# Analysis of LIBS amino acid spectra and the impact of neighborhood size on the efficacy of nonlinear analysis

Piyush Kumar Sharma[1], Gary Holness[2], and Poopalasingam Sivakumar, Yuri Markushin, Noureddine Melikechi[3]

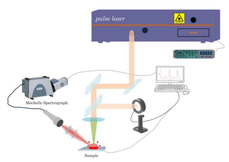
[1]Department of Mathematical Sciences, [2]Department of Computer and Information Sciences, [3]Department of Physics and Engineering

Delaware State University, Dover, DE 19901

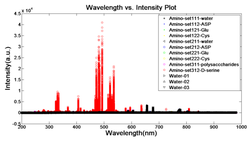
## Abstract

In recent work we explored the impact of LIBS spectra on suitable neighborhood size over which to consider pattern phenomena, if nonlinear methods capture pattern phenomena with increased efficacy, and how they improve classification and identification of compounds. We analyzed four amino acids, polysaccharide, and a control group, water. We developed an information theoretic method for measurement of Laser Induced Breakdown Spectroscopy (LIBS) energy spectra, implemented manifold methods for nonlinear dimensionality reduction, and found while clustering results were not statistically significantly different, nonlinear methods lead to increased classification accuracy. The goal of the poster is to discuss recently published results.

## Introduction




LIBS Experimental Setup.




High intensity peaks of D-Serine (red) at certain λ regions most likely due to micro-well influence.

LIBS is a powerful analytical technique that provides information about the elemental composition of a given sample. The method uses an intense short laser pulse to break down the matrix of the target and to create a short-lived micro-plasma. During the cooling of plasma, atomic, ionic, and occasionally molecular constituents emit the discrete spectra which is further collected and analyzed, allowing characterization of the sample content.

## Methods

For a given $D$ dimensional dataset $X$ find low dimensional, $d(\ll \mathrm{D})$, embedding $Y$ by algorithm(s):

### PCA & cMDS

PCA & cMDS are linear methods that employ pairwise comparisons. PCA uses covariance while cMDS uses pairwise distances ( dissimilarities). Both yield similar results when Euclidean distance is used.

### ISOMAP

- Find $K$ nearest nbhd for each point $x_i \in X$.
- Compute a nbhd graph $G$ with edge length as Euclidean distance $d_X(i,j)$ for neighbors $i$ & $j$.
- Compute geodesic distance $d_G(i,j)$ between all pairs using Dijkstra or Floyd-Warshall algorithm.
- Find embedding $Y$ by applying cMDS to $d_G$.

### LLE

- Find $K$ nearest nbhd for each point $x_i \in X$.
- Compute a set of weights $W$ for each point that best describe the point as a linear combination of its neighbors. Solve for reconstruction weights $W$.
- Find embedding $Y$ by using an eigenvector based optimization technique which minimizes reconstruction loss.

### Clustering

Davies-Bouldin Index = $\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\{D_{ij}\}$, where $D_{ij}=\frac{\overline{d_i}+\overline{d_j}}{d_{ij}}$ $\overline{d_i}$ is the within-to-between cluster distance ratio for the $i^{th}$ and $j^{th}$ clusters, $\overline{d_i}$ is the average distance between each point in the $i^{th}$ cluster and the centroid of the $i^{th}$ cluster, $\overline{d_j}$ is the average distance between each point in the $i^{th}$ cluster and the centroid of the $j^{th}$ cluster. $d_{ij}$ is the Euclidean distance between the centroids of the $i^{th}$ and $j^{th}$ clusters.

### Classification

In support vector machine (SVM), decision surface separating the classes is a hyperplane $w^T x + b = 0$ where $w$ is a weight vector, $x$ is an input vector, and $b$ is a bias. The vectors that define the hyperplane are the support vectors. The problem can be expressed in dual formulation with $w = \sum_{i=1}^{m}\alpha_i\varphi(x_i)$ as:

$$min_\alpha D(\alpha) = \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j\varphi(x_i)\varphi(x_j) - \sum_i y_i\,\alpha_i$$

such that $\sum_i \alpha_i = 0,\ \ 0 \le y_i\alpha_i \le C$

with dual decision function: $\mathrm{f(x)} = K(x_i, x) + \mathrm{b}$
where $K(x_i, x) = \varphi(x_i).\varphi(x)$ is the kernel function, $m$ is the number of support vectors, and $\alpha$ Lagrangian multipliers.

## Results

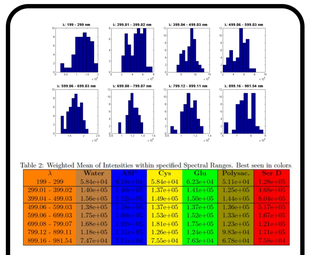
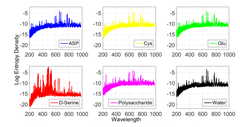


Table 2: Weighted Mean of Intensities within specified Spectral Ranges. Best seen in colors

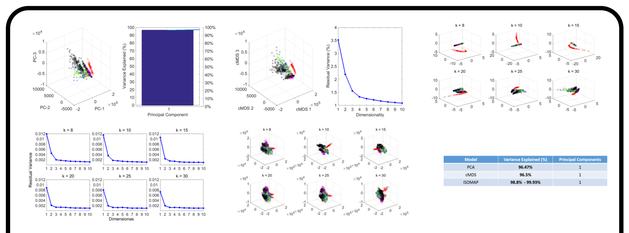
| | Water | Asp | Cys | Glu | Polysacch | Ser |
|---|---|---|---|---|---|---|
| 199 - 299 | 5.84e+04 | 6.23e+04 | 5.84e+04 | 6.23e+04 | 5.11e+04 | 7.29e+05 |
| 299.01 - 399.02 | 1.40e+05 | 1.32e+05 | 1.40e+05 | 1.32e+05 | 1.28e+05 | 4.68e+05 |
| 399.04 - 499.03 | 1.56e+05 | 1.49e+05 | 1.49e+05 | 1.56e+05 | 1.44e+05 | 8.66e+05 |
| 499.06 - 599.03 | 1.38e+05 | 1.35e+05 | 1.37e+05 | 1.35e+05 | 1.38e+05 | 5.37e+05 |
| 599.06 - 699.03 | 1.75e+05 | 1.53e+05 | 1.53e+05 | 1.53e+05 | 1.85e+05 | 1.85e+05 |
| 699.08 - 799.07 | 1.68e+05 | 1.62e+05 | 1.81e+05 | 1.75e+05 | 1.23e+05 | 1.21e+05 |
| 799.12 - 899.11 | 1.18e+05 | 1.26e+05 | 1.26e+05 | 1.24e+05 | 9.83e+04 | 1.11e+05 |
| 899.16 - 981.54 | 7.47e+04 | 7.55e+04 | 7.55e+04 | 7.63e+04 | 6.78e+04 | 7.56e+04 |

Expected Total Intensity (on x-axis) w.r.t. Spectral Ranges for D-Serine.

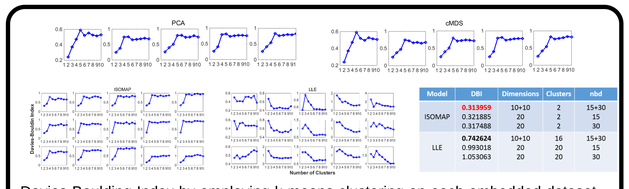
Shannon Entropy of a discrete random variable $X$ is:
$H(X) = -\sum_{i=1}^{N} p(x_i)\log p(x_i) \geq 0$
such that $\lim_{p\to 0^+} p\log p \to 0$.



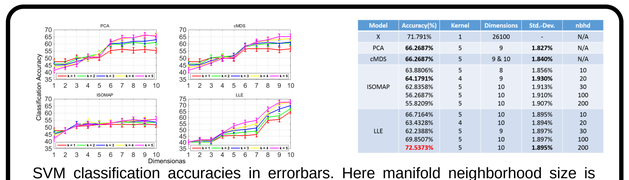

log Entropy Density Plots for each amino acid. Highest peaks appear for D-Serine (red) ranging up to −2.5.




PCA, cMDS, ISOMAP, and LLE 3D Scatter and Scree Plots. Residual variance decreasing for ISOMAP with neighborhood size increasing while the elbow of the curve unaltered at dimension $d = 3$. Notice D-Serine's (red) variation for $k$.




Davies-Boulding Index by employing k-means clustering on each embedded dataset. From left to right, plots correspond to 2, 3, 5, 7, & 10 dimensions respectively. The 1st, 2nd, & 3rd rows of ISOMAP & LLE plots correspond to k = 8, 15, & 30.




SVM classification accuracies in errorbars. Here manifold neighborhood size is 200. LLE achieves the highest accuracy for polynomial kernel with degree 5 (magenta) suggesting the intrinsic nonlinearity associated to LIBS dataset.

## Conclusion

1. LIBS pattern phenomena are nonlinear.
2. k-neighborhood size:
   a. for inter sample variabilities is [8; 15].
   b. for intra sample variabilities is $k > 15$.
3. Identified micro-wells influence on D-Serine spectra.
4. Future work will comprise remaining amino acids.

## Acknowledgement

This work was supported by NSF Award# HRD-1242067, Award# CNS-1205426, and Propinquo LLC

## Bibliography

[1] S. J. Rehse, H. Salimnia and A. W. Miziolek. Laser-induced breakdown spectroscopy (LIBS): an overview of recent progress and future potential for biomedical applications, Journal of Medical Engineering and Technology, 36, 77–89, 2012.

[2] You Z-H, Lei Y-K, Gui J, Huang D-S, Zhou X. Using Manifold Embedding for Assessing and Predicting Protein Interactions from High-Throughput Experimental Data. Bioinformatics. 2010; 26(21):2744–2751. PMC. Web. 11 Aug. 2015.

[3] Joshua B. Tenenbaum, Vin de Silva, John C Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science Vol 290, 2000.

[4] Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science Vol 290, 2323, 2000.

[5] Wendy L. Martinez, Angel R. Martinez, Jeffrey L. Solka. Exploratory Data Analysis with MATLAB. Chapman and Hall/CRC Computer Science and Data Analysis, 2nd Edition, 2010.