Astronomy 6523
Modeling, Inference & Data Mining in Astrophysics

Professor J. Cordes
622 Space Sciences Building
Tuesday – Thursday 1:25-2:40 pm
http://www.astro.cornell.edu/~cordes/A6523

- Bayesian and Frequentist methods, optimization
- Mining of weak signals in large-scale surveys
  (in space, time, frequency, …)
- Characterization, clustering, principal components
- Stochastic processes and modeling
- Precision timing, Doppler, and spatial localization
- Time-series analysis & image formation and analysis
- Mining and machine learning in Big Data contexts
- Visualization and data presentation

Examples and case studies:
- Exoplanet discovery and analysis
- Gravitational wave detection
- Red noise processes in geophysics, finance, and astrophysics
- Massive surveys in Astronomy (radio arrays, LSST)
- Cosmic bursts (milliseconds to months)
- Pulsar discovery and timing, GR tests

A6523
Signal Modeling, Statistical Inference and Data Mining in Astrophysics
Spring 2015

- What the course is about (syllabus)
- What I need to know about your background:
  Data ➔ Short quiz to help calibrate lectures
- What you need to do in the course
  • Homework, small projects, large project
  • Participation in class (interactive = better)
- Code:
  • Some will be supplied for small data challenges (python)
  • You are free to supply your own (and will have to)
Books etc.

• Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*

• Lee, *Bayesian Statistics, An Introduction*

• Ivezic et al., *Statistics, Data Mining, and Machine Learning in Astronomy*

• Notes by JMC.

• Articles and chapters from other books
Quiz to help define level of course

- Note that items in the quiz made no mention of astronomy.
- All are quite general methods or part of the lexicon of data mining
- Astronomy provides a context for virtually every known statistical method and provides unique challenges for new methods:
  - Sparse object populations in data sets
  - Low signal to noise ratio (not like data mining of credit-card usage or the internet archive)
  - Raw data that do not compress (in the usual way)
Reading for next week

• Chapters 1-2 from Gregory
• On course web site:
  – Course goals document
  – Intro lecture on probability and statistics (Penn State Astrostatistics)
  – Introduction to astroML: Machine Learning for Astrophysics

Some of my research
Arecibo Pulsar Survey Data Flow
Survey analysis + Einstein@Home extended survey analysis

Searching for Binary Pulsars with Short Orbital Periods

• Einstein@Home:
  • $2 \times 10^5$ worldwide volunteers process LIGO data
  • LIGO data served from U.S. (U. Wisc. Milwaukee)

• Now Einstein@Home volunteers to process Arecibo PALFA data
  • raw data Cornell -> Albert Einstein Institute (Hannover)
  • Dedispersed time series to each E@H client
  • Spend ~12 hr processing each time series to search for binary pulsars with $P_{\text{orb}}$ as short as ~10 min
  • Candidate signals sent back to servers at Hannover and Cornell, database at Cornell
  • Meta-analysis to select candidates for reobservation at Arecibo
Using Pulsars for Gravitational Wave Detection

A galactic-scale GW detector: the Pulsar Timing Array

GW perturbations are correlated among different pulsars.

Need to observe an ensemble of MSPs to extract the correlated signal from the noise.
Difficulties of GW Detection $\Delta L/L \sim h$

<table>
<thead>
<tr>
<th>Pulsar Timing Array</th>
<th>Ground-based Interferometer</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L \sim cT \sim 3 \text{ pc}$</td>
<td>$L \sim 4 \text{ km} \times 50 \text{ reflections}$</td>
</tr>
<tr>
<td>$h_{\text{min}} \sim 10^{-16} - 10^{-14}$</td>
<td>$h_{\text{min}} \sim 10^{-23}$</td>
</tr>
<tr>
<td>$\Delta L \sim 10^{3}$ to $10^{5} \text{ cm}$</td>
<td>$\Delta L \sim 10^{-16}$ cm</td>
</tr>
</tbody>
</table>

PTA: $\delta t$ includes:
- Translational motion of the NS $\sim 100$ to $1000 \text{ km/s}$
- Orbital motions of the pulsar and observatory: $10$s – $100$s km/s
- Interstellar propagation delays: ns to seconds

Using Pulsars as Clocks: Precision Timing of Pulsars

Glitches
Spin noise
Magnetosphere

Differential rotation, superfluid vortices
Interstellar dispersion and scattering
Emission region: beaming and motion

Uncertainties in planetary ephemerides and propagation in interplanetary medium
GPS time transfer
Additive noise
Instrumental polarization
# Course Emphasis

<table>
<thead>
<tr>
<th>Principles</th>
<th>Math and statistical methods</th>
<th>Algorithms</th>
<th>Applications and implementation</th>
</tr>
</thead>
</table>

- **Design vs. Inference**
  - **Engineering applications**
    - Physics + engineering → Devices, machines, software → Operations, signals
  - **Astrophysics and Space Science**
    - Measurements of photons, non-photonic messengers (GWs, cosmic rays, neutrinos) → Signal processing, statistical inference, hypothesis testing, classification → Physical models, testing of fundamental physics, understanding cosmic evolution
**Spectral analysis as a unifying thread**

Signals ⇔ Statistics

**Spectral analysis:**

1. Analysis of variance in a conjugate space
   
   \[ t \leftrightarrow f \] (time and frequency domains)
   
   \[ u,v \leftrightarrow \theta \] (interferometric images)

• Statistical questions about the nature of the signal in frequency space:
   
   a. Is there a signal?
   
   b. What is its frequency?
   
   c. What is the shape of the spectrum?

1. **Basis functions:**

   - Sinusoids \[ t \leftrightarrow f \]
   - Spherical harmonics \[ \theta, \varphi \leftrightarrow l,m \]
   - Wavelets time-frequency atoms
   - Principal components the data determine the basis

   The appropriate basis (often) is the one that most compactifies the signal in the conjugate domain