Bayesian Clustering and Topic Discovery: Adventures with Gene Expression Data

Karren Dai Yang, Skanda Koppula {KARREN, SKOPPULA}@MIT.EDU
Massachusetts Institute of Technology, Cambridge, MA 02139 USA

1. Introduction

Tumors cell lines are composed of different subpopulations 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 cell clusterings, and (2) can we identify clusters of biologically-related genes? Our goal with this project was to answer both these two questions using Bayesian methods.

In brief, we explored the use of three Bayesian clustering methods (a vanilla mixture, integrated topic-mixture, and non-parametric models) to address the first question, and two topic models (vanilla and a dynamic-topic LDA) to address the second.

1.1. Description of Data

Using a contemporary gene sequencing machine, we obtained samples of single-cell RNA-sequencing data (scRNA-seq) taken from tumors in mice. Our 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 estimation. 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 [1]. We eschewed dimensionality reduction techniques such as PCA because of loss of its direct feature interpretability. The complete dataset, as well our preprocessed and pruned versions, are available at (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; Borenszstein, 2017). Other techniques, including PCA followed by linear regression, has been used for expression-based cell clustering (Stegle, 2016). Yu et al. propose a 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 Heirarchical 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).

1.3. Structure of Report

We first discuss our experiments using Bayesian topic models to discover related sets of genes in a 'topic': LDA in Section 2 and Dynamic Time Models in Section 4. We discuss our experiments in clustering: Dirichlet Mixtures in Section 3, Integrated Topic-Clustering in Section 5, and non-parametric models in Section 6. Finally, we conclude our paper with our obversations from across all studies.

2. Latent Dirichlet Allocation

2.1. Model Description

In the generative process for LDA, the topic assigned to a word in a document is a drawn from the document's topic 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 genes that are biologically related. For exam-

[1] We recognize that this can bias towards noisy genes. We favor this method because it is simple and easy to implement, and a practice used in literature (Ling, 2012)

Algorithm 1 Mixture Model

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

ple, 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

2.3. Experiments

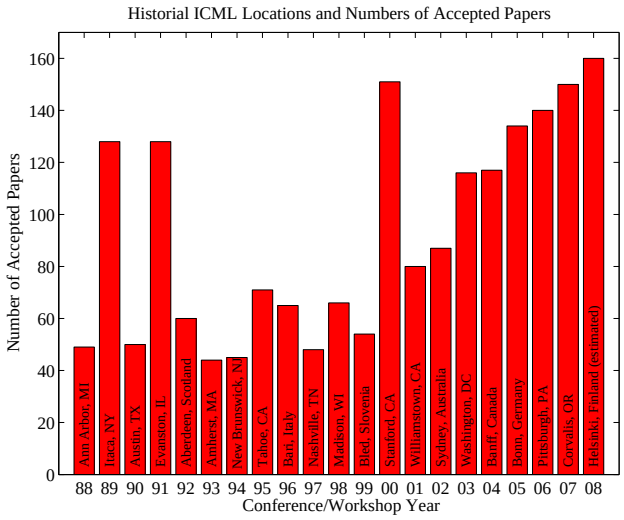

Figure 1. This is a demo figure.

3. Dirichlet Mixture Model

3.1. Model Description

Algorithm 1 describes the generative process for the mixture odel.

3.2. Implementation

3.3. Experiments

4. Dynamic Time Model

4.1. Model Description

4.2. Implementation

4.3. Experiments

5. Integrated Topic-Clustering Model

5.1. Model Description

5.2. Implementation

5.3. Experiments

6. Non-parametric Models: IBP and HDP

6.1. Model Description

6.2. Implementation

6.3. Experiments

References

Azizi, et al. Bayesian inference for single-cell clustering and imputing. https://genomicscomputbiol.org/ojs/index.php/GCB/article/view/46/35, Genomics 2017.

Blei, et al. Latent dirichlet allocation. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf, JMLR 2003.

Borenszstein, et al. Xist-dependent imprinted x inactivation and the early developmental consequences of its failure. http://www.nature.com/nsmb/journal/v24/n3/fig_tab/nsmb.3365_SF1.html, Nature Structural & Molecular Biology 2017.

Ling, et al. Improving relative-entropy pruning using statistical significance. https://www.cs.cmu.edu/~awb/papers/coling2012/rep_coling2012.pdf, ACL 2012.

Pe'er, et al. Using bayesian networks to analyze expression data. http://www.cs.huji.ac.il/~nir/Papers/FLNP1Full.pdf, Genomics 2000.

Prabhakaran, et al. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. http://www.c2b2.columbia.edu/danapeerlab/html/pub/prabhakaran16.pdf, ICML 2016.

Stegle, et al. Computational and analytical challenges in single-cell transcriptomics. https://www.nature.com/nrg/journal/v16/n3/full/nrg3833.html, Nature Methods 2016.

Xie, et al. SINCERA: A pipeline for single-cell rna-seq profiling analysis. http://months.plos.org/ploscompbiol/article?id=10.1371/month.pcbi.1004575, PLoS 2015.

Yang, Karren and Koppula, Skanda. Tumor cell melanoma data. https://www.dropbox.com/sh/tdrplhiu3mzvmeu/AADFlZITyPX_-I4O7jDuruJra?dl=0.

Yu, et al. SC3 - consensus clustering of single-cell rna-seq data. http://biorxiv.org/content/early/2016/09/02/036558, Nature Methods 2016.

7. Appendix

7.1. Latent Dirichlet Allocation