotting histograms of whitened gene expression values, we’ve found that for many genes the distribution is Gaussian across our data points. This hints that the generative process in a Gaussian mixture model is plausible.

5.2. Implementation

We implemented the mixture model in Edward, the probabilistic programming framework on top of Tensorflow. This allowed us to experiment with different samplers and variational inference approximations. The results we show below are for our runs using a Gibbs sampler, with a fixed burn-in of number of samples (200), a fixed number of sampling iterations (500).

5.3. Experiments

Figure 6 shows an estimated cluster assignment for one-hundred sampled data point. We see that while our learned assignments do capture some of the clustering, there is significant bias toward one group (the burgundy cluster).

Following this trend, we examined the set of unique clusters assigned in the posterior. The results are shown in Table 5.3. Interestingly, the model consistently uses less clusters than is allocated in the model. We are unsure why this is the case.

Finally, we examine the log-likelihood of a held-out test set three different runs of posterior estimation with our sampler. This is shown in Figure 5. The runs are consistent with each other, and the likelihood is largely invariant across the five different cluster counts we tried, \( k = \{10, 25, 50, 75, 100\} \). There are an estimated 9 true cell types in the original dataset, which could explain this result.

### Table 1

<table>
<thead>
<tr>
<th># of Clusters</th>
<th># of Clusters Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>50</td>
<td>16</td>
</tr>
<tr>
<td>75</td>
<td>18</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
</tr>
</tbody>
</table>

6. Integrated Topic-Clustering Model

6.1. Model Description

LDA is suitable for modeling corporuses with no special structure, but this is not the case of scRNA-seq data. Since the cells within a tissue sample may belong to many different sub-populations, we would expect cells within the same sub-population to have more similar topic proportions. We capture this additional structure using the clustering topic model (CTM), a unified framework for clustering and topic modeling based on LDA. In this model, which is similar to the one proposed by (Wallach, 2008), the following generative process is assumed:
1. Choose a distribution over the clusters, $\pi \sim \text{Dir}(\pi_0/C)$, where $\pi_0$ is a hyperparameter and $C$ is the number of clusters.

2. Choose a global distribution over the topics, $\phi_0 \sim \text{Dir}(\alpha_0/T)$, where $\alpha_0$ is a hyperparameter and $T$ is the number of topics.

3. For each cluster $c$, choose a cluster distribution over the topics, $\phi_c \sim \text{Dir}(\alpha_1 \phi_0)$, where $\alpha_1$ is a hyperparameter.

4. For each topic $k$, choose a distribution over the vocabulary, $\beta_k \sim \text{Dir}(\eta)$, where $\eta$ is a hyperparameter.

5. For each document $d$,
   
   (a) Choose a cluster assignment $\xi_d \sim \text{Cat}(\pi)$
   
   (b) Choose a distribution over topics based on the cluster assignment, $\theta_d \sim \text{Dir}(\alpha \phi_{\xi_d})$, where $\alpha$ is a hyperparameter.
   
   (c) Choose the number of words in this document, $N_d \sim \text{Poisson}(\kappa)$, where $\kappa$ is a hyperparameter.
   
   (d) For the $i$th word in the $d$th document,
      
      i. Choose a topic, $z_{di} \sim \text{Categorical}(\theta_d)$, based on the distribution over topics
      
      ii. Choose a word, $w_{di} \sim \text{Categorical}(\beta_{z_{di}})$, based on the distribution over the vocabulary for topic $z_{di}$

This model is very similar to LDA, with the addition of cluster assignments that influence document topic proportions. Note that the documents in each cluster have more similar topic proportions, particularly if we choose a reasonably large value of $\alpha$.

### 6.2. Inference

We implemented a collapsed Gibbs sampling in MATLAB to perform inference for the CTM, integrating out all latent variables except the topic assignments $z$ and the cluster assignments $\xi$ to facilitate mixing. First, we establish the prior distribution of $z_{di}$:

$$p(z_{di} = k | z_{-di}, \xi_d) = \frac{N_{d,k} + \alpha}{N_d + \alpha}$$

where $z_{-di}$ refers to topic assignments other than that for word $i$ in document $d$, $N_{d,k}$ is the number of words in document $d$ assigned to topic $k$, $N_d$ is the total number of words in document $d$, $N_{\xi_d,k}$ is the number of words in the documents from cluster $\xi_d$ assigned to topic $k$, $N_{\xi_d}$ is the total number of words in the documents from cluster $\xi_d$, $N_k$ is the number of words in the corpus assigned to topic $k$, and $N$ is the total number of words in the corpus.

It follows that the full conditional for each $z_{di}$ is:

$$p(z_{di} = k | z_{-di}, \xi_d) = \frac{N_{w_{di},k} + \alpha N_{\xi_d,k} + \alpha}{N_{w_{di}} + \alpha N_{\xi_d} + \alpha}$$

where $N_{w_{di},k}$ is the number of times the word $w_{di}$ appears in topic $k$, $N_k$ is the total number of words assigned to topic $k$, and the prior of $z_{di}$ is given by equation 1. Finally, the full conditional for each $\xi_d$ is:

$$p(\xi_d = c | \xi_{-d}, z_{-d}; \pi) \propto p(z_d | \xi_d = c, z_{-d}) p(\xi_d = c | \xi_{-d})$$

$$= \frac{M_{c,e} + \pi_0/C}{M_d + \pi_0}$$

4When counting the number of words, we always exclude the word currently being considered, but we omit this in the equation to avoid clutter.

Figure 6. 2D-projection of sampled gene expression data points, colored by predicted cluster (left) and by true cluster assignments (right). Axis are the two learned PCA dimensions.
where $M_c$ is the number of documents assigned to topic $c$ and $M$ is the total number of documents in the corpus\(^5\). $\text{Multi}(z_d, p(z_d | \cdots))$ is the probability under the multinomial distribution of getting topic assignments $z_d$ when the topic proportions are given by $p(z_d | \cdots)$. The topic proportions are obtained essentially as described in equation 1, except we set the number of words in document $d$ to be 0 and adjust all the other counts accordingly.

6.3. Experiments

We trained the CTM on a subset of 500 cells and 3000 genes from the human melanoma scRNA-seq dataset using the collapsed Gibbs sampler until convergence (about 500 iterations). All hyperparameters were set to 1 with the exception of $\alpha$, which we set to 10 to encourage similarity of topic proportions for cells assigned to the same cluster, and the numbers of clusters and topics were set to 10 and 20 respectively. To determine the efficacy of the clustering, we generated a tSNE plot of the cells using the Euclidean distance between their topic proportions (Figure 7). Compared to a tSNE plot based on the Euclidean distance between the counts of all 3000 genes (Figure 16), in which there was no obvious grouping of cells, we found that the topic proportions were nicely grouped based on their learned cluster assignments. To determine if our clusters had any biological significance, we compared them with cell labels published by the original authors (Figure 8)\(^6\). Interestingly, our clusters corresponded nearly exactly to the different cell types, and we were even able to classify the cells with unknown labels.

Subsequently, we wanted to determine if the learned topics were meaningful. We first used the biological relevance score described earlier and found that the topics were significantly enriched for some known gene sets from MSigDB (Figure 17). We then did some examination of the specific genes appearing in select topics using a combination of the MSigDB analysis and GO-enrichment analysis\(^7\). We found that topic 18, which

---

\(^5\)When counting the number of documents, we exclude the document currently being considered, but omit this in the equation to avoid clutter.

\(^6\)The authors published the cell labels based on prior biological knowledge, i.e. thresholding the cells based on known gene markers of different cell types. Due to noise and dropout in scRNA-seq data, they were not able to label all cells using this approach.

\(^7\)GO enrichment analysis looks for enrichment of genes associated with cell processes. Whereas MSigDB is a database of various published gene sets, the GO database contains only those well-characterized gene modules that

---

Figure 7. tSNE plot of cells from human melanoma, plotted using Euclidean distances between topic proportions, with cluster assignments shown in color. The cells nicely group into their cluster assignments, suggesting that the clustering topic model inference algorithm worked as intended. Moreover, the cluster assignments correspond to the cell labels found by the group that published the dataset, see Figure 8.

Figure 8. tSNE plot of cells from human melanoma, plotted using Euclidean distances between topic proportions, with labels published by the group that collected the dataset. 0 - label unknown; 1 - T-cell; 2 - B-cell; 3 - Macrophage; 4 - Endoderm; 5 - Cancer-associated fibroblast; 6 - Natural Killer; 7 - Malignant. The labels correspond to the cluster assignments learned using the CTM model.
was prevalent in the T-cells and not in other cells (Figure 19), was highly enriched for genes involved in "myeloid dendritic cell activation involved in immune response". Topic 3, which was prevalent in the B-cells and not elsewhere (Figure 20), was highly enriched for genes involved in "lymphocyte activation" and "inflammatory response". Topic 4, which was highly prevalent in the T-cells and B-cells only (Figures 19 and 20), was highly enriched for genes involved in "somatic diversification of immune receptors". Finally, topic 8 which was mainly present in the tumor cells (18), was enriched for genes involved in "G1 to S transition of mitotic cell cycle" and "cell division".

The correspondence between the genes enriched in the topics and the cell sub-populations with those topics suggests that our learned topics have biological significance. As a result, the highly-ranked genes in some of these topics may be of interest to experimental biologists as candidate gene markers for specific cell sub-populations.

7. Discussion

In this paper, we sought to address the following two questions using Bayesian topic models: given gene expression data, (1) can we identify biologically meaningful gene modules, and (2) can we identify cell sub-populations and find gene modules that illuminate their differences.

7.1. Can we identify biologically meaningful gene modules?

We addressed the first question by trying to model scRNA-seq data with LDA as well as two non-parametric topic models, the HDP and the IBP-DP. The purpose of doing this was mainly to determine if topic models based on LDA are suitable for modeling gene expression; as far as we are aware, there are no publications applying LDA to scRNA-seq data. Our results from section 2 suggest that LDA could potentially be useful for scRNA-seq, since our topics were enriched for previously discovered biological gene sets. One major consideration is computational cost: although scRNA-seq data is sparse, it is not nearly as sparse as the bag-of-words representation of a corpus, so most inference algorithms take a long time to run on the entire dataset. Based on our results, however, it appears that this problem can be addressed either by using a faster posterior inference method such as online variational inference or by pre-processing the dataset for those genes with the greatest variability.

Compared to LDA, we found that HDP and IBP-DP were not as suitable for modeling scRNA-seq data (Section 3). The HDP underperformed in comparison to LDA in both log-likelihood on a held-out dataset and the biological significance scores of its topics. The IBP-DP did not seem to work better than LDA on a subset of the Reuters-21578 corpus, based on a qualitative analysis of the topics, so we did not try it on the gene expression data, as we do not expect it to perform better than LDA. The underperformance of the IBP-DP compared to LDA on the Reuters-21578 dataset does not surprise us, as the authors of the IBP-DP paper did not show any topics from their model, even though they claimed it may produce more 'focused' topics.

After we were confident that LDA could be useful for modeling scRNA-seq data, we wanted to try models that could capture higher-level structure of gene expression data - for instance, time. As cells proliferate, they may differentiate into different cell types. A useful topic model might learn not just the topics at a given time point, but also how the topics are evolving over time. We have presented some preliminary results applying a dynamic topic model to our scRNA-seq dataset and selecting an appropriate number of topics (Section 4). To further test this model, experiments could be done on simulated time-series scRNA-seq data with known evolving topics, and one could evaluate how well those evolving topics are picked up by the model.

7.2. Can we identify cell sub-populations and find gene modules that illuminate their differences?

In addition to time, another source of structure in the scRNA-seq dataset is the clustering of cells into sub-populations. In biology, there is much interest in identifying these sub-populations and finding gene modules or markers that are characteristic of different sub-populations. We first wanted to determine if a Bayesian finite mixture model would be appropriate for separating the cells into clusters (Section 5). We found that we were approximately able to learn the cell type labels published by (Tirosi, 2016)

Finally, we wanted to create a unified framework for cell clustering and topic modeling. We implemented the clustering topic model (Section 6) and found that it was able to correctly separate the cells into sub-populations. It appeared to outperform the vanilla mixture model without clustering, suggesting that
knowledge of the topics helps the model learn better clusters. We also found that several topics corresponded to the cell types that they were prevalent in; we believe these results suggest that biologists can use the topics - which are ranked lists of genes - as candidate gene markers for cell sub-populations of interest within the melanoma, such as malignant cells, T-cells and B-cells. Whether these candidate gene markers are truly useful must be confirmed by biologists experimentally. Looking forward, it would be very interesting to combine the dynamic topic model and the clustering topic model, so that we can capture the time-evolution of the cell sub-populations within a scRNA-seq dataset.

8. Author Contributions
The implementations and experiments involving LDA, Gaussian Mixtures, and the Dynamic Topic Model was completed by SK (Section 2, Section 4, and Section 5). KY completed the implementations and experiments involved non-parametric topic models, the clustering topic model, and the gene set enrichment test implementation (Section 3, Section 6, and Section 9.1). The authors contributed equally in this work.

References


9. Appendix

9.1. Biological Significance Score / Gene Set Enrichment

In general, it is difficult to evaluate the ‘coherence’ of topics of genes, since unlike topics of words, we cannot just read them and get a sense of how similar they are. Therefore, nothing like the intrusion test exists for evaluating topics of genes. To overcome this, we developed our own test for the biological significance of gene topics. The MSigDB contains thousands of gene sets discovered through previous research, catalogued by biologists from the Broad Institute. We deduced that if a topic is ‘coherent’, then it is likely significantly enriched for some of these gene sets. For each topic, our biological significance scoring method iterates through the MSigDB database and tests each gene set for enrichment using the minimum hyper-geometric test (Wagner, 2015). The most significant p-values for each topic are recorded and used as a proxy for its biological coherence.

9.2. Latent Dirichlet Allocation

For our Gibb’s sampler, we had a fixed burn-in of number of samples (200), a fixed number of sampling iterations after that (500). We didn’t extensively explore varying these values, but trying out significantly more iterations (700) didn’t seem to change the topic’s word distributions significantly. There was one sampling chain on each of four cores.

9.3. HDP

HDP is an infinite topic extension of LDA based on the Dirichlet process (Blei, 2005). Briefly, words in a corpus are assumed to be generated as follows:

1. Sample the global distribution over topics, $G_0 \sim \text{DP} (\gamma, H)$, from a Dirichlet process with concentration $\gamma$ and base Dirichlet distribution $H$.

2. For each document $d$,
   - (a) Sample the local distribution over topics, $G_d \sim \text{DP} (\alpha_0, G_0)$, from a Dirichlet process with concentration $\alpha_0$ and base distribution $G_0$.
   - (b) Choose the number of words in this document, $N_d \sim \text{Poisson}(\xi)$, where $\xi$ is a hyper-parameter.
   - (c) For the $i$th word in the $d$th document,
     - i. Choose a topic (i.e. distribution over words in vocab), $\beta_{di} \sim G_d$, based on the local distribution over topics
     - ii. Choose a word, $w_{di} \sim \text{Categorical}(\beta_{di})$, based on the distribution over the vocabulary in the topic

9.4. IBP-DP

One drawback to the HDP is that there tends to be a correlation between how frequently a topic appears across all documents and how prevalent this topic is within documents that it appears in. Williamson et al. (Blei, 2010) proposed the ‘focused topic model’ to overcome this drawback. In their model, the frequency of a topic $k$ depends on two separate variables, its relative prevalence within a document ($\phi_k$) and the probability that a given document contains this topic ($\pi_k$), thus reducing correlation between the two. Briefly, the words in the corpus are assumed to be generated as follows:

---

8More specifically, $H$ is a Dirichlet distribution over the $(V-1)$-dimensional simplex, where $V$ is the size of the vocabulary, and $G_0$ is a countably infinite set of point masses over this simplex whose weights sum to 1. $G_0$ in HDP plays a role analogous to $\alpha$ in LDA.
1. For each topic $k = 1, 2, \ldots$
   (a) Choose a population frequency (i.e. probability that a document contains topic) $\pi_k = \prod_{j=1}^{k} \mu_k$, where each $\mu_k \sim \text{Beta}(\alpha, 1)$, where $\alpha$ is a hyperparameter.
   (b) Choose a relative prevalence (i.e. how often the topic appears within a document containing the topic) $\phi_k \sim \text{Gamma}(\gamma, 1)$, where $\gamma$ is a hyperparameter.
   (c) Choose a distribution over the vocabulary, $\beta_k \sim \text{Dir}(\eta)$, where $\eta$ is a hyperparameter.

2. For each document $d$,
   (a) Choose the topics that will appear as a binary vector, $b_d \sim \text{Bernoulli}(\pi)$
   (b) Choose the topic proportion as $\theta_d \sim \text{Dirichlet}(b_d \cdot \phi)$
   (c) Choose the number of words in this document, $N_d \sim \text{NegativeBin}\left(\sum_k b_{dk} \phi_k, \delta\right)$, where $\delta$ is a hyperparameter.
   (d) For the $i$th word in the $d$th document,
      i. Choose a topic, $z_{di} \sim \text{Categorical}(\theta_d)$, based on the distribution over topics
      ii. Choose a word, $w_{di} \sim \text{Categorical}(\beta_{z_{di}})$, based on the distribution over the vocabulary for topic $z_{di}$

9.5. Dynamic Topic Model

Here is a description of the generative process for our dynamic topic model:

1. Draw a new topic composition hyperparameter
   $\beta \leftarrow \mathcal{N}(\beta_{t-1}, \sigma I^2)$

2. Draw a new document composition hyperparameter
   $\alpha \leftarrow \mathcal{N}(\alpha_{t-1}, \delta I^2)$

3. For each document:
   (a) Draw a new document topic distribution
      $\theta \leftarrow \pi(\mathcal{N}(\alpha_t))$
   (b) For every word in the document:
      i. Draw the word topic assignment
         $Z \leftarrow \text{Mult}(\theta)$
      ii. Draw the word identity
         $Z \leftarrow \text{Mult}(\pi(\beta_{z_i}))$

Here, $\pi(x_i)$ is the softmax function $\frac{\exp(x_i)}{\sum_k \exp(x_k)}$ (Blei & Lafferty, 2006).
Clustering and Topic Discovery in Gene Expression Data

**GIBBS TOPICS**

<table>
<thead>
<tr>
<th>cluster_size: 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>match: topic 1</td>
</tr>
<tr>
<td>match: topic 1</td>
</tr>
<tr>
<td>match: topic 6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cluster_size: 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>match: topic 8</td>
</tr>
<tr>
<td>match: topic 22</td>
</tr>
<tr>
<td>match: topic 22</td>
</tr>
<tr>
<td>match: topic 22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cluster_size: 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>match: topic 2</td>
</tr>
<tr>
<td>match: topic 3</td>
</tr>
<tr>
<td>match: topic 6</td>
</tr>
<tr>
<td>match: topic 7</td>
</tr>
<tr>
<td>match: topic 17</td>
</tr>
<tr>
<td>match: topic 17</td>
</tr>
<tr>
<td>match: topic 22</td>
</tr>
<tr>
<td>match: topic 24</td>
</tr>
</tbody>
</table>

**OVB TOPICS**

<table>
<thead>
<tr>
<th>cluster_size: 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>match: topic 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cluster_size: 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>match: topic 5</td>
</tr>
<tr>
<td>match: topic 8</td>
</tr>
<tr>
<td>match: topic 15</td>
</tr>
<tr>
<td>match: topic 15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cluster_size: 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>match: topic 6</td>
</tr>
<tr>
<td>match: topic 15</td>
</tr>
<tr>
<td>match: topic 22</td>
</tr>
<tr>
<td>match: topic 28</td>
</tr>
<tr>
<td>match: topic 38</td>
</tr>
<tr>
<td>match: topic 41</td>
</tr>
<tr>
<td>match: topic 42</td>
</tr>
</tbody>
</table>

*Figure 9.* Matches between the gene collections found in LDA topics and published gene sets in MSigDB. ‘Cluster size’ refers to the number of topics in the model.
Clustering and Topic Discovery in Gene Expression Data

Figure 10. Matches between the gene collections found in DTM topics and published gene sets in MSigDB. $k$ refers to the number of topics in the model, $t$ refers to one of three time slices.

| $k=15$ | $t=1$ | match: topic 10, PILON_KLF1_TARGETS_UP, pval: 0.65484091166 |
| $t=2$ | match: topic 11, ST_TUMOR_NECROSIS_FACTOR_PATHWAY, pval: 0.1315961800 |
| $t=3$ | match: topic 2, ST_INTEGRIN_SIGNALING_PATHWAY, pval: 0.22324501174 |

| $k=30$ | $t=1$ | match: topic 26, WEST_ADRENOCORTICAL_TUMOR_MARKERS_UP, pval: 0.04716140993 |
| $t=2$ | match: topic 12, KEGG_PARKINSONS_DISEASE, pval: 0.19929937855 |
| $t=3$ | match: topic 26, WEST_ADRENOCORTICAL_TUMOR_MARKERS_UP, pval: 0.08559659069 |

| $k=50$ | $t=1$ | match: topic 13, WEST_ADRENOCORTICAL_TUMOR_MARKERS_UP, pval: 0.0092204210 |
| $t=2$ | match: topic 20, HOLLMANN_APOPTOSIS_VIA_CD40_DN, pval: 0.05797811165 |
| $t=3$ | match: topic 7, ST_TUMOR_NECROSIS_FACTOR_PATHWAY, pval: 0.20037158939 |

Figure 11. Comparison of log-likelihood of the held-out testing set, under various LDA models and the HDP model.
Figure 12. Comparison of distributions of p-values from gene set enrichment analysis between LDA models and the HDP model.

Figure 13. Beeswarm plot of number of words per topic, for 3 different IBP-DP models with different concentration parameters. Each point represents one topic from its model.
Clustering and Topic Discovery in Gene Expression Data

Figure 14. Topics from IBP-DP model trained on subset of Reuters dataset.

Figure 15. Topics from LDA model with 15 topics trained on subset of Reuters dataset.
Figure 16. tSNE plot of cells from human melanoma, plotted using Euclidean distances between gene count vectors, with cluster assignments shown in color.

Figure 17. Histogram of topics from CTM model based on p-value of enrichment of MSigDB gene sets.
Figure 18. Bar charts of topic proportions for malignant cell clusters.

Figure 19. Bar charts of topic proportions for T-cell clusters.
Figure 20. Bar charts of topic proportions for B-cell clusters