

# Biclustering EEG Data from Epileptic Patients Treated with Vagus Nerve Stimulation

Nikita Boyko, Stanislav Busygin, Michael Bewernitz,  
Panos Pardalos

Center for Applied Optimization  
Department of Industrial & Systems Engineering  
University of Florida

Data Mining, Systems Analysis and Optimization in  
Biomedicine



# Purpose of Studies

- ▶ Our ultimate goal is to develop a physiologic marker for optimal Vagus Nerve Stimulation (VNS) parameters based on scalp EEG signals
- ▶ We performed first step in this direction:
  - ▶ two different parameter sets are analyzed for two patients
- ▶ We study how VNS simulation affects EEG
  - ▶ EEG  $\rightarrow$  Lyapunov Exponents  $\rightarrow$  Biclustering Separation
- ▶ Interpretation of Results

# Epilepsy

- ▶ A symptom of a brain disorder distinguished by recurring seizures.
- ▶ Can begin at any age.
- ▶ Affects 1% of the population. World Health Organization estimates 50 million cases worldwide.
- ▶ Quality of Life: Affects self-esteem, career, social opportunities, restricted driving privileges.

# Epilepsy Treatment

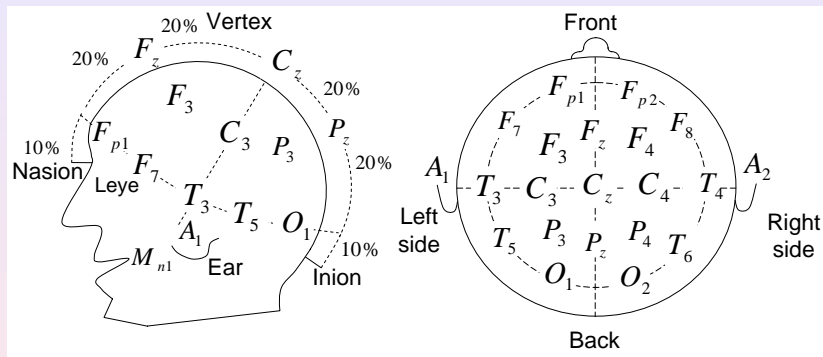
- ▶ Treatment: Drugs, Surgery, Electrical and Magnetic Stimulation.
- ▶ Vagus Nerve Stimulation
  - ▶ Electric stimulator implanted subcutaneously in the chest
  - ▶ Connected, via subcutaneous electrical wires, to the cervical left vagus nerve.

# Vagus Nerve Stimulation parameters

- ▶ The VNS is programmed to deliver electrical stimulation at a set intensity, duration, pulse width, and frequency.
- ▶ Optimal parameters are presently determined on a case by case basis, depending on clinical efficacy (seizure frequency) and tolerability.
- ▶ Such parameter adjustment is time consuming and costly.
- ▶ There is a need to develop a reliable, objective and rapid method of determining the optimal stimulation parameters for each patient

# EEG Data

- ▶ General Clinical Research Center in Shands Hospital at The University of Florida.
- ▶ Two patients A and B.
- ▶ 25 scalp-EEG channels
- ▶ Sampling rate 512 Hz



**Figure:** Montage for scalp electrode placement (10-20)

# VNS Stimulation Settings

Parameter	Patient A	Patient B
Recorded time	$\approx 24$ hours	
Signal duration	30 sec	
Rest duration	5 min	
Pulse width	500 $\mu$ sec	250 $\mu$ sec
Output current	1.75 mA	1.5 mA
Frequency	30 Hz	20 Hz
# of VNS cycles	255	237

# Manually Activated Stimulation

- ▶ Stimulation may be activated manually, for example in case of seizure.
- ▶ During EEG recording session
  - ▶ Patient A did not undergo seizures.
  - ▶ Patient B experienced 14 seizures (manual stimulation was activated 14 times)
- ▶ Parameters for patient B's manual simulation
  - ▶ stimulation activated after 19-37 sec after seizure, output current 1.75 mA, signal frequency 20 Hz, pulse width 500  $\mu$ sec, duration 60 sec.
- ▶ Manual stimulation is not included

# Lyapunov Exponents: what is it and why?

- ▶ Important measure that characterizes chaotic behavior of nonlinear system.
- ▶ Global Lyapunov Exponent: how fast nearby orbits of the system converge or diverge in infinitely large time interval.
- ▶ Local Lyapunov exponent characterize local predictability around a point  $x_0$  in phase space
- ▶ Lyapunov Exponent has proven its efficiency in EEG analysis for predicting epileptic seizures

## Formal Definition

Let a system be set by

$$\dot{X}(t) = F(X), \text{ where } X : \mathbb{R} \mapsto \mathbb{R}^n, F : \mathbb{R}^n \mapsto \mathbb{R}^n$$

The maximal Lyapunov Exponent  $\lambda$  can be defined as follows:

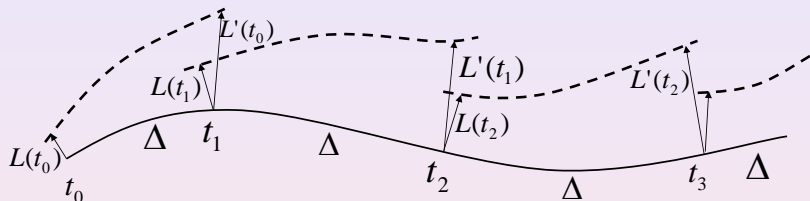
$$\lambda = \lim_{t \rightarrow \infty} \lim_{\delta X(0) \rightarrow 0} \frac{1}{t} \log_2 \frac{\delta X(t)}{\delta X(0)}$$

For short term maximal Lyapunov Exponent (STLmax) we can take “reasonable”  $t$  instead of external limit.

# Estimation from Time Series

- ▶ In real life we often deal with one dimensional time series of noisy data (such as EEG signal) instead of explicit system of equations
- ▶ Wolf *et al* suggested algorithm for Lyapunov Exponent calculation from time series
  - ▶ A Wolf, J B Swift, H L Swinney, J A Vastano, Determining Lyapunov Exponents from a Time Series, Physica 16D (1985), pp. 285 - 317.
- ▶ We used Sackellares *et al* modification of Wolf's algorithm for STLmax calculation that handles noisy non-stationary data
  - ▶ L D Iasemidis, J C Principe, J C Sackellares, Measurement and Quantification of Spatiotemporal Dynamics of Human Epileptic Seizures. Nonlinear Signal Processing in Medicine, Ed. M. Akay, IEEE Press, 1999

# Approach to Estimation



**Figure:** Evolution in phase space and replacement procedure used to estimate Lyapunov Exponents from experimental data

$$STL_{\max} = \frac{1}{t_M - t_0} \sum_{k=1}^M \log_2 \frac{L'(t_k)}{L(t_{k-1})}.$$

# Algorithm set up

- ▶ *Input*: EEG signal recorded with 512Hz ( $\Delta t = 1.95\text{msec}$ )
- ▶ *Output*: STLmax series computed for every 4sec window of the source data
- ▶ *Algorithm parameters*: reconstructed dimension  $p = 7$ , lag step  $\tau = 7$  (14 msec), evolution time  $\Delta = 21$  (41 msec)
- ▶ 25 channels for patients A and B are processed

# Biclustering

- ▶ Biclustering is a methodology allowing for simultaneous classification of samples and features (may be supervised or unsupervised).
- ▶ It finds clusters of samples possessing similar characteristics while at the same time revealing features responsible for creating these similarities.
- ▶ The required consistency of sample and feature classification gives biclustering an advantage over other methodologies treating samples and features of a dataset separately of each other.

# Biclustering

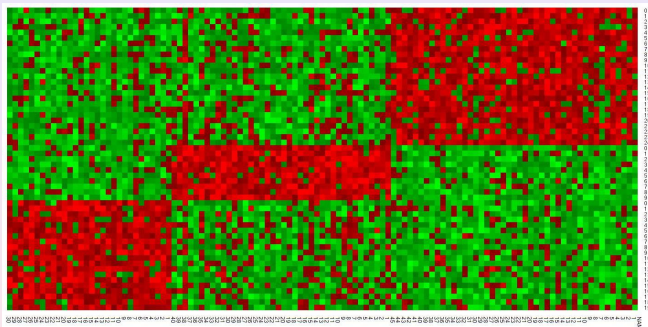




Figure: Partitioning of samples and features into 3 classes.

# Biclustering: Applications

- ▶ Biological and Medical:
  - ▶ Microarray data analysis
  - ▶ Analysis of drug activity, Liu and Wang (2003)
  - ▶ Analysis of nutritional data, Lazzeroni et al. (2000)
- ▶ Text Mining: Dhillon (2001, 2003)
- ▶ Marketing: Gaul and Schader (1996)
- ▶ Dimensionality Reduction in Databases: Agrawal et al. (1998)
- ▶ Others:
  - ▶ electoral data - Hartigan (1972)
  - ▶ currency exchange - Lazzeroni et al. (2000)

# Biclustering Surveys

-  A. Tanay, R. Sharan, R. Shamir, Biclustering Algorithms: A Survey, to appear in the Handbook of Bioinformatics, 2004.
-  S. Busygin, O.A. Prokopyev, and P.M. Pardalos, Biclustering in Data Mining, to appear in C&OR, 2007.

# Definitions

- ▶ Data set of  $n$  samples and  $m$  features is a matrix

$$A = (a_{ij})_{m \times n},$$

where the value  $a_{ij}$  is the expression of  $i$ -th feature in  $j$ -th sample.

- ▶ A *biclustering* of a data set is a collection of pairs of sample and feature subsets

$$\mathcal{B} = ((\mathcal{S}_1, \mathcal{F}_1), (\mathcal{S}_2, \mathcal{F}_2), \dots, (\mathcal{S}_r, \mathcal{F}_r)),$$

where  $r$  is the number of classes. Such that the collection  $(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_r)$  forms a partition of the set of samples, and the collection  $(\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_r)$  forms a partition of the set of features.

## Intuition Behind Biclustering

- ▶ Let us distribute features among the classes of training set such that each feature belongs to the class where its average expression among the training samples is highest.

$$(\mathcal{S}_1^0, \mathcal{S}_2^0, \dots, \mathcal{S}_r^0) \rightarrow (\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_r)$$

- ▶ Now, if we transpose the matrix, take the feature classification as given, and re-classify the training samples according to highest average expression values in feature classes.

$$(\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_r) \rightarrow (\mathcal{S}_1^1, \mathcal{S}_2^1, \dots, \mathcal{S}_r^1)$$

# Intuition Behind Biclustering

- ▶ Will we obtain the same training set classification?

$$? (\mathcal{S}_1^0, \mathcal{S}_2^0, \dots, \mathcal{S}_r^0) = (\mathcal{S}_1^1, \mathcal{S}_2^1, \dots, \mathcal{S}_r^1) ?$$

- ▶ If yes, we will say that we obtained a **consistent biclustering**.

# Consistent Biclustering

- ▶ Let each sample be already assigned somehow to one of the classes  $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_r$ . Introduce a 0–1 matrix  $S = (s_{jk})_{n \times r}$  such that  $s_{jk} = 1$  if  $j \in \mathcal{S}_k$ , and  $s_{jk} = 0$  otherwise.
- ▶ The sample class *centroids* can be computed as the matrix  $C = (c_{ik})_{m \times r}$ :

$$C = AS(S^T S)^{-1}, \quad \left( c_{ik} = \frac{\sum_{j \in \mathcal{S}_k} a_{ij}}{|\mathcal{S}_k|} \right)$$

whose  $k$ -th column represents the centroid of the class  $\mathcal{S}_k$ .

# Consistent Biclustering

- ▶ Consider a row  $i$  of the matrix  $C$ . Each value in it gives us the average expression of the  $i$ -th feature in one of the sample classes. As we want to identify the checkerboard pattern in the data, we have to assign the feature to the class where it is most expressed. So, let us classify the  $i$ -th feature to the class  $\hat{k}$  with the maximal value  $c_{i\hat{k}}$ :

$$i \in \mathcal{F}_{\hat{k}} \Rightarrow \forall k = 1 \dots r, k \neq \hat{k} : c_{i\hat{k}} > c_{ik}$$

## Consistent Biclustering

- ▶ Using the classification of all features into classes  $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_r$ , let us construct a classification of samples using the same principle of maximal average expression. We construct a 0–1 matrix  $F = (f_{ik})_{m \times r}$  such that  $f_{ik} = 1$  if  $i \in \mathcal{F}_k$  and  $f_{ik} = 0$  otherwise. Then, the feature class centroids can be computed in form of matrix  $D = (d_{jk})_{n \times r}$ :

$$D = A^T F (F^T F)^{-1}, \quad \left( d_{jk} = \frac{\sum_{i \in \mathcal{F}_k} a_{ij}}{|\mathcal{F}_k|} \right)$$

whose  $k$ -th column represents the centroid of the class  $\mathcal{F}_k$ .

# Consistent Biclustering

- ▶ The condition on sample classification we need to verify is

$$j \in \mathcal{S}_{\hat{k}} \Rightarrow \forall k = 1 \dots r, k \neq \hat{k} : d_{j\hat{k}} > d_{jk}$$

# Consistent Biclustering

## Definition

A biclustering  $\mathcal{B}$  will be called **consistent** if the following relations hold:

$$i \in \mathcal{F}_{\hat{k}} \Rightarrow \forall k = 1 \dots r, k \neq \hat{k} : c_{i\hat{k}} > c_{ik}$$

$$j \in \mathcal{S}_{\hat{k}} \Rightarrow \forall k = 1 \dots r, k \neq \hat{k} : d_{j\hat{k}} > d_{jk}$$

# Consistent Biclustering

## Definition

A data set is **biclustering-admitting** if some consistent biclustering for it exists.

## Definition

The data set will be called **conditionally biclustering-admitting** with respect to a given (partial) classification of some samples and/or features if there exists a consistent biclustering preserving the given (partial) classification.

# Consistent Biclustering

- ▶ **A consistent biclustering implies separability of the classes by convex cones.**

## Theorem

*Let  $\mathcal{B}$  be a consistent biclustering. Then there exist convex cones  $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_r \subseteq \mathbb{R}^m$  such that all samples from  $\mathcal{S}_k$  belong to the cone  $\mathcal{P}_k$  and no other sample belongs to it,  $k = 1 \dots r$ . Similarly, there exist convex cones  $\mathcal{Q}_1, \mathcal{Q}_2, \dots, \mathcal{Q}_r \subseteq \mathbb{R}^n$  such that all features from  $\mathcal{F}_k$  belong to the cone  $\mathcal{Q}_k$  and no other feature belongs to it,  $k = 1 \dots r$ .*

# Separation by Cones

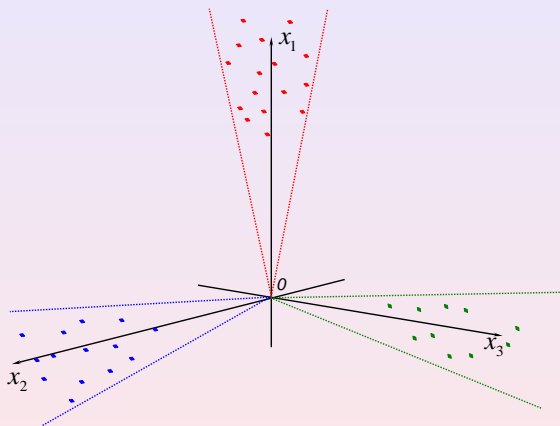


Figure: 3 classes are separated in 3D-space

# Supervised Biclustering

- ▶ In general, only a subset of features is relevant to the particular phenomenon of interest.
- ▶ We handle such a task utilizing the notion of consistent biclustering. *Namely, we select a subset of features of the original data set in such a way that the obtained subset of data becomes conditionally biclustering-admitting with respect to the given classification of training samples.*

# Fractional 0–1 Programming Formulation

- ▶ Formally, let us introduce a vector of 0–1 variables  $x = (x_i)$ ,  $i = 1 \dots m$  and consider the  $i$ -th feature selected if  $x_i = 1$ .
- ▶ The condition of biclustering consistency, when only the selected features are used, becomes

$$\frac{\sum_{i=1}^m a_{ij} f_{i\hat{k}} x_i}{\sum_{i=1}^m f_{i\hat{k}} x_i} > \frac{\sum_{i=1}^m a_{ij} f_{ik} x_i}{\sum_{i=1}^m f_{ik} x_i}, \forall j \in \mathcal{S}_{\hat{k}}, \hat{k}, k = 1 \dots r, \hat{k} \neq k.$$

# Fractional 0–1 Programming Formulation

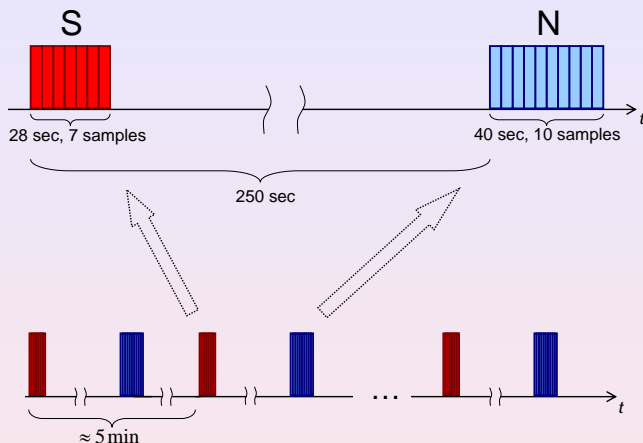
- ▶ We will use the fractional relations as constraints of an optimization task selecting the feature set. It may incorporate various objective functions over  $x$ , depending on the desirable properties of the selected features, but one general choice is **to select the maximal possible number of features in order to lose minimal amount of information provided by the training set**. In this case, the objective function is

$$\max \sum_{i=1}^m x_i$$

# References



S. Busygin, O.A. Prokopyev, P.M. Pardalos, Feature Selection for Consistent Biclustering via Fractional 0-1 Programming, Journal of Combinatorial Optimization, Vol. 10/1 (2005), pp. 7-21.



**Figure:** Building dataset for biclustering:  $STL_{max}$  points that are included for the analysis for each channel.

# Biclustering Experiment

- ▶ Positive (stimulation) class: Each 30 sec stimulation provided 7 data points (STLmax that each correspond to a four seconds window)
- ▶ Negative (non-stimulation) class: 10 consecutive Lyapunov Exponents 250 sec after stimulation
- ▶ We averaged corresponding data points across all stimulation cases
- ▶ Thus, we obtained a  $17 \times 25$  matrix. 17 samples (7 stimulation + 10 non-stimulation) and 25 features (channels)

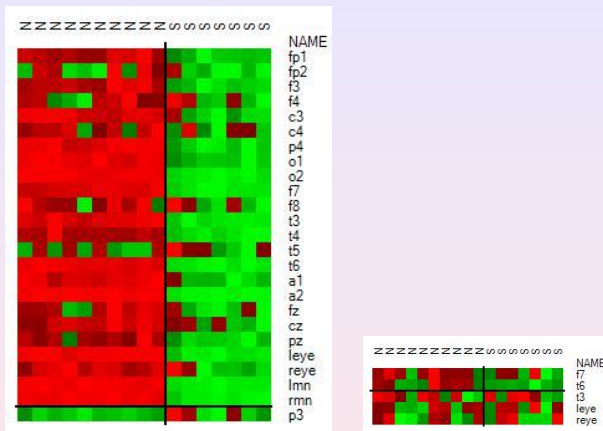


Figure: Heatmaps for patients A and B

# Feature selection

- ▶ Patient A
  - ▶ No features were excluded, i.e. patient's data were conditionally biclustering-admitting
  - ▶ STLmax data were consistently decreasing during the simulation except for channel  $p3$
  - ▶ All samples and all classes are confirmed by cross-validation
- ▶ Patient B
  - ▶ five features selected
  - ▶ The leave-one-out cross-validation was passed for all but four samples

# Further Plans

- ▶ Obtain data on more patients and
  - ▶ figure out influence seizures on separation quality
  - ▶ distinguish between stimulation parameters setting
- ▶ Utilize full spectrum of Lyapunov Exponents
- ▶ Apply other classification techniques
  - ▶ Logistic regression
  - ▶ ...
- ▶ Look into sleep disorders

# Thank you!

- ▶ Dr. Basim Uthman
  - ▶ Department of Neurology, University of Florida
- ▶ Dr. Georges Ghacibeh
  - ▶ Department of Neurology, University of Florida
- ▶ Dr. Oleg Prokopyev
  - ▶ Department of Industrial Engineering, University of Pittsburgh