Novel machine learning approach to differential flow cytometry analysis base on projection pursuit

Mahan Dastgiri [1], Yajie Duan[1], Davit Sargsyan[2,3,4], Abraham Adkwei[4,5], Rebecca Mary Peters[2,3], PoChung Chou[2,3], Ge Cheng[1], Chun-Pang Lin[1], Jocelyn Sendecki[4], Helena Geys[4], Kanaka Tatikola[4], Ah-Ng Kong[2,3] and Javier Cabrera[1]

[1]Department of Statistics, School of Arts and Sciences, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[2]Department of Pharmaceutics, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[3]Graduate Program in Pharmaceutical Science, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[4]Janssen Pharmaceuticals, Johnson and Johnson, Spring House, PA, USA, and Beerse, BE

[5]Epidemiology and Statistics Department, University of Georgia, GA, USA

SUBMIT TO: Special Issue of the Journal of Biopharmaceutical Statistics (JBS) special issue on AI/ML

Correspondence

Professor Javier Cabrera

Rutgers, the State University of New Jersey

School of Arts and Sciences, Room XXX

XXXXXXXX, Piscataway, NJ 08854

Phone: +1-XXX-XXX-XXXX

Email: xavier.cabrera@gmail.com

RESOURCES (delete before submission)

1. Wiki page:

https://en.wikipedia.org/wiki/Flow_cytometry#

2. FCS standards (also, see publications):

https://isac-net.org/

3.

Table of Contents

Abstract

Summary here. Write last.

1 Background

1.1 Key principals of flow cytometry

Multicolor flow cytometry (FC) is a laboratory technique used in biological studies to measure individual cell properties such as size, granularity and molecular composition. To measure specific proteins on the surface or inside a cell, fluorescent chemical compounds called fluorochromes or fluorophores are added to the suspension. The fluorochromes are attached to molecules with affinity to specific proteins, hence labeling these proteins. Cells are first separated and suspended in a liquid, and the suspension is pass through narrow tubes, one cell at a time. The instruments contain a large number of such tubes for parallel processing. As a cell moves through a tube, it is hit by a beam of light from a lamp or a laser. The light-excites fluorochromes then emit light in a relatively narrow band of wavelengths. The emitted light passes through a series of optical filters and dichroic mirrors deflecting it onto detectors (Figure X <DRAW CYTOMETER PICTURE>). Besides measuring light emitted by fluorochromes, flow cytometers also detect light scattered by the cells forward or to the side (FSC and SSC, respectively). The FSC and SSC measurements provide information about the cells' physical properties and are used to separate single, live cells from cell clusters and debris during data preprocessing. Additionally, the instruments are able to measure electrical current impedance, i.e., the opposition to alternative current as the cells travel through the tubes. This allows for calculation of the cell size and additional physical properties. As of 2023, flow cytometers may contain as many as 10 lasers and up to 30 fluorochrome detectors.

The detectors convert the analog signal into digital and send the data to the instrument's computer. The data collection process in flow cytometer is called ***acquisition***. The data is typically saved in Flow Cytometry Standards (FCS) format as a matrix, with rows representing individual cells and columns the markers [1]. FCS specifications were developed and are maintained by the international Society for Advancement of Cytometry (ISAC).

Flow cytometry is used in biology to achieve a variety of goals including cell genotyping, sorting and studying apoptosis but, in this work, only one specific, widely used experimental design will be considered, namely, studying treatment effects on immune cell differentiation. Administering potent test compounds to naïve immune cells leads to their specialization that is manifested through changes in cell surface markers. Cytometers identify and quantify these markers allowing for differential analysis of the samples across treatment groups.

Following acquisition, the data is processed, traditionally using a technique called ***gating***. Specialized tools such as FlowJo and … import FSC files and plot the data, 2 dimensions at the time. The investigator draws areas of interests, or gates, to manually identify clusters of cells that they are interested in. This process of gating goes on sequentially as the investigator focuses on specific subpopulations of cells. The gating strategy follows current understanding of differentiation process, with major differentiating proteins gated first (Figure X <ADD PLOT EXAMPLE OF GATING STRATEGY>). Once gating is completed, the software will count the number of cells in each gate and output a processed data file. Often, the interest is not or not only the counts, but the ratios of child-parent populations as defined by the gating strategy, i.e., frequencies.

2 Materials and Methods

2.1 Data Source and Experimental Design

The data was obtained from the Flow Repository website (https://flowrepository.org/) …<@Mahan: provide the link and details on the specific dataset used, including experimental design>

The HIV-exposed-uninfected versus unexposed (HEUvsUE) data was collected to find cell populations that can be used to differentiate between HIV-exposed-uninfected (HEU) and unexposed (UE) infants. For this purpose, blood samples were taken from infants 6 months after birth, where some were stimulated with six Toll-like–receptor ligands and some were left unstimulated for control. There are 308 FCS files from 40 patients. In this study, we aimed to find the region within the data where stimulated and unstimulated cells differ the most. We chose data files from 2 patients, one HIV-exposed-uninfected and one unexposed. For each patient we considered LPS stimulated and unstimulated. Each data file provides information on 8 cell markers. The details can be seen in Table 1.

Table 1: List of reporter and analytes for the datasets.

Channel	Reagent
FSC-A	
SSC-A	
FITC-A	IFNa
PE-A	CD123
PerCP-Cy5-5-A	MHCII
PE-Cy7-A	CD14
APC-A	CD11c
APC-Cy7-A	IL6
Pacific Blue-A	IL12
Alex 700-A	TNFa

2.2 Data compression with Data Nuggets. (Javier and Kanaka)

Our calculations in the rest of this paper are not possible to do based on the raw data because when you have 1.5 million observations any calculation of order n^2 is not computable. For this reason we apply a compression algorithm call data nuggets (see Beavers e.a. 2022) that represents a dataset of millions of observations with a weighted set of a few thousand observation that are called data-nugets. Data-nuggets preserve the structure of the data much better than random samples and for this reason they are more suitable to find the true data structures using low dimensional projections. In summary data nuggets compression reduces a large dataset into a smaller collection of data nuggets while preserving the underlying structure.

2.3 Projection pursuit and differential projection pursuit(Javier)

Projection Pursuit is a technique that searchers for projections of multivariate p-dimensional data into lower d-dimensional projections containing the main structure of the data. By main structure we mean clusters, outliers and any other low dimensional nonlinear structure. These methods were introduced by Friedman, Tukey (1972) for finding structure, while exploring a 9-dimensional data from particle physics. Later Friedman (1982) introduce Friedman index as an example of projection pursuit (PP) indices. Cook e.a (1993) made substantial progress in this area by introducing several new PP indices. In particular, they introduced the natural Hermite index that became very popular and will be the center tool of this paper. The natural Hermite index measures the distance between any *d*-dimensional distribution $f(y)$ and a *d*-dimensional normal distribution $\phi(y)$.

$$I^N = \int_{\mathbb{R}^d} \{f(\mathbf{y}) - \phi(\mathbf{y})\}^2 \phi(\mathbf{y}) d\mathbf{y}$$

The computational burden of these indices is satisfactory for small to moderate data sets but not attainable for large datasets. Therefore Duan, Cabrera (2023) introduced weighted versions of the PP indices computed over data nuggets. They showed that the “most interesting” projections found by the Natural Hermite index on large datasets are identical to those found by the weighted version of the index over data-nuggets.

In our case we are interested in comparing d-dimensional projections of two p-dimensional samples where d < p. For this we are introducing a differential version of the Hermite index for comparing any two d-dimensional distributions $f_1(y)$ and $f_2(y)$. Let $f(y) = \frac{f_1(y)+f_2(y)}{2}$. The differential Hermite index for two distributions is given by the following formula:

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

This index is minimized over all d-dimensional projections of our data, where $f_1(y)$ and $f_2(y)$ are estimated by the d-dimensional Kernel density estimators of our projected data $\hat{f}_1(y)$ and $\hat{f}_2(y)$. The steps are

(i) Apply a projection matrix *P* to our two data sets: $Y_1 = PX_1$ and $Y_2 = PX_2$.

(ii) Calculate density estimators $\hat{f}_1(y)$ and $\hat{f}_2(y)$ from our projected data Y_1, Y_2 and their average $\hat{f}(y)$.

(iii) Calculate the index I^{dH} for projection P using $\hat{f}_1(y)$ and $\hat{f}_2(y)$ and .

(iv) Repeat step (i)-(iii) to maximize I^{dH} over all projections P.

This procedure usually finds a local maximum, so it needs to be repeated a few times with different initial P's to obtain some good local maxima. Usually, we will obtain 3 or 4 local maxima projections.

When we want to compare *k* >2 samples we require to evaluate *k*(*k*-1)/2 integrals. But in Weigle e.a.(2023) it was shown that

$$\sum_{i<j} d_H(\hat{f}_{\mathrm{i}}, \hat{f}_{\mathrm{j}})=\mathrm{c} \sum_i d_H(\hat{f}_{\mathrm{i}}, \hat{f}),$$

which requires to evaluate only *k* integrals. Therefore the algorithm to find the optimal projection for the difference between two groups can be extended to *k* groups.

2.4 Factor analysis and clustering

@Javier, @Yajie, @Mahan

3 Results

@Mahan @Javier and @Yajie to provide the results, @Davit and others to interpret.

After obtaining 3… refined data nuggets from the HEUvsUE dataset, we obtained six projections and conducted varimax rotations on them which are pictured in Figure 1. We have estimated density plot for the Stimulated, Unstimulated data points as well as difference between them which is plotted in the third column. The blue region shows the positive difference, which is the area we are interested to focus on in this study. We conducted SVM based on the 2-d projections and based on the projected nugget centers and predicted the blue region. In Table 2, we see the proportion of Stimulated and Unstimulated data points exists in the predicted blue clusters for each projection, as well as the ratio of the stimulated and unstimulated, and number of stimulated and unstimulated cells. For the rest of the analysis, we focused on projections 1 and 3. The loading for each protein for the projection 1 and 3 are represented in the loadings plots shown in Figure 4. For the first projection, protein … contributes strongly to the first component, protein … to the second component, and protein … have strong relationships with both components. For the third projection, protein … strongly correlated with the first component, protein … with the second component, and protein… contributes significantly to both directions. In order to detect clusters of interest, weighted K-means was applied to the predicted blue regions for the two projections, and the clustering results are pictured in Figure 2. The optimal number of clusters were 5 and 7 for projections 1 and 3, respectively. The box plots in Figure 3 shows the protein expression levels for each cluster. Tables 3 and 4 represents the proportion of Stimulated and Unstimulated data points exists in the predicted blue clusters for each projection, as well as the ratio of the stimulated and unstimulated, and number of stimulated and unstimulated cells.

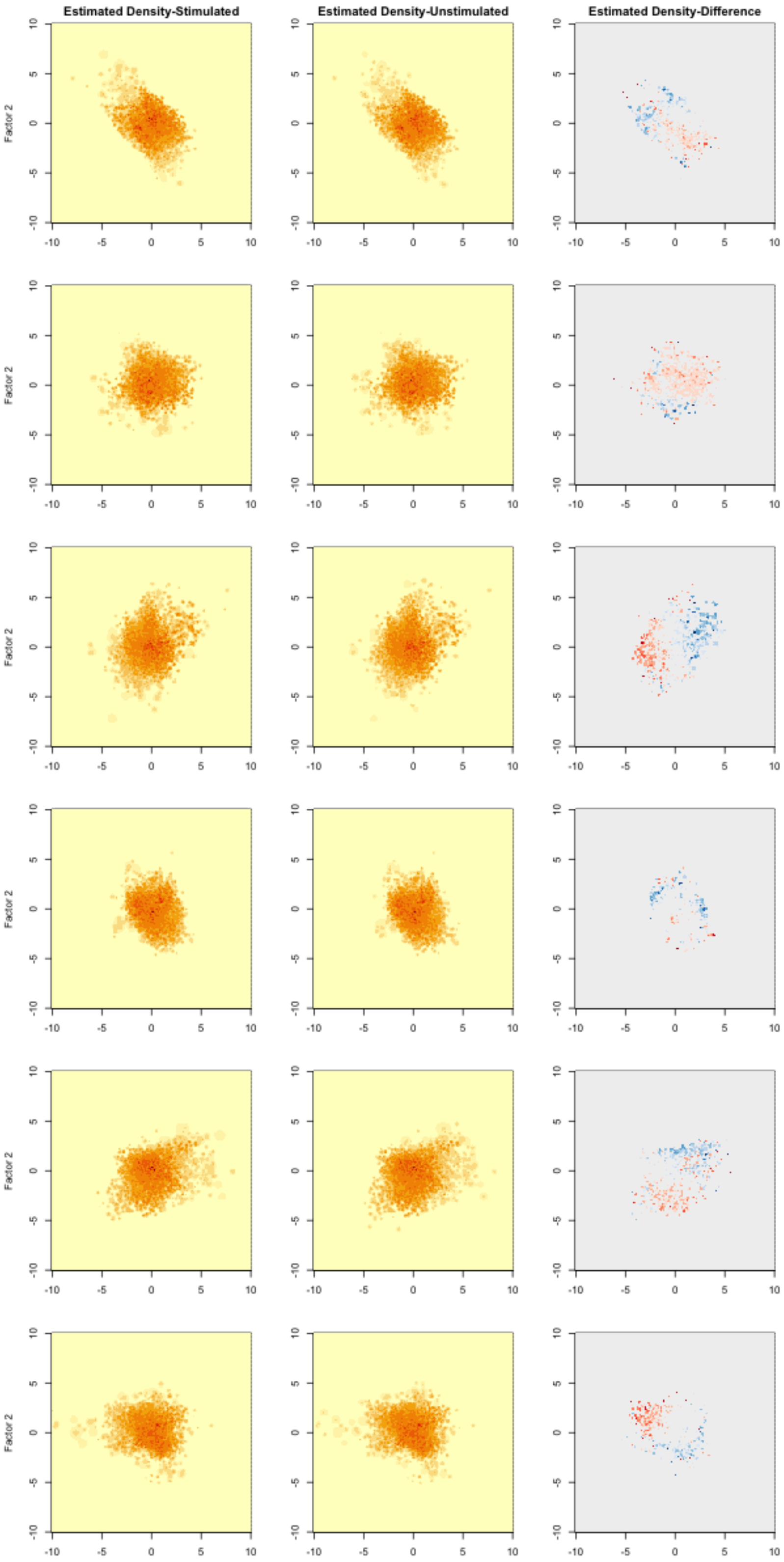
Estimated Density-Stimulated
Estimated Density-Unstimulated
Estimated Density-Difference
Factor 2
10
5
0
-5
-10

Figure 1: Density plot of Stimulated (first column), Unstimulated (second column) data nuggets, and the difference between them (third column) for 6 projections that were obtained by optimizing the data nuggets projection pursuit Hermite index.

In Figure 1, we display six projections that were produced by the optimization of the data nuggets Hermite index. The column of graphs on the right shows several blue clusters where the stimulated cells were more abundant than the unstimulated. Alternatively, the red clusters are the regions where the unstimulated cells are more abundant than stimulated.

Table 2: Proportion of the stimulated and unstimulated cells in the predicted blue region. The three numbers in the third column are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each of the six projections.

(the two tables are the same with different layout; will keep one)

Projection		Predicted blue cluster based on 2-d projection	Predicted blue cluster based on nugget centers
1	Stimulated	0.020 (53.5%) 10870	0.087 (53%) 46624
	Unstimulated	0.010 (46.5%) 9464	0.044 (47%) 41328
2	Stimulated	0.014 (51.5%) 7678	0.110 (51.9%) 59130
	Unstimulated	0.008 (48.5%) 7238	0.059 (48.1%) 54700
3	Stimulated	0.017 (52.9%) 8879	0.253(49.2%) 135400
	Unstimulated	0.009 (47.1%) 7921	0.15 (50.8%) 139596
4	Stimulated	0.008 (54.4%) 4539	0.075 (54.4%) 40324
	Unstimulated	0.004 (45.6%) 3804	0.036 (45.6%) 33823
5	Stimulated	0.017 (50.5%) 8850	0.196 (49.6%) 104937
	Unstimulated	0.009 (49.5%) 8682	0.115 (50.4%) 106823
6	Stimulated	0.032 (51.4%) 17221	0.100 (53%) 53714
	Unstimulated	0.017 (48.6%) 16256	0.051 (47%) 47675

Projection	Predicted blue cluster based on 2-d projection		Predicted blue cluster based on nugget centers	
	Stimulated	Unstimulated	Stimulated	Unstimulated
1	0.020 (53.5%) 10870	0.010 (46.5%) 9464	0.087 (53%) 46624	0.044 (47%) 41328
2	0.014 (51.5%) 7678	0.008 (48.5%) 7238	0.110 (51.9%) 59130	0.059 (48.1%) 54700
3	0.017 (52.9%) 8879	0.009 (47.1%) 7921	0.253(49.2%) 135400	0.15 (50.8%) 139596
4	0.008 (54.4%) 4539	0.004 (45.6%) 3804	0.075 (54.4%) 40324	0.036 (45.6%) 33823
5	0.017 (50.5%) 8850	0.009 (49.5%) 8682	0.196 (49.6%) 104937	0.115 (50.4%) 106823
6	0.032 (51.4%) 17221	0.017 (48.6%) 16256	0.100 (53%) 53714	0.051 (47%) 47675

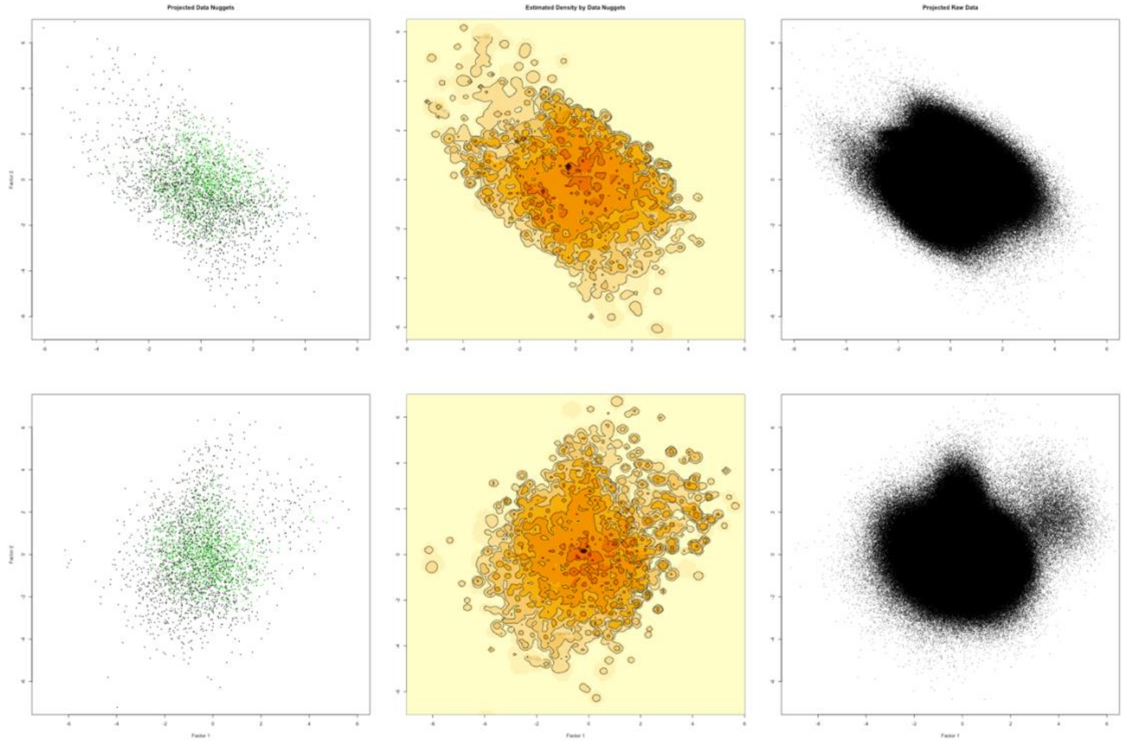

Figure 2: Projected data nuggets (first column), estimated density plot for the data nuggets (second column), and projected raw data (third column) for the first and third projections

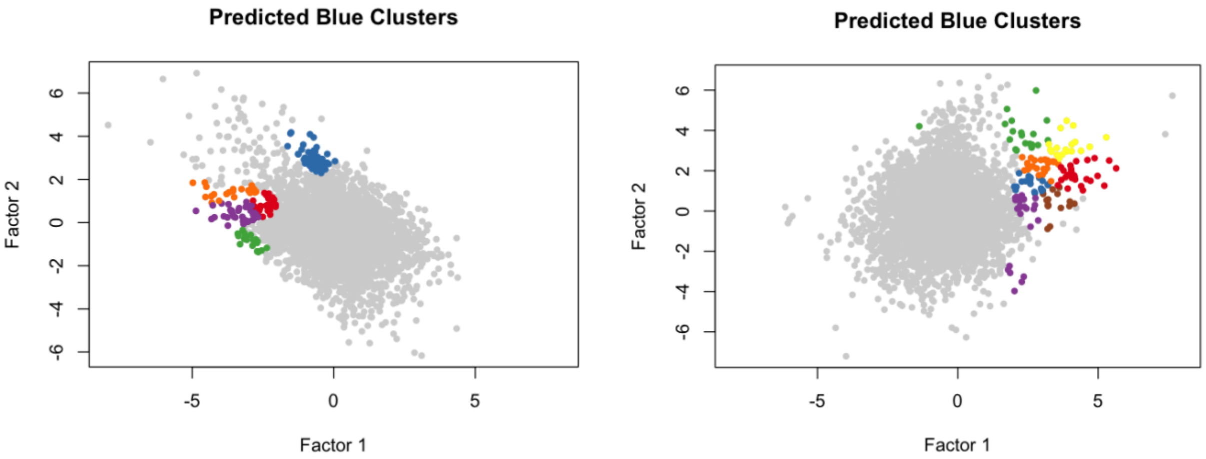

Figure 3: Clusters of the predicted blue region for projection 1 (left picture) and projection 3 (right picture)

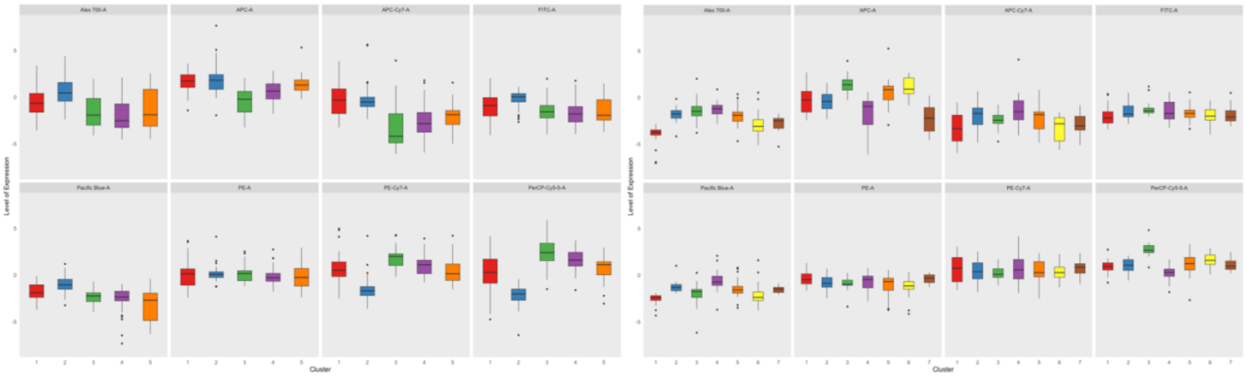

Figure 4: Expression level for each protein in the 5 clusters of the first projection (left picture) and 7 clusters of the third projection (right picture) in the predicted blue region

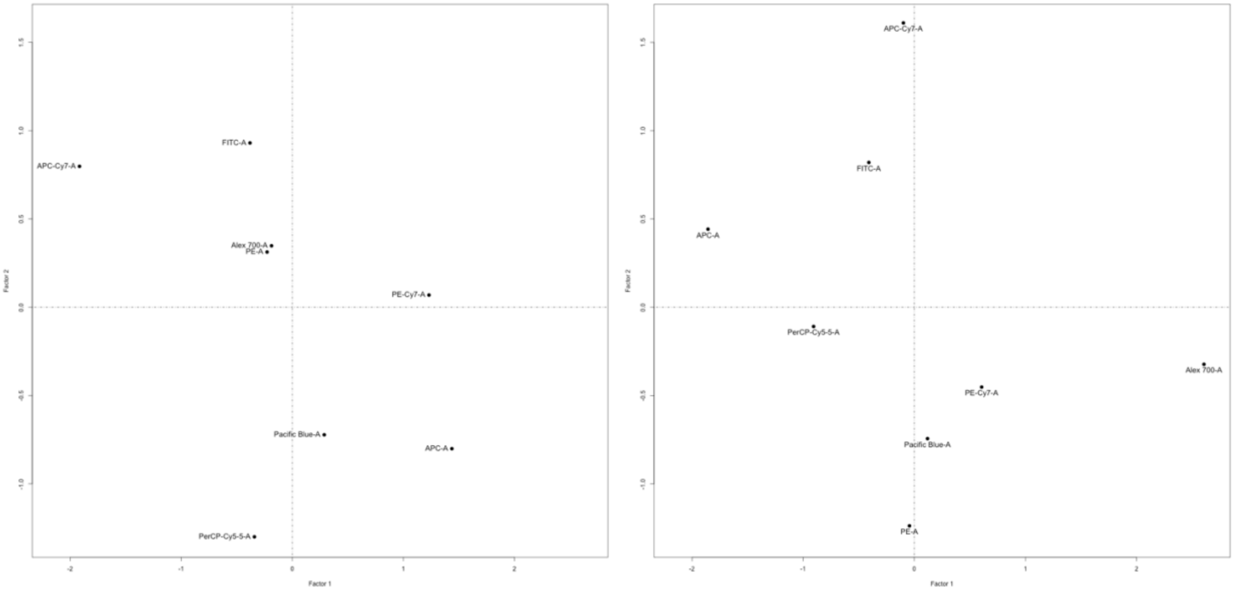

Figure 5: Loading plots for the first (left) and third (right) projection.

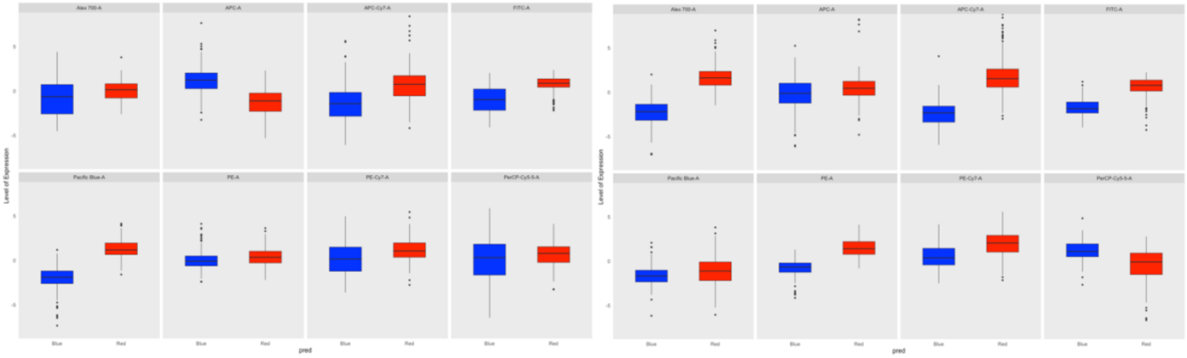

Figure 6: Expression level for each protein in the predicted blue and red regions of the first projection (left picture), and the third projection (right picture)

Table 3: Proportion of the stimulated and unstimulated cells in each 5 clusters of the predicted blue region of the first projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Stimulated	0.0038 (55.2%) 2013	0.0098 (53.9%) 5259	0.0011 (51.1%) 582	0.0039 (51.3%) 2098	0.0017 (54.0%) 920
Unstimulated	0.0018 (44.8%) 1634	0.0048 (46.1%) 4498	0.0006 (48.9%) 558	0.0021 (48.7%) 1990	0.0008 (46.0%) 784

Table 4: Proportion of the stimulated and unstimulated cells in each 7 clusters of the predicted blue region of the third projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Stimulated	0.0041 (50.5%) 2216	0.0036 (59.3%) 1904	0.0012 (55.2%) 665	0.0027 (46.3%) 1428	0.0018 (56.5%) 977	0.0013 (58.2%) 703	0.0018 (49.9%) 986
Unstimulated	0.0023 (49.5%) 2170	0.0014 (40.7%) 1307	0.0006 (44.8%) 539	0.0018 (53.7%) 1659	0.0008 (43.5%) 752	0.0005 (41.8%) 504	0.0011 (50.1%) 990

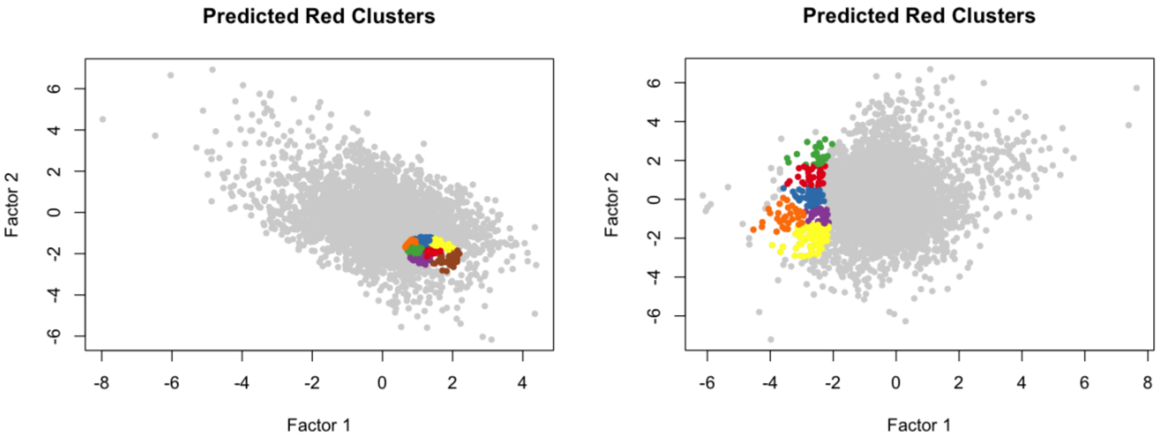

Figure 7: Clusters of the predicted red region for projection 1 (left picture) and projection 3 (right picture)

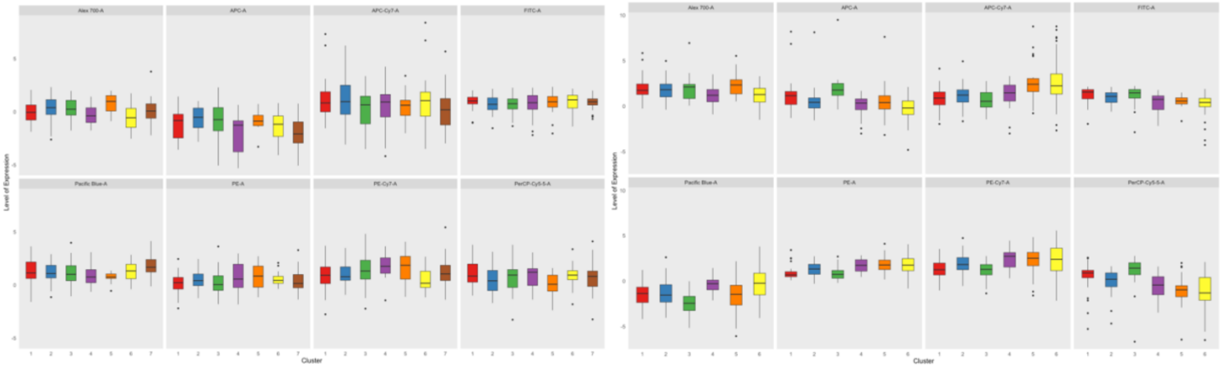

Figure 8: Expression level for each protein in the 7 clusters of the first projection (left picture) and 6 clusters of the third projection (right picture) in the predicted red region

Table 5: Proportion of the stimulated and unstimulated cells in each 7 clusters of the predicted red region of the first projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Stimulated	0.0013 (26.8%) 709	0.0039 (29.3%) 2100	0.0018 (28.9%) 966	0.0010 (27.4%) 512	0.0014 (20.1%) 774	0.0018 (27.6%) 986	0.0008 (28.3%) 429
Unstimulated	0.0021 (73.2%) 1941	0.0054 (70.7%) 5056	0.0026 (71.1%) 2372	0.0015 (72.6%) 1358	0.0033 (79.9%) 3086	0.0028 (72.4%) 2585	0.0012 (71.7%) 1089

Table 6: Proportion of the stimulated and unstimulated cells in each 6 clusters of the predicted red region of the third projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Stimulated	0.0029 (26.6%) 1546	0.0033 (25.5%) 1777	0.0007 (26.9%) 384	0.0014 (25.3%) 753	0.0005 (19.8%) 252	0.0008 (24.1%) 435
Unstimulated	0.0046 (73.4%) 4269	0.0056 (74.5%) 5196	0.0011 (73.1%) 1043	0.0024 (74.7%) 2225	0.0011 (80.2%) 1023	0.0015 (75.9%) 1372

4. Discussion

Gating approach to flow data cytometry was determined in part by biology but also by limitation of computing power and tools that would allow multidimensional data visualization and analysis. Plotting and clustering such data two dimensions at the time went around these limitations. However, such projections can present severely distorted images of a multidimensional object, masking important patterns. Additionally, results from manual gating are highly dependent on the investigator’s perception and experience and are almost certainly non-reproducible. Automated gating can improve reproducibility, but it still does not address the dimensionality issue. In this worked, we stepped back from gating and instead examined the data in its true dimensional space. Applying data nuggets reduced the amount of data by grouping individual cells into typical groups. Projection pursuit found optimal projections that revealed the most information about the data. Finally, by comparing projections of samples with different treatments, we identified cell subpopulations that had significantly different densities between the treatment groups. Some of these subgroups were identifiable using current classification of immune cells based on surface markers and physical characteristics of the cells while several of the subgroups could represent new subtypes or reveal previously unknown mechanism. The latter would require more careful examination and interpretation by biologists, as well as conformation from other data sets.

@Davit:

Based on discussion with Maggie, discuss experimental design - Stain Index, color assignment, compensation, Data from different instruments is not comparable as each instrument have its own setting, and lasers need to be calibrated whole the time…

@Javier:

1. The use of data nuggets is to make this work with very large data tables
2. Cons of using 2D for N-D data
3. WH use PP? PP finds best projections and is not attached to individual markers although we try to find projections with axis using minimal number of markers
4. Discuss data: different treatment groups so we are interested in differential analysis
5. Did we find any markers (in the results) that are interesting? Are any of the clusters interpretable? Can we find an example (or synthetic data) where a cluster is masked if you look at simple projections but revealed in PP? E.g., 4D data with 4 clusters along the 4D diagonal, e.g., a cylinder along the 4D diagonal. The clusters will be masked in simple 2D projections but if w project on the diagonal, the difference is visible even in 1D. More complicated – 4D diagonal and some linear combination. Individual 2D scatter plots will look bad; hence, gating would not work.

@ALL: please contribute

5. Figures and Tables

6. References

1 Spidlen, J., Moore, W., Parks, D., Goldberg, M., Bray, C., Bierre, P., Gorombey, P., Hyun, B., Hubbard, M., Lange, S., Lefebvre, R., Leif, R., Novo, D., Ostruszka, L., Treister, A., Wood, J., Murphy, R.F., Roederer, M., Sudar, D., Zigon, R., and Brinkman, R.R.: ‘Data File Standard for Flow Cytometry, version FCS 3.1’, Cytometry A, 2010, 77, (1), pp. 97-100

2. Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9), 881-890.

3. Friedman, J. H. (1987). Exploratory projection pursuit. Journal of the American statistical association, 82(397), 249-266.

4. Cook, D., Buja, A., & Cabrera, J. (1993). Projection pursuit indices based on orthonormal function expansions. Journal of Computational and Graphical Statistics, 2(3), 225-250.

5. Duan Y, Cabrera J, A New Projection Pursuit Index for Big Data. (in revision)

6. Beavers, T., Cabrera, J., Chen G, Duan Y, Lubomirski, Tiggler, S, M. (2019). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. (in revision)