Survey of the Mathematics of Big Data

Philippe B. Laval

KSU

September 12, 2014
We survey some mathematical techniques used with Big Data. The goal here is to make you aware of these techniques rather than giving you detail about them. That task would take several semesters for each technique.
1. How Big is Big Data?
2. How Can Mathematics Help
   - Types of Data
   - Mathematics to the Rescue!
3. Acquisition of Data
   - Information Content
   - Compressed Sensing
4. Data Analysis
   - High Dimensional Data
   - Imaging Data
5. Conclusion
How Big is Big Data?

- We live in a digital world, which generates a lot of data.
- New technologies produce enormous amount of data.
- We are gathering more data than ever, even from old technologies.
- **Problem:** Acquisition/storage, analysis and transmission of data.
  - Total data generated $>\text{ total storage}$.
  - Increase in generation rate $>>$ increase in communication rate.
  - Analysis can be very complex.
- **Problem:** Data is noisy, unstructured and dynamic.
How Can Mathematics Help?

- One answer is given in a video from a previous lecture, remembering that many of the skills mathematicians have are needed in programming.

- But also, mathematics ...
  - ...allows us to formalize both the data and the problem.
  - ...provides a big "chest" of tools or methodologies.
  - ...allows validation of the methodologies (proof of functionality).
Signals. Can be represented by a function $f : \mathbb{R} \rightarrow \mathbb{R}$.
Images. Can be represented by a function \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \)
Data on manifolds. Can be represented by a function $f : S^2 \rightarrow \mathbb{R}$.
Acquisition of Data

- As noted above, the main problem is the **size of the data**.
- However, this cannot be solved by increasing storage capacity.
- Possible solutions:
  - Search for better ways to compress.
  - Acquire less data.
Data Compression saves storage but also decreases bandwidth.

General ideas behind compression:
- Take advantage of redundant information.
- Only keep the most relevant information.

Einstein: "Not everything that can be counted counts, and not everything that counts can be counted."
The standard for data compression was JPEG. It is based on the discrete cosine transform (DCT) (see Fourier analysis, Fourier transform, fast Fourier transform (FFT) and the discrete fast Fourier transform (DFFT). It expresses the image in terms of building blocks, the functions \( \cos \left[ \frac{\pi}{n} \left( i + \frac{1}{2} \right) k \right] \).

**Definition**

The DCT is a linear, invertible function \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \). It transforms the numbers \((x_0, x_1, ..., x_{n-1})\) into \((X_0, X_1, ..., X_{n-1})\) by the formula

\[
X_k = \sum_{i=0}^{n-1} x_i \cos \left[ \frac{\pi}{n} \left( i + \frac{1}{2} \right) k \right] \quad \text{for } k = 0, 1, ..., n - 1
\]
The latest standard for data compression is JPEG2000. It is similar to JPEG, but it uses wavelets instead of the discrete cosine transform.

The branch of mathematics which studies wavelets is called Harmonic Analysis.

Like the DCT, wavelets decompose the image in terms of building blocks. But the building blocks are not limited to the cosine function.

Both the DCT and wavelets use the divide and conquer method.

- Decompose an image in terms of its building blocks.
- Only keep the most relevant building blocks.
Problem: The raw data still has to be stored first
Solution: Compressed sensing (Donoho, Candès, Romberg, Tao (2006))
Main idea:

- **Assumption**: Only a small percentage of the data to be captured is really relevant (sparse data).
- **Problem**: We don’t know which.
- **Solution**: Use random building blocks.
- This is pushing the **sampling theorem** beyond its limits!

Goal:

- **Goal**: To capture a signal with as few points as possible. Thus the raw data will be already compressed.
- **Requires** a lot of linear algebra, solving sparse systems.
Applications of Compressed Sensing

- Business
- Compression
- Astronomy
- Radar
- Biology
- Communications

- Imaging Science
- Geology
- Medicine
- Information theory
- Remote sensing
- Optics
There are many techniques, statistical and mathematical which already exist.

Problem:
- The analysis maybe be too complex for current computing power.
- Data can be noisy, unstructured and dynamic.

However, these problems cannot be simply solved by more computing power.

Possible solutions include:
- Graph Theory.
- TDA (Topological Data Analysis).
Data Analysis - High-Dimensional Data

Strategy:
- Detection of inner structure.
- Reduction of dimension

Examples:
- A circle can be approximated by lines.
- A sphere can be approximated by flat surfaces.
**Question:** How do we identify inner structure?

- **By using topology.** The link will open a website. Scroll down to play the video.
Data Analysis - Imaging Data

Tasks:

- Removing Noise.
- Pattern recognition.
- Reconstructing missing data.
- ...

Tools:

- Statistics.
- Fourier and harmonic analysis.
- Partial differential equations.
- Compressed sensing.
- Linear Algebra.
Big Data is here to stay. We need to tame it!
The techniques presented are fairly difficult and require a lot of knowledge in various branches of mathematics, statistics and science.
Not everybody needs to be a mathematician. But chances are you will work with one. Knowing at least some mathematics will make the communication and the teamwork easier.
Questions?
Some Questions

1. Name some new technologies which generate data which were not mentioned during the talk.
2. Give examples of unstructured data not mentioned during the talk.
3. Why is removing noise from data (images as well as other data) important?
4. Give applications of pattern recognition in images.
Online Articles to Read

1. The Mathematical Shape of Things to Come from Quanta Magazine.
3. The New Shape of Big Data.
4. New Mathematical Method May Help Tame Big Data from Communications of the ACM.