Automatic determination of the number of mixture components in Flow Cytometry with Variational Bayes

Hannes Bretschneider[1,2], Andrew Roth[3]
[1] Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada
[2] School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, 10099, Germany
[3] Department Of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

Modern flow cytometry platforms allow for the collection of data sets of increasing dimension and size. This poses a major challenge to manual analysis of flow cytometry data. In particular, accurately gating high dimensional data that cannot be directly visualized is difficult.

Recently, mixture models have been proposed as an automated means of gating flow cytometry data. Using Gaussian densities, or more robust Student-t densities, mixture models can cluster flow data in a statistically meaningful way. A major challenge to the use of mixture models is the requirement of a-priori specification of the number of clusters. If the number of clusters is unknown, we can treat determination of the correct number of clusters as a model selection problem. Using this approach can be computationally expensive, requiring multiple runs of the software with varying numbers of clusters specified.

We propose a computationally cheaper alternative that allows us to fit a Student-t mixture model (SMM) in a single run using a Variational Bayes (VB) inference algorithm. SMM's have been previously used to analyse flow cytometry data, and generally outperform Gaussian based solutions because of the robustness of the Student-t distribution to outliers. Our contribution is the implementation of an efficient inference algorithm, which through the use of sparsity promoting priors allows for automatic determination of the number of clusters. In contrast to model selection based methods, we can determine the number of clusters in a single run leading to dramatic decrease in run times.

References:
[1] C Archambeau and M Verleysen. Robust bayesian clustering. *Neural Networks*, 20(1):129–138, 2007.
[2] K Lo, RR Brinkman, and R Gottardo. Automated gating of flow cytometry data via robust model-based clustering. Cytometry Part A, 73(4):321–332, 2008.
[3] S Pyne, X Hu, K Wang, E Rossin, T Lin, LM Maier, C Baecher-Allan, GJ McLachlan, P Tamayo, DA Hafler, PL De Jager, and JP Mesirov. Automated high- dimensional flow cytometric data analysis. *Proc Natl Acad Sci U S A*, 106(21):8519–24, May 2009.
[4] G. Finak, A. Bashashati, R. Brinkman, and R. Gottardo. Merging Mixture Components for Cell Population Identification in Flow Cytometry. *Advances in Bioinformatics,* 2009.
[5] A Rosenberg and J Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* 410–420, 2007.