Statistical Learning and Data Mining for Pediatric Obstructive Sleep Apnea (OSA)

Session on Women in Data Science, SIAM Mathematics of Data Science Virtual Conference

Emily T. Winn
Division of Applied Mathematics, Brown University

Website: www.emilytwinn.com
Twitter : @EmilyTWinn13

June 2020

BROWN

#BlackLivesMatter

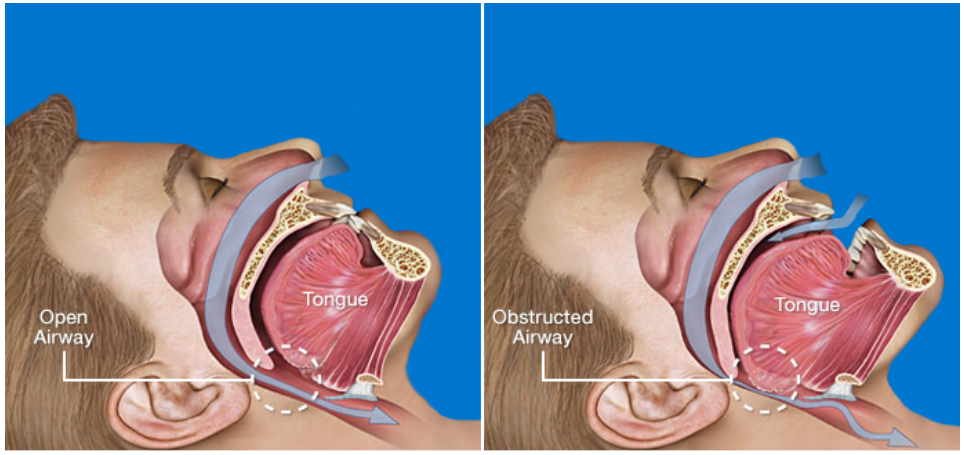
I acknowledge that I am on the traditional homelands of the Narragansett and Wampanoag peoples.

Outline of Talk

- Background on Pediatric Obstructive Sleep Apnea (OSA)
- Work with structured data (Winn et al, 2020)
- Work with unstructured data (Tymochko et al, 2020)
- Future Directions

Background

Obstructive Sleep Apnea (OSA)



Non-Obstructed Airway Obstructed Airway

- A form of sleep-disordered breathing characterized by recurrent episodes of partial or complete airway obstruction during sleep.
- OSA affects 1-5% of elementary school-age children.
- Health care utilization 240%.
- Behavioural issues, reflux, obesity, hypertension, cardiovascular dysfunction, and neurocognitive dysfunction.

Classes of OSA

One of most important measurements is Apnea-Hypopnea Index (AHI), which determines the severity of OSA.

- $\text{AHI} \leq 1$: None
- $1 \leq \text{AHI} \leq 5$: Mild
- $5 < \text{AHI} \leq 10$: Moderate
- $10 \leq \text{AHI}$: Severe

Challenge of Diagnosis

Diversity of symptoms in children makes diagnosis of OSA not so clear cut

- Surveys/Questionnaires
  - Pros: Fast, inexpensive
  - Cons: Missing Data, usually baseline
- Polysomnography (PSG)
  - Pros: Gold standard for OSA
  - Cons: Expensive, takes a lot of time, limited access
- Alternative diagnostic tools: biomarkers, genes, modelling airflow in upper-airway, airway shape, facial morphology

Research Objectives

- Use classical statistical and machine learning methods to demonstrate survey data can be effective in detecting hidden signals in complex data
- Use persistent homology and Markov chains to classify sleep states from polysomnography tests
- **Ultimate Goal**: Build algorithm to expedite and aid clinicians in diagnosing OSA

Structured Data

$ Raw data

- 200 subjects from two different clinics
- 172 variables from 6 surveys, 1 set of facial measurements
  - Child's Sleep habits
  - Quality of Life Survey
  - Pediatric Quality of Life Child Report
  - Pediatric Quality of Life Parent Report
  - Pediatric Sleep Survey
  - Health Screening
  - Craniofacial Index Measures (Complete data set)

Methods: Data Splitting

- **173 Subjects (67 Controls, 106 Patients)**
- **157 input variables**
- **Goal: Classify OSA vs No OSA** (patient vs control)
- Missing values imputed using MissForest from *MissingPy* package for Python
- Split the data set so 70% contained in training set, 30% in testing set.
- Each method trained/tested on 10 different splits to measure stability in addition to other success measures
- Python packages: *MissingPy*, *scikit-learn*
- R packages: *mass*, *randomForest*, *nnet*, *bnlearn*, *gRain*, *glnet*

Statistical/Machine Learning Methods Applied

- **Supervised Learning Methods**
  - Linear/Quadratic Discriminant Analysis, Logistic Regression, Decision Trees, Random Forests, Neural Networks, Support Vector Machines, supervised k-nearest neighbors
  - Used cross validation grid search to optimize parameters
- **Bayesian Classifiers**
  - Naïve Bayes, Tree augmented Bayesian classifier, Semi-Hierarchical Bayesian classifier
  - Did have to discretize some continuous random variables (eg time asleep)
- **Density Based Clustering** (no train/test split)
  - DBSCAN, Spectral, CkNN, Cut-Cluster-Classify, Threshold sample density

Results

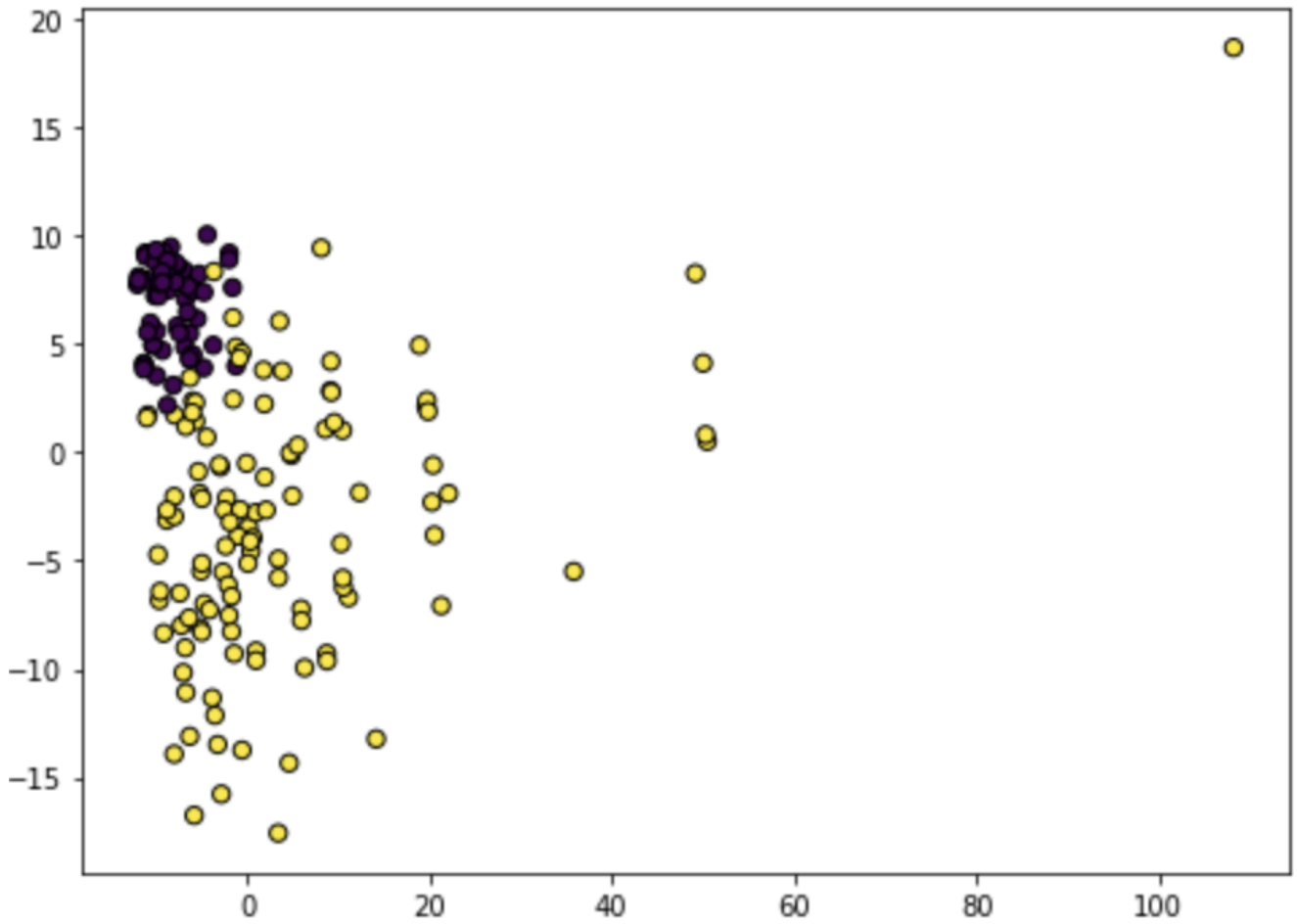


Figure: Kernel PCA projection of density based clustering with threshold sample density. Purple marks the controls, yellow marks the patients. (Winn et al, 2020)

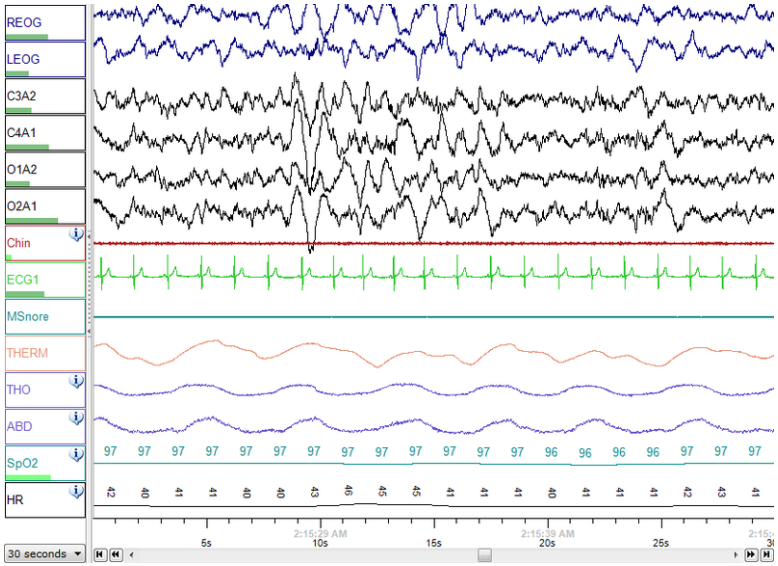
Results

Measure	Best Score	Method
Accuracy	0.78 ± 0.03	Random Forests,
	0.78035	DBC Threshold
Positive Predictive Value	0.89 ± 0.02	k-Nearest Neighbors
Negative Predictive Value	0.81 ± 0.06	Random Forests
Sensitivity	0.90 ± 0.06	Naïve Bayes
	0.87736	DBC Threshold
Specificity	0.85 ± 0.04	Random Forests

Table: Best methods by each unit of measure used in standard clinical literature (Winn et al, 2020)

Unstructured Data

$$$ Raw Polysomnography (PSG)



- 78 children recruited who had taken PSG
- Project approved by Health Research Ethics Board at the University of Alberta
- Analyze subset of 8 patients - 2 from each OSA class (none, mild, moderate, severe)

Sleep Stages

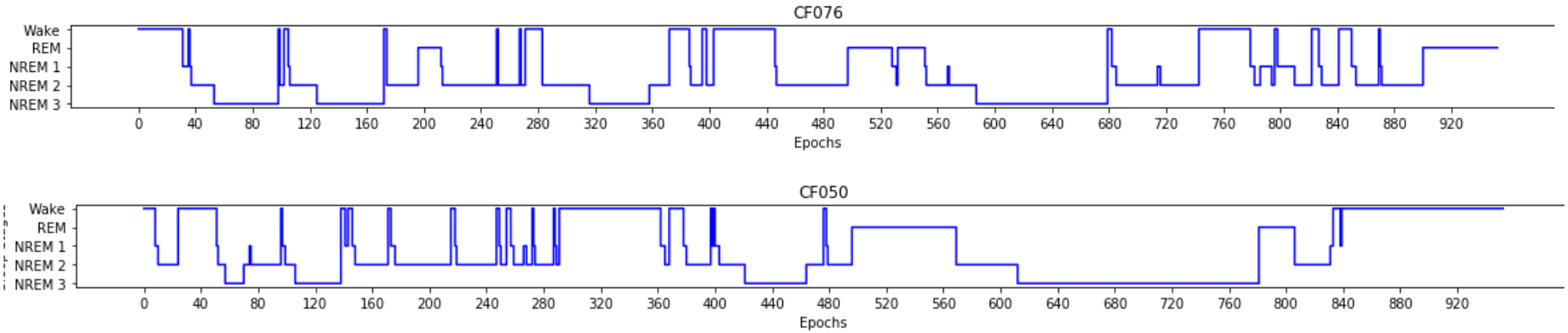


Figure: Hypnograms of a patient with no OSA (top) and severe OSA (bottom) (Tymochko et al, 2020)

- Wake
- Rapid Eye Movement (REM)
- Non Rapid Eye Movement 1 (NREM1) (light sleep)
- Non Rapid Eye Movement 2 (NREM2)
- Non Rapid Eye Movement 3 (NREM3) (deep sleep)

Methods of Exploration

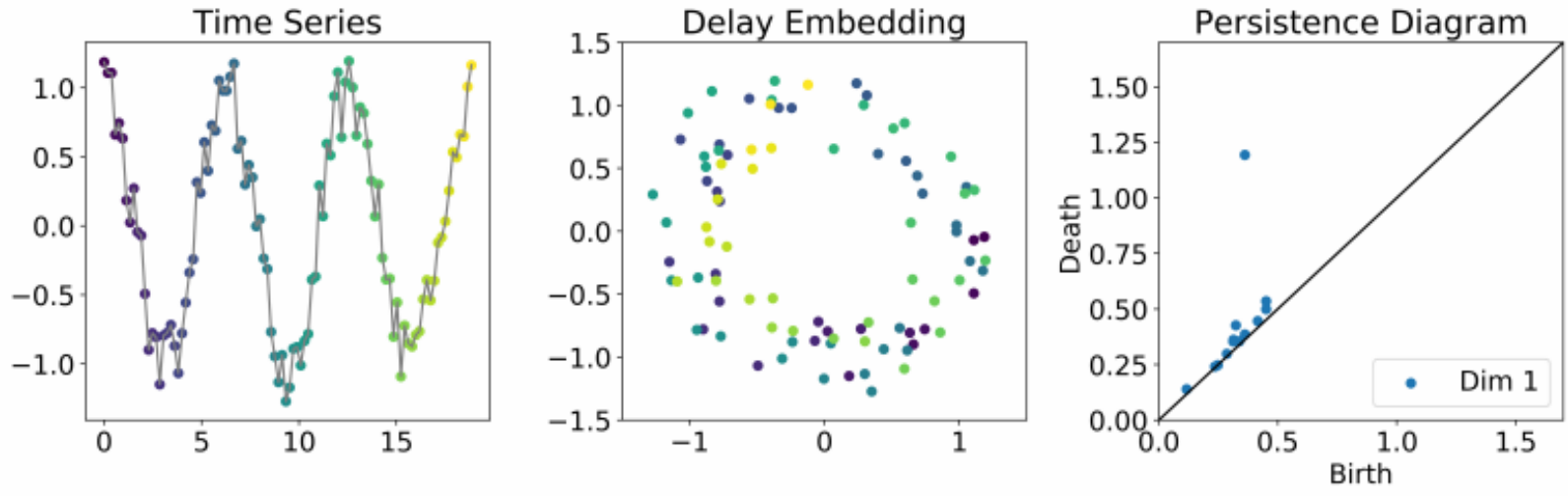


Figure: Time series data can be embedded in a point cloud, which can then be used to generate a persistence diagram. Python packages used: *ripser*, *scikit-tda*, *Persim*. (Tymochko et al, 2020)

- Train and test sleep states on each individual patient, as encoded as persistence images
- Do for 2-5 classes, using several classifiers, compare results

Methods of Exploration

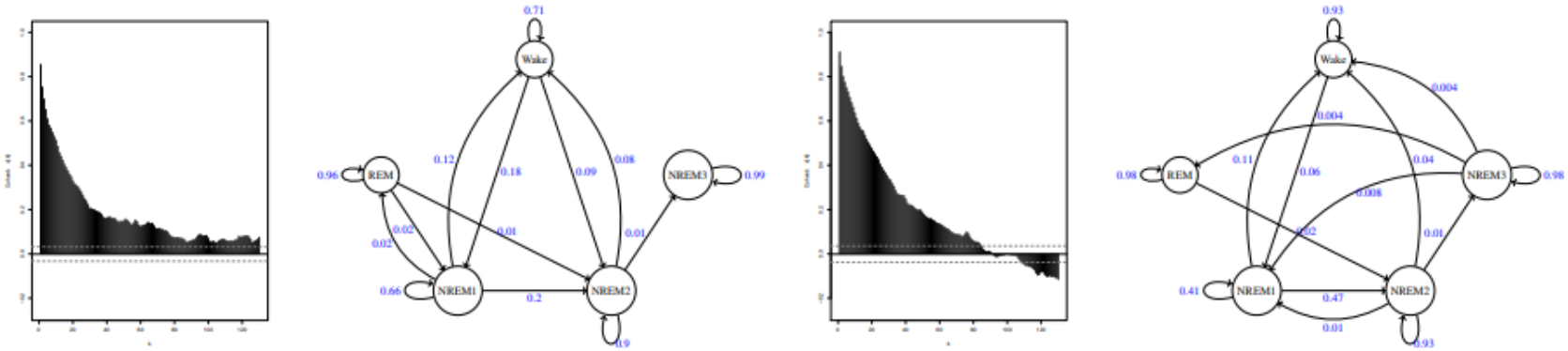


Figure: Cohen's κ plots and representations of transition probabilities for a subject with No OSA (left) and a subject with severe OSA (right) (Tymochko et al, 2020)

- **Goal:** Explore automatic classification of sleep states, observing relationship between OSA and sleep patterns

Results

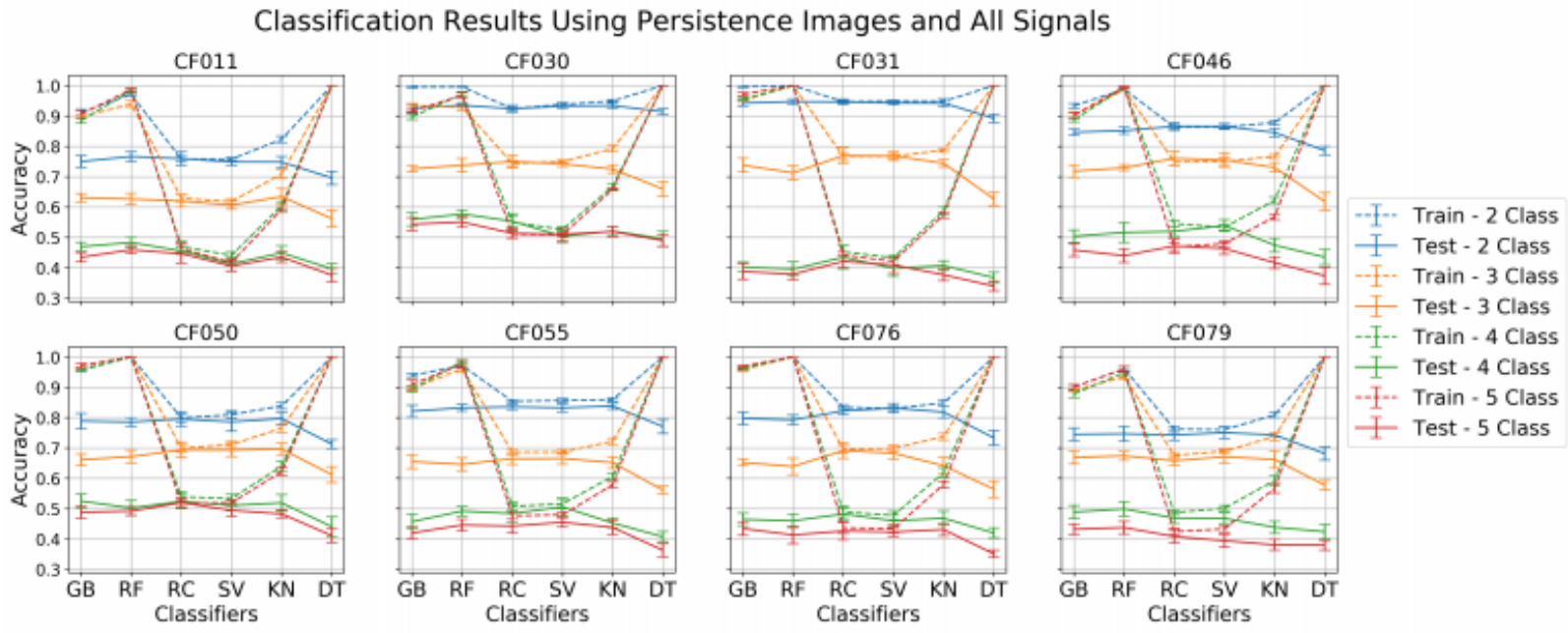


Figure: Classifiers include gradient boosting (GB), random forests (RF), ridge classification (RC) support vector classifier (SV), K-neighbors classifier (KN), and decision trees (DT). (Tymochko et al, 2020)

Cohen's κ and Markov Chains were unable to distinguish severe OSA from no OSA.

Future Directions

Future: Structured Data

- Want to combine the best of the best methods to maximize accuracy, PPV, NPV, sensitivity, specificity
- Want to categorize via the true AHI bins
- Delve further into the driving factors behind these methods

Future: Unstructured Data

- Explore issue of class imbalance in performance, subsampling data to get relatively equal distribution of classes
- Other featurization techniques of persistence diagrams
- Explore application of convolutional neural networks (CNN) and recurrent neural networks (RNN) to classification

Preprints!

- Work started at Women in Data Science and Mathematics (WiSDM) Workshop at Institute for Comupational and Experimental Mathematics (ICERM), Providence, RI, July 2019
- E.T. Winn, M. Vazquez, P. Loliencar, K. Taipale, X. Wang, G. Heo. *A survey of statistical learning techniques as applied to inexpensive pediatric Obstructive Sleep Apnea data.* Preprint, 2020.
- S. Tymochko, K. Singhal, G. Heo. *Classifying sleep states using persistent homology and Markov chain: a Pilot Study.* Preprint, 2020.

Acknowledgements

- WiSDM 2019 @ ICERM Organizers: Ellen Gasparovic, Kathryn Leonard, Linda Ness, ICERM for hosting
- SIAM MDS Women in Data Science Session Organizers: Carlotta Domeniconi and Andrea Bertozzi
- National Science Foundation:
  - EW: NSF Grant No. 1644760
  - ST: NSF Grant DMS 1800446, CMMI-1800466
  - KS: NSF Grant DMS 1547357
- National Sciences and Engineering Research Council in Canada
  - GH: NSERC DG 2016-05167
  - XW: NSERC DG 2019-05917
- Seed grant from Women and Children's Health Research Institute, Biomedical Research Award from American Association of Orthodontists Foundation, and the McIntyre Memorial fund from the School of Dentistry at the University of Alberta

Acknowledgements

Figure: From Left: Giseon Heo, Marilyn Vazquez, Kritika Singhal, Brenda Praggastis, Sarah Tymochko, Emily Winn, Kaisa Taipale, Xu (Sunny) Wang, Melissa Stockman. Not pictured: Pranchi Loliencar