Postdoc Journal
Vol. 4, No. 1, January 2016
Journal of Postdoctoral Research
www.PostdocJournal.Com

How Mathematics is Evolving for Big Data

Ashley M. Fenn
Center for Systems Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA. Email: afenn@mgh.harvard.edu

Outside the world of mathematics it appears that math has remained a steady and consistent field since Newton and Leibniz discovered calculus in the 17th century. After all, 2 x 2 always equals 4 and the slope of a vertical line is always undefined. So how can mathematics really transform? In this month's highlight article, the January Postdoc of the Month Winner, Dr. Brittany Fasy, describes her work in the new mathematical field of Topical Data Analysis (TDA). TDA is a unique combination of algebraic topology and pure mathematics that allows for topological organization of large data sets to identify areas of persistence and thus, relevance. Applications of TDA range from identifying areas of high human traffic for strategic advertising, to characterizing the organization of the universe, to diagnosing cancer.

Today the world is accumulating a gargantuan amount of data. The arrival of smart phones, integrated mobile GPS systems, and thousands of Apps gathering hundreds of thousands of data points per day equate to roughly 2.5 quintillion bytes of data collected *per day*. These so-called "Big Data" sets require a more efficient and effective method of analysis. Analyzing these Big Data sets is precisely why TDA was developed nearly 15 years ago. Large data sets can provide a seemingly endless supply of useful information, but understanding which relationships are meaningful and what information can be discarded is difficult to parse out when dealing with hundreds of millions of data points. In essence, the exciting and relevant data can get lost in the vast amount of noise.

TDA, and specifically persistent homology TDA, uses topological signatures for a data set to identify which data features are the most significant. For the past four years Dr. Fasy has focused her research on the theoretical and practical applications of persistent homology. Persistent homology assumes that real and meaningful relationships will persist longer than noise, and thus isolates points for further analysis by identifying persistent homological features. The first step requires obtaining a large data set and simplifying it into a finite set of points on a plane. Using statistical methods developed by Dr. Fasy and colleagues (Fasy et al., 2014), these "points on a plane" are converted into functions and topological arrangements that correspond to a birth and death rate of a homological feature. A homological feature is a mathematical technique that distinguishes two shapes based on their connectivity and higher-order connectivity – e.g., the shape of a jelly filled doughnut is different from a glazed doughnut because the glazed doughnut has a hole in the middle of it. Now, over a defined interval of time these homological features can be born or die, or in other words they can appear or disappear/converge with other components, and this occurs on a 45-degree angle called the birth/death rate. The points (representing topological features) are then plotted. Points that are on or close to the birth/death rate not statistically significant and are likely topological noise, whereas points statistically distant from the line (having so-called long persistence intervals) are potentially relevant (Fasy et al., 2014).

In her present article (Fasy 2016), Dr. Fasy summarizes the concepts of persistent

homology and points to several real world applications for this mathematical tool. For example, physicians currently screen for prostate cancer through visual confirmation of abnormal or carcinogenic prostate tissue. As expected appropriate classification of prostate cancer from 1 (normal) to 5 (extremely irregular and carcinogenic) can vary depending on the expertise of the physician. Use of a defined mathematical approach like TDA would standardize the prostate cancer grading system and perhaps allow for earlier and more accurate diagnoses.

TDA is not the only mathematical approach to try and standardize our biomedical diagnoses systems using analysis of "Big Data." Researchers at the Ohio State University are using the mathematical Wavelet-Chaos theory to detect and diagnose epilepsy by evaluating small perturbations in brain activity from a large data set of EEG recordings. Because EEG recordings of epileptic individuals can appear normal in the absence of a seizure (interictal EEG), doctors often require positive EEG recordings during a seizure (ictal EEG) or more invasive diagnostic tests including lumbar puncture or a sleep test to reach a diagnosis. Using the Wavelet-Chaos theory and multiple EEG recordings (i.e., millions of data points) from epileptic patients during interictal and ictal periods, researchers are able to train their mathematical formulas to detect small changes in an epileptic patient's interictal EEG (Adeli et al., 2007). Using this mathematical formula, researchers could diagnose future epileptic patients based on a "normal," interictal EEG allowing for early detection and rapid treatment.

There is a note of caution when using mathematical models to summarize large data sets, however. The standardization required to map or evaluate patterns relies on inherent assumptions. In TDA persistent homology, mathematicians assume that relevant data will persist longer than topological noise. In the Wavelet-Chaos theory mathematic formula, the data set used to train the mathematical model uses pre-established epileptic and non-epileptic patient EEGs which may or may not account for the extreme variability within the human population. Thus, some inherent error may be encoded into the formulas based on these assumptions that cause us to misinterpret the data or ignore other important relationships. Nonetheless, given the option of using a mathematical formula with assumptions established by well-respected and studied people in the field versus filing through hundreds of millions of raw data points looking for significant interactions and absolute conclusions, the mathematic models sound like the better option and most likely to succeed. It will be thrilling to see where Dr. Fasy and others in the field of persistent homology TDA take their mathematical models and how they will be applied to learn more about our universe and expand the human experience.

References

Adeli H, Ghosh-Dastidar S, Dadmehr N. 2007. A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy. *IEEE Trans Biomed Eng*, 54(2):205-11.

Fasy BT, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, Singh A. 2014. Confidence sets for persistence diagrams. *Ann Stat*, 42(6): 2301-39.

Fasy BT. 2016. Statistical Techniques in TDA with Applicatiosn to Real Data. *Journal of Postdoctoral Research*, 4 (1) : 1-6. http://doi.org/bdj4