On the Use of NMF and curvHDR to Cluster Flow Cytometry Data

José M. Maisog[1, 2], Andrea ~~G~~A. Barbo[2], George ~~L~~Luta[* 2]
[1]Medical Numerics, Inc., Germantown, MD, 20876 USA; ,~~-~~[2]Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington, DC 25700, USA

We used non-negative matrix factorization (NMF) in combination with the curvHDR method to identify cell populations in flow cytometry data for F~~f~~lowCAP Challenge 2, in which the problem was unsupervised clustering with an unknown number of clusters. NMF (Lee and Seung, 1999; for review, see Devarajan, 2008; http://cran.r-project.org/web/packages/NMF/index.html) is a relatively new matrix factorization approach, for which many algorithms have been developed; we are aware of only one paper that used NMF in the analysis of flow cytometry data (de Jager et al., 2008). Given an M x N matrix X with non-negative values, NMF factors X into an M x k matrix W and a k x N matrix H, such that W * H approximates X as closely as possible, all values of W and H are non-negative, and k < M,N. If the rows of X are the observations / samples, and the columns of X are the variables / "channels", then the k rows of H can be considered components / latent "parts" into which X has been decomposed, and the rows of W can be considered the "expression patterns" / encodings of the corresponding samples in the basis defined by the rows of H. NMF can be used for both clustering (especially sparse NMF methods) and dimensionality reduction. curvHDR (Naumann et al., 2010;http://www.uow.edu.au/~mwand/Rpacks.html) is a very new method developed specifically for unsupervised clustering of flow cytometry data. To perform gating of flow cytometry data, the curvHDR method makes use of the notions of significant high negative curvature regions and highest density regions. One feature of curvHDR is that it does not force all data into clusters; that is, some data is allowed to fall outside the defined gates. In our submission for ~~f~~FlowCAP Challenge 2, we used NMF to reduce the dimensionality to 2, and then we applied curvHDR on the resulting encodings to perform unsupervised clustering with an unknown number of clusters. (curvHDR was available only in a 2D implementation at the time of our work, hence it was necessary to reduce the dimensionality to 2; a 3D implementation of curvHDR will be released very soon.) Our primary interest was to determine whether the combined use of NMF and curvHDR might yield clustering results similar to manual gating of flow cytometry data.

References:

De Jager PL, Rossin E, Pyne S, Tamayo P, Ottoboni L, Viglietta V, Weiner M, Soler D, Izmailova E, Faron-Yowe L, O'Brien C, Freeman S, Granados S, Parker A, Roubenoff R, Mesirov JP, Khoury SJ, Hafler DA, Weiner HL. Cytometric profiling in multiple sclerosis uncovers patient population structure and a reduction of CD8low cells. Brain. 2008 Jul;131(Pt 7):1701-11.

Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol. 2008 Jul 25;4(7):e1000029.

Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999 Oct 21;401(6755):788-91.

Naumann U, Luta G, Wand MP. The curvHDR method for gating flow cytometry samples. BMC Bioinformatics. 2010 Jan 22;11:44.

Formatted: Font: Bold

Formatted: Portuguese (Brazil)

Formatted: Portuguese (Brazil)

Formatted: Portuguese (Brazil)