Building Tools for Data-Driven Discovery

Genevera I. Allen

Dobelman Family Junior Chair,
Department of Statistics and Electrical and Computer Engineering, Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children’s Hospital.

April 22, 2017
90% of the world's data was created in the last 2 years. We are here.
Data Science
The art and science of extracting meaningful and actionable information from data.
Process of Data Science

- Data Collection
- Data Management
- Data Analysis
- Communicating Results
Process of Data Science

Data Collection

Data Management

Data Analysis

Communicating Results

Domain Expert

Data Scientist
Process of Data Science

Data Collection

Data Management

Data Analysis

Communicating Results

Prediction.

Data-Driven Discoveries.
What is Data Science?

2 Building Tools for Data-Driven Discovery
- Data Integration & Diverse Data
- Graphical Models for Diverse Data

Results
Diverse Data

- National Security.
- Health-care.
- Internet Advertising.
- Energy Industry.
- Environmental Studies.
- Biomedical Technologies.
Mixed, Multi-Modal Data

Mixed or heterogeneous sources of data (sets of features) measured for the same set of subjects or observations.
Statistical Data Integration

Uses probabilistic models to jointly model multiple types of measurements taken on the same set of subjects.
Motivation: The Cancer Genome Atlas

- 33 different cancer types.
- Over 11,000 patients!
- 7 different types of genetics data.
- 2.5 Petabytes worth of data.
Motivation: The Cancer Genome Atlas
Motivation: The Cancer Genome Atlas

Gene Expression

How much is a gene turned off or on in a cell?

MicroRNA Expression

DNA Methylation

Epigenetics:
- Can turn off gene expression
- Changes with Lifestyle & Environment
Motivation: The Cancer Genome Atlas

Gene Expression

DNA → transcription → RNA → translation → folding → protein

MicroRNA Expression

DNA Methylation

RNAseq

DLK1

RNA

mRNA

pre-mRNA

DICER

Degradation
Block Translation

Transcription possible
Transcription impeded

mRNA

miRNA

miRNA expression

DNA methylation
Big-Data & Networks
Motivation: Gene Expression Networks

Glioblastoma gene expression network estimated from microarrays using a Gaussian Graphical Model.
Motivation: Gene Expression Networks
Objective

Objectives

- Develop new families of graphical models that can integrate mixed, multi-modal data.
- Develop new statistical machine learning techniques to estimate integrative graphical models.

Why proper graphical models (Markov Networks)?

- Probabilistic interpretability.
- Statistical guarantees on network recovery.
- Characterize uncertainty.

Reproducible Data-Driven Discoveries.
What is Data Science?

Building Tools for Data-Driven Discovery

- Data Integration & Diverse Data
- Graphical Models for Diverse Data

Results
Networks for Different Data Types

Existing (Markov) Network Types:

- Gaussian Graphical Models (Continuous-Valued).
- Ising Models (Binary-Valued).

What about count-valued data? Skewed? Bounded?

RNA-sequencing data? Methylation data?
Networks for Different Data Types

Our Framework: Graphical Models via Exponential Families.

- Idea: Leverage univariate distributions appropriate for different data types.
- Key Assumption: Conditional distributions are Exponential Families.
  - Ex: Gaussian, Bernoulli, Poisson, Exponential, Negative Binomial, etc.

Skipping the math …

Theorem
Joint network distribution exists and has a closed form!

- Dependencies parameterized by products of sufficient statistics.
- Strong statistical guarantees for network inference.
Networks for Different Data Types

Our Framework: Graphical Models via Exponential Families.
Networks for Different Data Types

Our Framework: Graphical Models via Exponential Families.

Lung Cancer Gene Expression Network
(from RNA-Seq via Poisson Graphical Models).
Integrated Network Models

Gene Expression

DNA → Transcription → RNA → Translation → Amino Acid Chain → Protein

MicroRNA Expression

DNA Methylation

Transcription possible

Transcription impeded

Degradation → Block Translation
Integrated Network Models

Mixed Chain Graphical Models via Exponential Families.

- Key Assumption:
  - Conditional distributions are different Exponential Families.

Skipping the math …

Theorem
Joint integrated network distribution exists and has a closed form!

- Dependencies parameterized by products of sufficient statistics from different distributions.
- Strong statistical guarantees for network inference.
Integrated Network Models

Mixed Chain Graphical Models via Exponential Families.
Integrated Network Models

Mixed Chain Graphical Models via Exponential Families.
Integrated Network Models

**Implication**

First multivariate distribution for mixed data types.
1. What is Data Science?

2. Building Tools for Data-Driven Discovery
   - Data Integration & Diverse Data
   - Graphical Models for Diverse Data
   - Results
Case Study: Ovarian Cancer Regulatory Network

\[ \mathbb{P}_{Pois} (\text{Genes} \mid \text{miRNAs, CpGs}) \mathbb{P}_{Gauss} (\text{miRNAs} \mid \text{CpGs}) \mathbb{P}_{Ising} (\text{CpGs}) \]

\( n = 293 \) patients & \( p = 1016 \) genes, miRNAs or CpG sites selected for association with survival, recurrence, chemoresistance or mutation rate.
Case Study: Ovarian Cancer Regulatory Network
Case Study: Ovarian Cancer Regulatory Network
Future Work

Tools:

- Feature Selection.
- Clustering.
- Pattern Recognition & Dimension Reduction.

Applications:
Software

**XMRF**: An R Package to Fit Markov Networks to High-Throughput Genomics Data.

**TCGA2STAT**: Simple TCGA Data Access for Integrated Statistical Analysis in R.
Acknowledgments

Collaborators:

- **Pradeep Ravikumar**, University of Texas, Austin.
- **Eunho Yang**, IBM Watson Research Center.
- **Yulia Baker**, PhD Candidate, Statistics, Rice University.
- **Zhandong Liu**, Baylor College of Medicine.
- **Ying-Wooi Wan**, Baylor College of Medicine.
- **Matthew Anderson**, Baylor College of Medicine.
Thank You!