Data analysis for science

Lee Samuel Finn
Center for Gravitational Wave Physics
Overview

• “Data analysis” goals

• Distinguishing signal from noise: Examples

• What this means for you
Data Analysis v. Source Simulation

- **Source Simulation**
  - Goal: *Identify source science* impressed on gravitational wave signal
  - Important question: how is source science encoded in radiation?

- **Data Analysis**
  - Goal:
    - Distinguish between signal and noise
    - Discriminate to identify source science in signal
      - E.g., source parameters like ns/bh masses, or spins, population statistics, etc.
    - Interpretation: place observations in (astro)physical context
  - Important question: how to *maximize contrast* between signal, noise?
What distinguishes?

- Measure of distinction: likelihood or sampling distribution
  - P(d | \(\Theta\), I) : prob of observing d given \(\Theta\), I
  - d – quantity (not necessarily \(h[t_k]\)) calculated from measurement at detector
  - \(\Theta\) - all the parameters that distinguish among signals
    - Amplitude, population, etc.
  - I - everything relevant about detector and noise

- What is d?
  - Depends on noise, sought-for signal
  - We’ll return to this point!

- P(d | 0, I)
  - Probability that observation is of noise alone (no signal)

- P(d | \(\Theta\), I)
  - Probability observation is of noise + signal \(\Theta\)

- Likelihood of d: “odds” signal v. noise:
  - \(\Lambda = P(d | \Theta, I) / P(d | 0, I)\)

Goal: Make probabilities P(d | \(\Theta\), I), P(d | 0, I) as different as possible!
The Detection Game

• **The Game:**
  – Observe $h[t_k]$
  – Calculate $d$
  – Calculate $P(d | \Theta, I)$
  – If $P(d | \Theta, I) < P_0$, buy tux, tickets to Stockholm

• **Choice $P_0$:** false alarms, false dismissal
  – False alarm prob: frequency with which noise alone (no signal) would give $d$ such that $P(d | \Theta, I) < P_0$
  – False dismissal prob: frequency with which noise + signal $\Theta$ would give $d$ such that $P(d | \Theta, I) > P_0$
  – Efficiency $:= 1 – (\text{false dismissal prob})$

• **Note:** the more different $P(d | \Theta, I), P(d | \Theta, I)$, the smaller the false dismissal for a given false alarm
Expressing the contrast: False alarm v. efficiency

- Guessing: pick a random number between 1 and 100
  - If less than N+1 then say detected
- False alarm probability?
  - N/100
- Efficiency?
  - N/100
- Close to diagonal is close to random guessing
- Better tests have greater lift off diagonal
  - High efficiency for low false alarm probability
Clearing the Clutter

• Goal: make contrast $P(d | \Theta, I) / P(d | 0, I)$ large
  – How? Can’t choose, change signal, noise
  – Only possibility: choose $d$!
  – Choice of $d$ based on signal characteristics and their uncertainty (in nature or knowledge)

• Examples:
  – Stochastic gravitational wave signal
  – Periodic signals
  – Gravitational waves from $\gamma$-ray bursts
  – Bursts: things that go “bump” in the night
Stochastic gravitational wave signal

• “Signal” is noise
  – How do we distinguish gw contribution to total “noise”?

• What’s distinguishes signal, instrumental contributions?
  – Physically distinct detectors respond coherently to gravitational waves

• Quantity that distinguishes
  – Cross correlation: \( d = \iint dt_1 dt_2 h_1(t) h_2(f) Q(t_1 - t_2) \)
  – Choose kernel Q to extremize contrast in d between signal present, absent cases

• Key point: look for, choose measure that draws the greatest contrast between signal, noise
(nearly) Periodic Signals

- **Signal**
  - $s(t) = A \sin [\Phi(t) + \phi_0]$
  - Know $\Phi(t)$ accurately, unknown $\phi_0$, $A$

- **What distinguishes?**
  - Noise not periodic with known phase
  - Signal has no power except at frequencies near $d\Phi/dt$
  - Phase $\phi_0$ not important

- **Identify a quantity that large for signal, small for noise:**
  \[ \rho^2 = x^2 + y^2 \]
  \[ x = \frac{1}{T} \int_0^T dt \, h(t) \cos \Phi(t) \]
  \[ y = \frac{1}{T} \int_0^T dt \, h(t) \sin \Phi(t) \]

**Key point:** phase must be known s.t. $\Delta \Phi \ll \pi$ for all $t$
The $\gamma$-ray Burst Story

Key Facts:
- Multiple, indistinguishable triggers
- Rapidly rotating (Jc/GM$^2$~1) BH
- $\gamma$-ray production far from BH
- Sources likely too distant (z~1) to detect individuals
- Gravitational wave strength, time dependence unknown

Hypernovae; collapsars; NS/BH, He/BH, WD/BH mergers; AIC; …

Black hole + debris torus

Relativistic fireball

$\gamma$-rays generated by internal or external shocks
What science might we learn?

- Progenitor mass, angular momentum
  - Expect radiated power to peak at frequency related to black hole $M, J$
- Differentiate among progenitors
  - Radiation originating from stellar collapse, binary coalescence have different gw intensity, spectra
- Internal vs. external shocks
  - Elapsed time between gw, g-ray burst depends on whether shocks are internal or external
- Analysts goal: describe an analysis that brings science into contrast
  - Spectra, elapsed time between g, gw bursts

Hypernovae; collapsars; NS/BH, He/BH, WD/BH mergers; AIC; …

Relativistic fireball

Black hole + debris torus

$\gamma$-rays generated by internal or external shocks
The LIGO Lock-in

Incident waves give correlated detector output

Integrated cross-correlated detector output:

\[
\langle s_H, s_L \rangle \equiv \int_0^T dt dt' s_H(t_H - t) s_L(t_L - t') K(t - t')
\]

\[
x_{\text{off}} = \langle n_H, n_L \rangle
\]

\[
x_{\text{on}} \equiv \langle n_H + h_H, n_L + h_L \rangle
\]

\[
\equiv x_{\text{off}} + \langle h_H, n_L \rangle + \langle n_H, h_L \rangle + \langle h_H, h_L \rangle
\]

Collect catalog of \(<s,s>\)
associated, not associated
with GRBs,

Source population average

\[
x_{\text{off}} = \mu_{\text{off}}
\]

\[
x = x_{\text{off}} + \langle h_H, h_L \rangle
\]

Collect & compare on-burst, off-burst catalogs:
Are distributions different?

Incident waves give correlated detector output

\[ h_{ij} \]

- Accuracy of estimated means increases the more samples are available
- Gather enough samples and any difference becomes distinguishable from zero

\[
\overline{x}_{\text{off}} = \mu_{\text{off}} \\
\overline{x}_{\text{on}} = \mu_{\text{off}} + \langle h_H, h_L \rangle
\]
Discovery: Things that go bump in the night

- How to discover?
- What distinguishes signal, noise?
  - Signal time-limited
  - Signal (s), noise (n) uncorrelated: \(<sn>=0\>
    - Important: \(<(n+s)^2>\) greater than \(<n^2>\>
- Analysis method: look for anomalies
  - Where is detector output unusual?
  - Where do noise statistics change?
Example: Power

- How does power in a frequency band evolve?
  - $|h(f)|^2$ measured over short intervals of time
- Spectrogram
  - Band-limited signals
  - Signals that exhibit interesting “time-frequency” behavior
- Refs.
  - Sylvestre. Phys Rev. in press. (gr-qc/0210043)
Example: Change-point analyses

- Look for places where statistics change
  - Statistics? Mean, variance
  - Needn’t assume any particular mean, variance: look for changes
- Refs.
  - Time series: Finn & Stuver in progress
What does this mean for you?

• **Source simulator’s job**
  – *Identify science* reflected in the gravitational waves
    • The science is the signal!
  – Find wave description draws the *science* into sharpest focus
    • Frequency, bandwidth, duration, polarization, …?
  – Connect the source to an astrophysical context
    • Amplitude, rate, space density, etc.
  – Don’t forget uncertainties!

• **The data analyst’s job**
  – Develop analyses that makes *science* stand-out
    • The science is the signal!
  – Provide astrophysical interpretation of observations