Docket #: S18-146

Unsupervised identification, matching, display and quantitation of subsets (clusters) by exhaustive projection pursuit methods

When examining one or higher dimensional data, researchers frequently aim to identify individual subsets (clusters) of objects within the dataset. With high-dimensional data (>3 dimensions), the data become progressively more sparsly distributed in space. This, in turn, progressively increases the influence of the “curse of dimensionality” and thereby makes subset identification progressively more difficult. To address this issue, we developed a pipeline of fully-automated sequential analysis methods that together provide statistically robust cluster idenfication and cluster matching tools for data generated with flow/mass cytometry and/or other technologies.

Applications

- We foresee that the methods we describe here with respect to flow cytometry data will be applicable to data generated by a wide variety of commercial and academic technologies including, for example, demographic data, retail marketing data, census data, etc.

Advantages

- The traditional approach to locating clusters (subsets) in high-dimensional (Hi-D) datasets such as those acquired by flow cytometry is to reduce their dimensionality, usually by linear and/or nonlinear one/two dimensional mapping or projection strategies. This Projection Pursuit approach has been known since

1974 to be a very efficient method of analyzing high dimensional data in a way that avoids the curse of dimensionality '1". Indeed, much of what is known about stem cells, blood cells and diseases such as leukemia and AIDS relies on flow cytometry data analyzed with manual Projection Pursuit methods (i.e., using "manual gating" such as that offered by FlowJo and other flow data analysis packages). However, the resolution of such subsets is by no means routine with the available manual analysis tools. Further, because such analysis methods ultimately rely on user skills to manually define subset boundaries and other properties, subset identification and quantitation is still more appropriately recognized as art rather than science. Automating this data analysis process and making it more objective is clearly desirable. In fact, several groups have recently developed intensive computational approaches aimed at simultaneously identifying the subsets (clusters) within a given Hi-D dataset '2". These attempts at Hi-D clustering methods are well motivated from a biomedical and user functionality point of view. However, they are perforce highly sensitive to compromise by what statisticians refer to as "the curse of dimensionality", a well-known statistical problem that compromises both statistical validity and computational performance of Hi-D clustering methods 'Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning. (Springer-Verlag, 2009)". Indeed, as we have shown '3", the curse of dimensionality clearly militates against the use of simultaneous Hi-D clustering methods for flow/mass cytometry and other high-dimensional data. Nevertheless, biomedical and other disciplines that rest on cluster analysis of multiparameter data require a solution to this problem. To develop objective, statistically valid and computationally efficient cluster identification methods, we started with the basic ideas underlying previous automated two-dimensional (2D) Exhaustive Projection Pursuit approaches '1". Basically, in such approaches 1) Hi-D data is presented as a collection of 2D linear projections; 2) every 2D projection is then characterized by a numerical index that indicates the amount of structure that is present '4"; 3) this index is then used as the basis for a heuristic search to locate the most "useful" 2D projection; 4) once the projection with the most useful structure has been found, this structure is then segmented and each portion is recursively analyzed until there is no remaining structure detectable. In general, these Projection Pursuit methods are a big step forward towards solving the problem of Hi-D data analysis since they avoid the curse of dimensionality. However, the approaches advanced thus far have some key limitations, e.g., it is neither obvious nor trivial to

specify what constitutes structures in data and how to make inferences from such identified structures. To overcome these limitations, we have developed unsupervised Exhaustive Projection Pursuit (EPP) methods that use the smallest misclassification error across a decision boundary between identified clusters (using DBM method '5") as an index to identify the most profitable 2D projection. To facilitate inference from recursive EPP outcomes, we introduced multidimensional cluster matching and display methods that together with the EPP provide rapid unsupervised cluster (subset) recognition, display and characterization. These EPP analysis methods, which we describe here, provide statistically robust clustering tools for flow/mass cytometry and are readily applicable to similar types of single- or multi-dimensional data generated by other technologies.

Innovators

- Leonore Herzenberg
- Darya Orlova
- Guenther Walther
- Stephen Meehan
- Wayne Moore
- David Parks

Licensing Contact

Imelda Oropeza

Senior Licensing Manager, Physical Sciences

Email