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

Mahan Dastgiri 1, Yajie Duan1, Davit Sargsyan2,3,4, Abraham Adkwei4,5, Rebecca Mary Peters2,3, PoChung Chou2,3, Ge Cheng1, Chun-Pang Lin1, Jocelyn Sendecki4, Helena Geys4, Kanaka Tatikola4, Ah-Ng Kong2,3 and Javier Cabrera1

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

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

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

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

5Epidemiology 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](mailto:xavier.cabrera@gmail.com)

RESOURCES (delete before submission)

1. Wiki page:

[https://en.wikipedia.org/wiki/Flow\_cytometry#](https://en.wikipedia.org/wiki/Flow_cytometry)

1. FCS standards (also, see publications):

<https://isac-net.org/>

Table of Contents

[Abstract 2](#_Toc137151707)

[1. Background 2](#_Toc137151708)

[2. Materials and Methods 2](#_Toc137151709)

[2.1 Experimental Design 2](#_Toc137151710)

[3 Results 2](#_Toc137151711)

[4 Discussion 2](#_Toc137151712)

[5 Figures and Tables 2](#_Toc137151713)

[6. References 2](#_Toc137151714)

# Abstract

Summary here. Write last.

# 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>

## 2.2 Dimensionality Reduction (Rows) with Data Nuggets

@Javier and @Kanaka: can you please write this section?

## 2.3 Projection Pursuit

@Mahan and @Yajie: can you please write this section?

Start describing what are proj. pur indices. Site Yajie’s paper. Talk about different indices including LDA index (we don’t have it, but we can implement. Javier thinks it might be better, but this is for future).

## 2.4 Factor analysis and clustering

@Javier, @Yajie, @Mahan

# 3 Results

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

# 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