Differential Projection Pursuit and its Application to Cell Flow Cytometry Analysis

Mahan Dastgiri[1], Yajie Duan[1], Davit Sargsyan[2,3,4], Ge Cheng[1], Kanaka Tatikola[4], and Javier Cabrera[1]

Department of Statistics, Rutgers University-New Brunswick

Abstract

Differential projection pursuit (DPP) is a proposed method to find regions with maximal difference between distributions. Multicolor flow cytometry is a laboratory technique to identify cell subpopulations by measuring their physical and biochemical characteristics. Data analysis in flow cytometry relies on gating, the process of manually selecting successive subpopulations of cells using two-dimensional plots. Plotting variables two at a time could mask hidden structures present in the data and manual selection makes the analysis inconsistent. The new method could automate flow cytometry analysis by utilizing projection pursuit, data nuggets and factor analysis. When applied to flow cytometry data, DPP allows researchers to quickly identify differences in cell populations exposed to different experimental conditions. DPP creates a platform to explore differences in large datasets and improves the analysis clarity and reproducibility by considering the data in its true dimensional space and through automation, respectively.

Flow Cytometry

- Multicolor flow cytometry measures individual cell properties by using fluorescence and forward (FSC) or side scatter (SSC) light.
- Tools such as FlowJo plot the data, two dimensions at a time.
- Clusters of cells are identified manually by drawing gates.

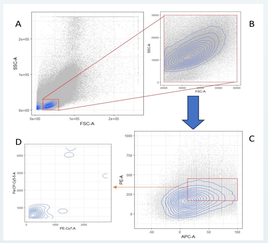

Figure 1: An example of flow cytometry gating.

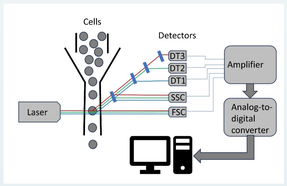

Figure 2: Schematics of a flow cytometer.

Projection Pursuit

- Projection pursuit (PP) is a technique to find low-dimensional projections revealing interesting structures within the data.
- A PP index is a function to numerically measure features of low-dimensional projections.
- Natural Hermite index (1) is a PP index to find non-linear structures and departures from normality.[1]
- Let $X \in R^{nxp}$ be a p-dimensional dataset, $P \in R^{pxd}$ be a random orthogonal projection matrix that projects X into a d-dimensional space ($d < p$), $Y = (y_1, y_2, \dots, y_n)' = XP \in R^{nxd}$ be the projected data in the d-dimensional space, $f(y)$ be the density of the projected data Y, and $\phi(y)$ be the standard multivariate Gaussian density. Then, the natural Hermite index is:

$$I^N = \int_{\mathbb{R}^d} [f(y) - \phi(y)]^2 \phi(y) dy \qquad (1)$$

Data Nuggets

- Data nuggets is a data compression algorithm that preserves the structure of the data.[2] It represents a dataset with millions of observations as a weighted set of a few thousand nuggets.
- Each nugget is represented by its center, weight, and scale parameters.

Weighted natural Hermite index

- A weighted version of natural Hermite index for large datasets was introduced using data nuggets,[3] where the density $f(y)$ is estimated by:

$$\hat{f}_B(y) = \sum_{i=1}^{m} \frac{w_i}{\sum_{i=1}^{m} w_i} |S_i|^{-\frac{1}{2}} \phi\left(S_i^{-\frac{1}{2}}(y - y_i)\right) \qquad (2)$$

Differential Projection Pursuit

- DPP finds projections with maximal difference between ≥ 2 distributions.
- Let $f_1(y)$ and $f_2(y)$ be any two d-dimensional distributions, and $f(y) = \frac{l_1 f_1(y) + l_2 f_2(y)}{l_1 + l_2}$.
- Differential Hermite index is is given by the following formula:

$$I^{dH} = d_H(f_1, f_2) = \int_{\mathbb{R}^d} [f_1(y) - f_2(y)]^2 f(y) dy \qquad (3)$$

- Let $s_k, w_k \in \mathbb{R}^m$ be the scale and the weight vector of data nuggets for $k = 1, 2$, respectively, and m be the number of data nuggets. Then, $f_1(y)$ and $f_2(y)$ are estimated by the Kernel density estimator $\hat{f}_k(y)$, where

$$\hat{f}_k(y) = \sum_{i=1}^{m} \frac{w_{k,i}}{\sum_{i=1}^{m} w_{k,i}} |S_{k,i}|^{-\frac{1}{2}} \phi\left(S_{k,i}^{-\frac{1}{2}}(y - y_{k,i})\right), k = 1,2 \qquad (4)$$

- The steps of the DPP algorithm are:
  i. Process the raw data into data nuggets and spherize the nuggets
  ii. Apply a projection matrix P to the two datasets: $Y_1 = X_1 P$ and $Y_2 = X_2 P$.
  iii. Calculate density estimators $\hat{f}_1(y)$ and $\hat{f}_2(y)$ and their average $\hat{f}(y)$ from the projected data Y_1, Y_2.
  iv. Calculate the index I^{dH} for projection P using $\hat{f}_1(y)$ and $\hat{f}_2(y)$.
  v. Repeat step (ii) through (iv) to maximize I^{dH} over all projections P.
- For comparison of $k > 2$ samples, the algorithm can be extended using the following formula[4]:

$$\sum_{i<j} d_H(\hat{f}_i, \hat{f}_j) = c \sum_i d_H(\hat{f}_i, \hat{f}) \qquad (5)$$

Data

- The dataset contains LPS-stimulated and unstimulated blood samples from 20 HIV-exposed uninfected (HEU) and 24 unexposed (UE) African infants.
- These files contained a total of 42M+ cells, reduced to ~22M cells after excluding debris, and ~11M after centering the LPS-stimulated samples using unstimulated samples.

Data (continued)

- The final ~11M cells were compressed into 4,858 data nuggets.
- Each file contains FSC, SSC and 8 protein marker measurements.

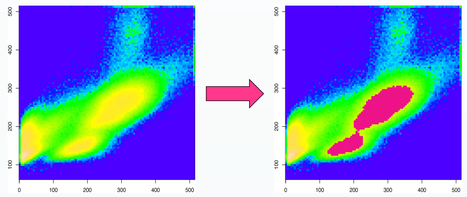

Figure 3: Excluding debris from live cells using density landmarks.

Multivariate clustering on the data nuggets

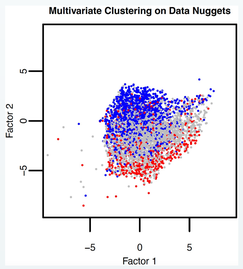

Figure 4: Multivariate weighted K-Means clustering done on 4,858 nuggets with K=3.

DPP: Regions of maximum difference

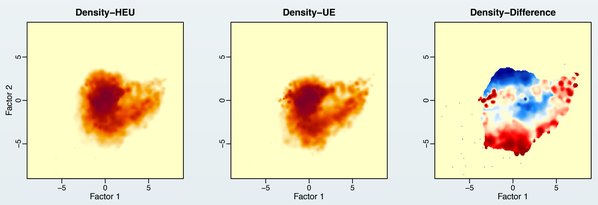

Figure 5: Density plot of the projected data nuggets. The projection was rotated using the varimax procedure. The blue and red in the third plot correspond to the positive and negative difference between the two densities, respectively.

DPP: Defining regions based of density

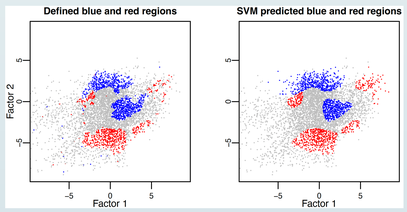

Figure 6: Blue and red nuggets are defined using density values (left), SVM was used to predict the blue and red regions (right).

Alignment with expected changes

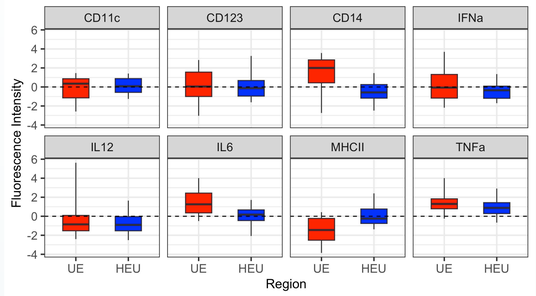

Figure 7: Weighted boxplots of the fluorescence intensity for the eight flow cytometry channels in the predicted blue and red region.

Clusters with distinct protein levels

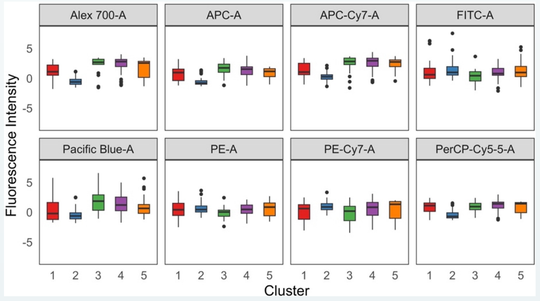

Figure 8: Example of weighted boxplots of the fluorescence intensity for each cluster of the blue region across the eight channels.

Conclusion

- DPP identifies regions of maximal difference between distributions.
- DPP proposes a multidimensional and reproducible approach for analyzing flow cytometry data.
- DPP application results align with biological studies.

References

1. Cook, D., A. Buja, and J. Cabrera. 1993. Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics 2*: 225-250.
2. Beavers, T., J. Cabrera, G. Chen, Y. Duan, M. Lubomirski, and J. E. Tiggler. 2023. Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. *arXiv preprint arXiv:2403.03099.*
3. Duan, Y., J. Cabrera, and B. Emir. 2023. New Projection Pursuit Index for Big Data. *arXiv preprint arXiv:2312.06465*.
4. Weigle, S., D. Sargsyan, J. Cabrera. 2023. A Hermite index randomization method for pre-clinical studies. (submitted)

Affiliations

[1]Department of Statistics, School of Arts and Sciences, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[2]Department of Pharmaceutics, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[3]Graduate Program in Pharmaceutical Science, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[4]Janssen Pharmaceuticals, Johnson and Johnson, Spring House, PA, USA, and Beerse, BE