Clustering Biological Data Sets

Kevin R. Coombes[1], Gerard Lozanski[2], Steven M. Kornblau[3]

[1]*Department of Biomedical Informatics, The Ohio State University Wexner Medical Center, Columbus, OH 43210*

[2]*Department of Pathology, The Ohio State University Wexner Medical Center, Columbus, OH 43210*

[3]*Department of Bone Marrow Transplantation, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030*

1 abstract

something should go here

Keywords hierarchical clustering, principal components analysis

Since the earliest days of gene expression microarrays, two-way clustered heatmaps have been a standard feature of most papers studying genome-wide biological datasets [1,2]. Such heatmaps remain ubiquitous, in spite of numerous difficulties in interpretation, reproducibility, and in assigning statistical significance.

In this article, we present a comprehensive method for combining principal comnponents analysis (PCA) with hierachical clustering to analyze biological datasets. This method includes

- Statistical tools for removing "outlier" features (i.e., genes, proteins, etc.) that do not contribute to clustering.
- Statistical tools for determining the significant number of principal components and the number of feature clusters.
- A method for automatically selecting the metric for hierarchical clustering features that best matches the clusters derived from PCA.
- more cool stuff

To deermine the theoretical performance characteristics of the method, we perform two sets of simulations. In the first set of simulations, each data set includes approximately 15 proteins, belonging to one or two significant principal components (PCs) and between one and four feaure clusters, and two outlier features. In the second set of simulations, each data set contains approximately 100 features, belonging to five signifcant PCs and ten feature clusters, and approximately 30 outlier features. We then apply the method to two actual biological datasets. The first consists of protein expression data collected on samples from patients with acute myelogenous leukemia (AML) using reverse phase protein lysate arrays (RPPA). The second consists of complex flow cytometry data from either peripheral blood or apheresis samples of patients with a variety of conditions and treatments.

2 Results

First Simuklation We simulated

Second Simulation

RPPA Data From AML Patients

Flow Cytometry Data

3 Methods

All analyses were performed using vesion 3.0.0 of the R statistical software environment [3] with version 1.0.0 of the `Thresher` package, which we developed.

Here we describe the `Thresher` algorithms that we use to cluster biological datasets. Throughout, we assume that previous feature selection methods have reduced the number of features to less than the number of independent samples in the data set. (The feature selection step may be statistical in nature or biological. For example, one could use only the genes or proteins that are par of a knwon biological pathway.) The data from each feature are then standardized to have mean zero and standard deviation one.

Number of Principal Components To determine the number of significant PCs, we begin with the visual Bayesian method proposed by Auer and Gervini [4]. Briefly, they place an exponential prior on the number of signficant components. This prior distribution depends on a hyperparameter, $\theta \in [0, \infty]$, that governs how quickly the distribution decays. When $\theta = 0$, the prior is flat, and the maximum a posteriori (MAP) estimate of the number of PCs is always equal to the number of features. As $\theta \mapsto \infty$, the prior drops off more rapidly, and the MAP estimate of the number of PCs will $\mapsto 0$. Auer and Gervini propose plotting the MAP estimate as a (step) function of θ, and then selecting the highest step whose length is "nontrivial".

We have operationalized the final subjective step in their approach by putting an upper bound on the largest reasonable value of θ (specifically, BLAH) and then defining “nontrivial” to mean at least twice the median length of a step.

Outlier Detection Having established the number N of significant PCs, we work in the principal component space containing the first N PCs. Each feature F is given a “loading” that describes its contribution to each PC, and so can be represented by an N-dimensional vector. The length, $||F||$, of this vector (in the sense of Euclidean distance in PC space) summarizes the full extent of its contributions to describing any structure present in the data. Based on the result of simulations, we identify a feature to be an outlier if $||F|| < 0.3$.

Number of Feature Clusters by PCA After removing outliers, we repeat PCA on the reduced data set, and normalize the features in N-dimensional PC space to have unit length ($F_{(1)} = F/||F||$); each normalized vector records the direction of a feature as a point on a unit hypersphere. We cluster these points using a mixture of von Mises-Fisher distributions [5,6] as implemented in version 0.1-2 of the `movMF` package for the R statistcial software environment.

4 Conclusions

1. Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace Jr, A. J., Kohn, K. W., Fojo, T., Bates, S. E., Rubinstein, L. V., Anderson, N. L., Buolamwini, J. K., van Osdol, W. W., Monks, A. P., Scudiero, D. A., Sausville, E. A., Zaharevitz, D. W., Bunow, B., Viswanadhan, V. N., Johnson, G. S., Wittes, R. E., and Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–9 (1997).

2. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863–8 (1998).

3. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2013).

4. Auer, P. and Gervini, D. Choosing principal components: a new graphical method based on bayesian model selection. *Communications in Statistics - Simulation and Computation* **37**, 962–977 (2008).

5. Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research* **6**, 1–39 (2005).

6. Hornik, K. and Grün, B. On conjugate families and Jeffreys priors for von Mises-Fisher distributions. *Journal of Statistical Planning and Inference* **143**(5), 992–999, May (2013).