Comparative Analysis of Linear and Nonlinear Dimension Reduction Techniques on Mass Cytometry Data

Nathan Jekel1*, Emily Vidal2*, Anna Konstorum3, Reinhard Laubenbacher3
1Penn State Harrisburg, 2Angelo State University, 3Center for Quantitative Medicine, UCONN Health
*These authors contributed equally to this work

Introduction

- Mass cytometry is a newly developed technology for quantification and classification of immune cells that can analyze up to 100 markers per cell. High dimensional data resulting from these experiments require innovative methods for analysis and visualization.

- Mathematical dimension reduction techniques map data from the input space to a lower dimensional subspace. Each technique aims to preserve a specific characteristic of the data during this process.

Objectives

1. Implement 4 dimension reduction techniques on a benchmark manually gated (divided into cell subtypes) mass cytometry data set
2. Compare techniques using 3 metrics: Computation Time, Neighborhood Proportion Error, and Residual Variance
3. Apply best techniques to non-gated mass cytometry data

Methods

<table>
<thead>
<tr>
<th>METHOD</th>
<th>TYPE</th>
<th>PRESERVED CHARACTERISTIC</th>
<th>COST FUNCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isomap</td>
<td>Nonlinear</td>
<td>Geometric structure of data</td>
<td>$| \tau(D_{ij}) - \tau(D_{ij}) |_2$</td>
</tr>
<tr>
<td>t-SNE</td>
<td>Nonlinear</td>
<td>Gaussian based similarity measure</td>
<td>$KL(P</td>
</tr>
<tr>
<td>Diffusion Maps</td>
<td>Nonlinear</td>
<td>Distance based on Gaussian random walk</td>
<td>$\sum_{i,j} (D_{ij}(x_i, x_j) - | y_i - y_j |^2)$</td>
</tr>
<tr>
<td>PCA</td>
<td>Linear</td>
<td>Variability</td>
<td>$\text{argmax } \text{var}()_{i=1}$</td>
</tr>
</tbody>
</table>

Comparison Metrics

1. Computation Time
2. Neighborhood Proportion Error (NPE):
   - Find the $k$ nearest neighbors for each data point and count the number of neighbors that are of the same cell subtype
   - Convert the counts for each subtype into a probability distribution
   - Calculate Total Variation between the high and low dimensional distributions
   - Sum the Total Variation for each subtype to get NPE
   - Convert the counts for each subtype into a probability distribution

3. Residual Variance:
   Residual variance is defined as
   $E = \sum_{i} \delta (P_i, Q_i)$
   where $r$ is the Pearson correlation coefficient within a distance of the original space ($D_{ij}$) and a distance matrix in the dimension reduced space ($D'_{ij}$).

Results (Benchmark Data)

- All dimension reductions were performed using the statistical computing software R [2]. The benchmark data set consists of 100,044 cells from healthy human bone marrow. Its purpose is to measure healthy human hematopoiesis [1].

Conclusion I (Benchmark Data)

- The acute response to toxin (ART) dataset consists of 66,662 control and toxin-stimulated immune cells. Its purpose is to examine the underlying molecular basis of TCRγδ cell activation. Upon infection, a special type of T cell, TCRγδ, is activated immediately, subseqently secreting cytokines that activate other immune cells.

- The cells in the control and experimental data occupy different regions of the low dimensional space, indicating that dimension reduction preserves differences in experimental conditions.

- High TCRγδ cells are projected to the edge of the dimension reduced space, which indicates that they represent a phenotype that is distinct from the other cell types.

Acknowledgements

Funded by National Science Foundation award #1460967. Special thanks to Dr. Miranda Lynch of the UCONN Center for Quantitative Medicine for statistical consultation and Dr. Antoine Menoret of the UCONN Health Immunology Department for providing the ART data.

References