Bayesian Clustering and Topic Discovery in Melonoma Cell Gene Expression Data

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

Abstract

We present a report of the methods and results of our use of various Bayesian models to analyze gene expression data. We explored parallelized LDA, mixture models, dynamic-time topic models, topic clustering models, and non-parametric models, implemented in `numpy`, `C++`, and `Edward`. Overcoming problems associated with the size and dimensionality of our dataset and difficulty of posterior inference, we report held-out likelihood, performance metrics, posterior predictive checks and, most notably, meaningful biologically-meaningful topics.

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.

2. Prior work

3. Format of the Paper

3.1. Partitioning the Text

3.2. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

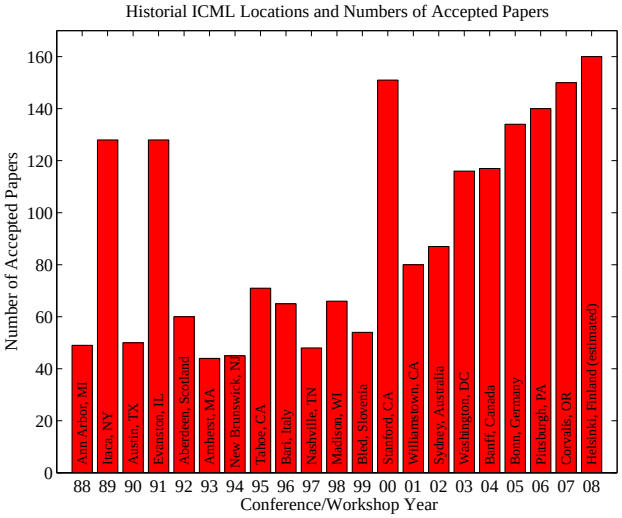

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

3.3. Citations and References

Acknowledgments

support.

References

[1] Spearman-based hierarchical clustering of scRNA-seq. http://www.nature.com/nsmb/journal/v24/n3/fig_tab/nsmb.3365_SF1.html

[2] Exploiting single-cell expression to characterize co-expression replicability. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0964-6

[3] SC3 - consensus clustering of single-cell RNA-

Algorithm 1 Bubble Sort

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 $true$

Seq data. http://biorxiv.org/content/early/2016/09/02/036558

[4] SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004575

[5] Integrating Document Clustering and Topic Modeling. https://arxiv.org/pdf/1309.6874.pdf

[6] Latent Dirichlet Allocation. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[7] Hierarchical Dirichlet Processes http://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf

[8] The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling http://www.cs.columbia.edu/~blei/papers/WilliamsonWangHellerBlei2010.pdf

[9] A Time-Series DDP for Functional Proteomics Profiles https://www.ma.utexas.edu/users/pmueller/pap/NM12.pdf