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 are composed of different sub-populations of cells, and these sub-populations often exhibit shared patterns of gene expression. With contemporary sequencing machines, it is possible to obtain the expression levels of 10,000+ genes for 1000+ cells in a single experiment.

1.1. Prior work

Prior research has applied basic measures of statistical distance to cluster cell groups or gene modules based on samples of single-cell RNA-sequencing (scRNA-seq) data taken from tumors (Borenszstein, 2017). Good clustering of scRNA-seq data often has biologically meaningful results, and as such, a wide variety of correlation based methods have been used in prior work to derive meaning from scRNA-seq data (Crow, 2016; Yu, 2016; Xie, 2015). We believe that Bayesian methods for clustering and topic modeling can be more illuminating than these previous methods, due to their consideration of higher-order relationships within the data.

1.2. Description of Data

1.3. Structure of Report

In this project, we applied Bayesian methods for clustering and topic discovery to discover meaningful cell clusters and gene modules from single-cell RNA-sequencing (scRNA-seq) datasets from tumors. We first intend to explore the use of a hierarchical topic model, Latent Dirichlet Allocation (LDA) (Blei, 2003), as well as two non-parametric topic models, Hierarchical Dirichlet Processes (HDP) (Blei, 2005) and the Indian Buffet Process Compound Dirichlet Process (IBP-CDP) (Blei, 2010) for analyzing scRNA-seq data. We will evaluate these models based on their ability to discover meaningful functional gene modules. Subsequently, we plan to explore the use of mixture models to cluster cells based on the output of the topic models, which we will evaluate based on their ability to differentiate cell types (e.g. immune, cancer, non-cancerous). Finally, we will train a combined clustering-topic model (e.g. MGCTM (Xie, 2013)) to see if it outperforms the individual models. We will compare the results of these methods with standard methods used in prior work, and evaluate these methods' robustness to noise. Time permitting, we will study extensions of these models appropriate for time-series scRNA-seq data (Nieto-Barajas, 2012).

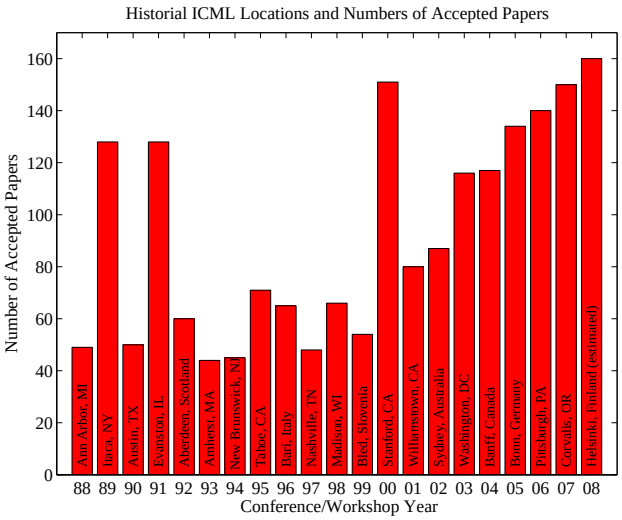

Figure 1. This is a demo figure.

We report the results using various Bayesian models to analyze gene expression data. Our exploration includes parallelized LDA, mixture models, dynamic-time topic models, topic-clustering models, and non-parametric models, implemented in `numpy`, `C++`, and `Edward`. We evaluate our methods using held-out likelihood, posterior predictive checks, and biological meaningfulness testing.

Algorithm 1 Latent Dirichlet Allocation

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 $true$

Algorithm 2 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 $true$

2. Latent Dirichlet Allocation

2.1. Model Description

Algorithm 2 describes the generative process for LDA.

2.2. Implementation

2.3. Experiments

3. Dirichlet Mixture Model

3.1. Model Description

Algorithm 2 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. Document 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

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

Blei, et al. Hierarchical dirichlet processes. http://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf, NIPS 2005.

Blei, et al. The IBP compound dirichlet process and its application to focused topic modeling. http://www.cs.columbia.edu/~blei/papers/WilliamsonWangHellerBlei2010.pdf, ICML 2010.

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

Crow, et al. Exploiting single-cell expression to characterize co-expression replicability. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0964-6, Genome Biology 2016.

Nieto-Barajas, et al. A time-series DDP for functional proteomics profiles. https://www.ma.utexas.edu/users/pmueller/pap/NM12.pdf, Biometrics 2012.

Xie, et al. Integrating document clustering and topic modeling. https://arxiv.org/pdf/1309.6874.pdf, NIPS 2013.

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.

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