Gaometric Data Analysis and Machine Learning for Prognostics

Michael Kirby
Department of Mathematics
Department of Computer Science
Colorado State University
Fort Collins, Colorado

May 5, 2017
Collaborators:

- Tomojit Ghosh (CS), Shannon Stiverson (Math), Daniel Jonas, Lara Kassab (Math), Kartikay Swarma (CS), Nate Mankovich (Math), Ariel Liu (Math/CS).
- Faculty and Postdocs CSU: Manucher Aminian, Henry Adams, Chris Peterson, (Math), John Belisle, Richard Bowen, Ric Slayden (MIP)
- Helene Andrews-Polymenis, David Threadgill (Texas A&M University), Angie Rasmussen, (Columbia), Corey O’Hern (Yale), Mark Shattuck (CCNY), Greg Huber (Chan-Zuckerberg Biohub),
- New Math Postdocs CSU: Amy Peterson, Eric Kehoe, Duy Hoang Thai
Philosophy of Geometric Data Analysis

- Understanding the geometric structure of data is critical for algorithm design, e.g., low D illumination cones.

- Mathematical theory provides insight into data reduction, e.g., Whitney’s theorem, Takens’ theorem.

- Manifolds provide a useful data model that is useful for capturing nonlinear variations.

- The geometric framework of Grassmannians, Stiefels and flags provide robust approaches to capture structure in large volumes of data.
The dots are class of Meadows and the circles are class of Metal Sheets of the Pavia University data set.
Each point is 10 pixels of an HSI encoded on a Grassmannian and embedded in Euclidean space using the smallest angle pseudo-metric.
Biological Data Analysis
Iterative Feature Removal (IFR)

- Repeatedly build a sparse linear classifier.\(^1\)
- At each iteration, remove all features selected by previous iterations.
- Model must be built without the highly discriminative features of previous steps.
- Has the effect of uncovering the non-obvious features that may only be discriminative in a multivariate sense.

---

\(^1\)A sparse model assigns zero-weights to most features
Zero weights

200 Largest Weights from an SSVM Model

Feature Index Sorted by Weight Magnitude

Weight Magnitude
The cumulative prognosis time distribution for the selected signaling pathways.
**Table:** Pathways with uncorrelated errors can be combined to yield a classifier with superior accuracy.

<table>
<thead>
<tr>
<th>Acc</th>
<th>Pathway Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.0</td>
<td>B Cell Maturation and Activation + Cell Adhesion Molecules</td>
</tr>
<tr>
<td>98.2</td>
<td>Antigen Recognition Genes + Cell Adhesion Molecules</td>
</tr>
<tr>
<td>96.5</td>
<td>IL-1 Beta Receptor Family + Cell Adhesion Molecules</td>
</tr>
<tr>
<td>96.5</td>
<td>B Cell Maturation and Activation + Complement Pathway</td>
</tr>
<tr>
<td>94.7</td>
<td>IL-1 Beta Receptor Family + B Cell Maturation and Activation</td>
</tr>
<tr>
<td>94.7</td>
<td>Cell Cycle Related + Complement Pathway</td>
</tr>
<tr>
<td>94.7</td>
<td>Cell Cycle Related + Cell Adhesion Molecules</td>
</tr>
<tr>
<td>94.7</td>
<td>Antigen Recognition Genes + B Cell Maturation and Activation</td>
</tr>
<tr>
<td>93.0</td>
<td>IL-1 Beta Receptor Family + Complement Pathway</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
predict contagion within 24 hours of exposure to respiratory infection.

determine time elapsed from exposure to pathogen
A Motivational Example

PCA trajectories pathway404
Control vs Shedders Microarray Analysis (GSE73072)

**Experiment:** Classifying subjects as Controls vs Shedders in bins of 8 hours in the first 36 hours With Smaller Feature Set

<table>
<thead>
<tr>
<th>Bins</th>
<th># Features</th>
<th>SSVM</th>
<th>CE</th>
<th>ANN</th>
<th># Features</th>
<th>SSVM</th>
<th>CE</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_8</td>
<td>25</td>
<td>95.14 +/- 00.00</td>
<td>98.20 +/- 00.23</td>
<td>93.27 +/- 01.97</td>
<td>50</td>
<td>95.49 +/- 0.0</td>
<td>98.63 +/- 0.003</td>
<td>98.14 +/- 0.01</td>
</tr>
<tr>
<td>9_16</td>
<td>35</td>
<td>97.24 +/- 00.00</td>
<td>98.34 +/- 00.40</td>
<td>95.67 +/- 01.39</td>
<td>160</td>
<td>95.8 +/- 0</td>
<td>99.09 +/- 0.004</td>
<td>98.84 +/- 0.014</td>
</tr>
<tr>
<td>17_24</td>
<td>40</td>
<td>86.86 +/- 00.00</td>
<td>94.40 +/- 00.40</td>
<td>86.45 +/- 01.03</td>
<td>80</td>
<td>89.92 +/- 0</td>
<td>98.37 +/- 0.006</td>
<td>95.42 +/- 0.022</td>
</tr>
<tr>
<td>25_32</td>
<td>30</td>
<td>97.11 +/- 00.00</td>
<td>96.02 +/- 00.23</td>
<td>91.32 +/- 02.70</td>
<td>30</td>
<td>97.46 +/- 3.33</td>
<td>98.60 +/- 0.005</td>
<td>96.81 +/- 0.012</td>
</tr>
</tbody>
</table>
Preliminary Time of Exposure (ToE) Results - 11/17

Hour 5, 12, 21, 29

Preliminary Result
Results: 9 class confusion matrix

**Rows:** True label, ordered by time: 1-8, 9-16, 17-24, 25-32, 33-40, 41-48, 49-72, 73-96 and 97-120

**Columns:** Predicted label

Simple BSR: 68.8% (bin accuracy)

"First two days" vs "days 3-5" BSR: 90.94%:
Visualization: Centroid Encoder
Objective: Determine *mechanisms of tolerance*. 
Collaborative Cross Strains
Manifold Learning

2-D Embeddings of One-Day Windows of Temperature Time Series

Window starts 3.0 days pre infection

Window starts at infection

Window starts 2.0 days post infection
Takens’ theorem motivated time-delay embedding and nonlinear fitting

\[ x_{n+T} = f(x_n, x_{n-T}, x_{n-2T}, \ldots) \]
Telemetry Anomaly Detection

Time

Temperature

# of iterations
Mathematics and Data: *Hassler Whitney* (1907-89)
Acknowledgments

This presentation is based on research partially supported by the National Science Foundation under Grant DMS-1322508 and DARPA Award N66001-17-2-4020. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or DARPA.

Thank you!

Michael Kirby
Department of Mathematics
Colorado State University
Michael.Kirby@Colostate.Edu