A Blind Date with Big Data

Goal

- Measure correlations
- Rank relatedness

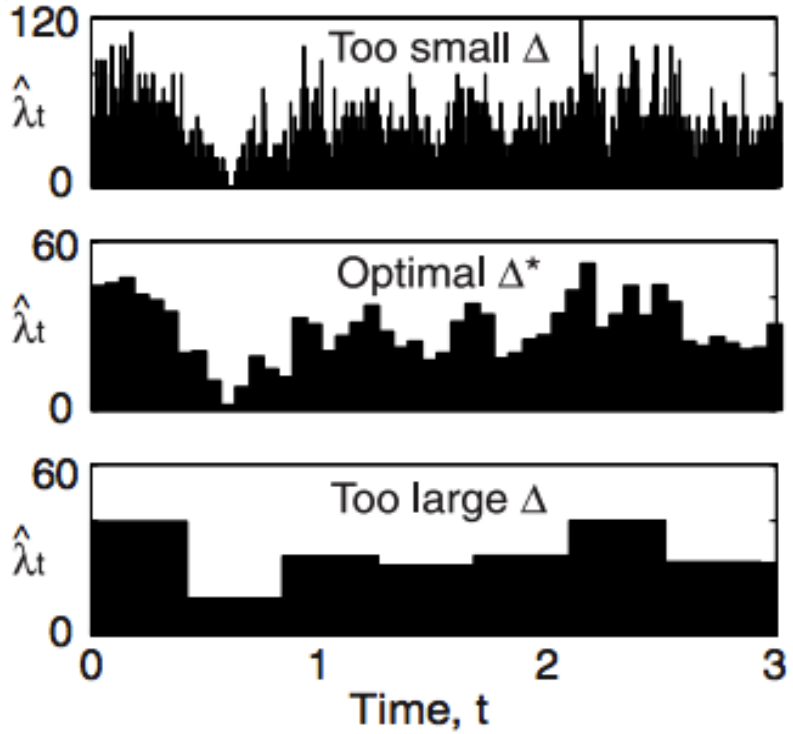
Challenges

- Conditional dependencies
- Missing data

Where We Are

Correlation Calculations

- Mutual Information
- Optimal Histograms
- Efficient Algorithm
  - Ranks NHANES in hours



$$\mathrm{MI}_{XY} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \frac{n_{(i,j)}}{n_{(:,:)}} \log \frac{n_{(:,:)} * n_{(i,j)}}{n_{(i,:)} * n_{(:,j)}}$$

Shimazaki H. and Shinomoto S., A method for selecting the bin size of a time histogram. Neural Computation. (2007) Vol. 19(6), 1503-1527

Where We're Going

Imputation

- Developing new method based on optimal bins

Correlations

- Explore Kernel Density Estimation

Conditionals

- Probabilistic Graphical Models

Personal Goals

- Practice programming with variety of languages
- Study statistics, bioinformatics, information theory
- Learn about all projects of the Sabeti Lab
- Have fun!