Big Data: Challenges and Opportunities

Daniel P. Palomar

Is Big Data An Empty Hype?

- At first, Big Data seems like the old stuff but with large dimension K and large observation N.
- Large K and N implies big matrices and an increase in computational complexity, but we have powerful computers.
- So what is really big Data?
- Let's look at some aspects...

Distributed Computation/Storage

- In the past, problems were small.
- If they are large, we have powerful computers.
- However, some applications imply huge dimensions and massive sample size, really big data.
- Such applications really require distributed computation (MapReduce paradigm).
- Even storage has to be distributed (HDFS)
- Example: Google's PageRank famous eigenvector computation.

Modeling and Representation

- When fitting the data to a model, we need to estimate some model parameters.
- With large dimensions, the number of parameters explodes, leading to overfitting.
- Solutions based on structure: sparsity (feature selection), low-rank, etc.
- Many dangers like spurious correlations.
- Voluminous large-scale data sets are often incomplete, prone to measurement corruption, and communication errors.
- Robustness to outliers is a must, as well as completion methods.

Small n - Large p Regime

- When the dimension K is huge, it may be the case that N<K.
- Most methods fail and have to be revisited.
- Some tools are very appropriate for this regime and have become extremely popular.
  - Random Matrix Theory
  - Statistical Physics (replica method)

Re-Examine Algorithms

- Low complexity algorithms for large matrices is a must.
- Off-the-shelf interior point methods are computationally prohibited and not amenable to distributed or parallel implementation.
- Approximate algorithms.
- Example: Netflix contest with winner based on low-rank approximation of a huge matrix with many missing entries.

New Applications

- Google's PageRank algorithm (famous eigenvector problem of a matrix with billions of columns/rows)
- Biomedical applications (genomic data analysis, MRI imaging)
- Netflix and recommender systems (low-rank approximation of a huge incomplete matrix)
- Social network applications (massive amount of data to infer or forecast)
- Indoor localization (fusion of data from different sensors, machine learning)
- Waze (map making, traffic estimation)
- Sensor networks
- Massive MIMO (hundreds antennas at BS)
- Sports and health data (pedometer, HR, Formula 1)
- Video surveillance, network anomaly detection, face recognition (with shadows and specularities), singing voice separation from music acompaniment
- Economics and finance (high-frequency trading, social and news indicators)

Other Aspects

- Databases (CS people)
- Hardware for data acquisition (EE people)
- Vulnerability to cyber attacks
- Visualization
- …