Physics 290e
Tools of the Trade:
Getting the Most from Your Data

August 30, 2017

Organizational Isssues (I)

- Each semester we choose a topic
  - This semester: Tools for data analysis
- This is a seminar and not a lecture class
  - No problem sets and no exams
  - You will be expected to give a talk
- Plans for the semester
  - First few talks from faculty and LBL statt
  - Then, students talk
  - You choose the subject (as long as it fits with the topic for the semester)
  - Tag-teams of two people giving back-to-back talks on related topics encouraged
- Barbara Jacak and I will be organizing the seminar
  - email us with your proposed topic and date
  - First-come, first-serve

Organizational Issues (II)

- Postdocs, students and staff welcome
- All talks posted in advance on LBL indico page
  https://indico.physics.lbl.gov/indico/category/17/
- ReadyTalk connection so off-site students can call in remotely
  +18667401260
  access code: 4866608
- You are encouraged to pick a topic that will teach you something new rather than recycling an old talk (or talking about your thesis analysis)
- We'll go through some suggestions today (just to get you thinking)
- A list of possible topics will be posted on bCourses and in indico

Why have we chosen this topic for the semester?

- Data analysis is central to what all experimental physicists do
  - But the skills to do this are rarely taught in classes
- In particle physics, we deal with large amounts of data
  - Approaches such as cloud computing essential
  - Yncreases is available CPU, memory and storage allow us to do things we couldn't do before
  - We need to be able to react to changes in industry standards
    - New architectures
    - Commodity computing
    - More CPU, less memory
- And care about precise predictions and measurements
  - New methods can help
    - Multivariant: neural nets and BDTs
    - Other Machine learning techniques
- Also need to better estimate backgrounds
  - Better simulations
  - Data driven methods

These are topics to expore this Fall

How a talk might be structured
(a suggestion not a requirement)

- Pick a topic
- Identify a technique relevant for the topic
- Find a physics measurement that relies on that technique
- Explain in your talk
  - What physics measurement you are going to present
    - Why is it interesting?
    - Where can it be measured?
    - What is the the strategy for making the measurement
  - What limits the measurement?
    - Which uncertainties can be reduced using appropriate techniques
- Describe the technique
- Explain how it was used in the measurement of interest

Some possible directions follow

Pattern Recognition

- Most detectors measure hits.
  - Need to turn them into space points
  - Space points combined to form "objects"
- Tracking detectors
  - "connect the dots" to form tracks
    - Curved trajectory if magnetic field
  - Must remove "fake" tracks and/or misassigned hits
  - Figures of merit: efficiency and fake rate
  - Other things might matter as well: eg 2-track resolution

  How does this work for <*pick your favorite experiment*>?
- Calorimeters (Either total absorption or sampling)
  - How do we turn measured pulse height into estimate of incident particle's energy?
  - How do we turn energy into a 3- or 4-momentum? (what is the direction)?

  < *Pick an interesting measurement*> and illustrate one or more of these methods

Particle ID

- Techniques very experiment dependent
- Can measure
  - Momentum
  - Energy
  - Time of flight
  - Energy loss in material (dE/dx, brem, Transition radiation)
  - Shower shape in a calorimeter
- These then combined to separate species
- Some interesting examples
  - Using cherenkov pattern to identify electrons in SuperK
  - Using combined tracking and calorimeter information to find electrons in ATLAS or CMS
  - Distinguishing WIMPS from background in CDMS
  - Using dE/dx to identify π, K, p, e in Alice

Uses of Simulated Data

- Simulated data used for many things
  - Predicting cross sections for complicated processes
  - Understanding performance of reconstruction algorithms
  - Calculating acceptance and efficiency
  - Estimating background from other physics processes
  - Providing detailed response matrices for unfolding
    - Reconstructed distribution $\rightarrow$ true distribution
- Some topics of interest
  - How does a Monte Carlo generator such as Pythia8 work?
  - How do we use GEANT4 to simuate a detector?
  - Can Monte Carlo datasets help us understand our systematic uncertainties?

Machine Learning

- In the old days, people just separated signal and background by makign *cuts* on specified variables
- This works if there is good separation between signal and background
- But one can often do better (typically $\sim 30\%$ but sometimes much more) using multivariant techniquest
- This is an active area of research both in particle physics and elsewhere
- Some interesting topics
  - Using BDTs to identify B hadrons
  - Machine learning as a tool in simulation
  - Machine learning techniques to improve triggers

Fitting in the presence of backgrounds

- If we know the shape of our background and our signal, this is easy
- But what if we don't?
  - Can we use analytic forms to estimate the background and "bump hunt"?
  - Can we use Monte Carlos to estimate the background
    - And how do we handle the uncertainty on the estimate?
  - Can we use "control regions" to do the estimate
    - And how do we extrapolate from the control region to the signal?
- Examples of topics of interest (there are hundreds)
  - How do we find the flux of neutrinos from the sun?
  - What is the cross section for Higgs production?
  - Is there really a pentaquark?
  - How were the mixing parameters for B^0 (or B_s) mixing measured?

Estimating Signficance

- Suppose we see an excess. How significant is it?
  - Of course, there is the local statistical significance, but that isn't all
  - Need to know how many places might have an excess
    - "Look elsewhere effect"
  - What is the systematic uncertainty on the shape and normalization of the background?
- Suppose we don't see an excess? What can we rule out
  - Many of the same issues apply
- Topics of interest
  - How is the 95% limit set for $<$ *Pick your favorite experiment* $>$
  - Do I really believe that the $<$ *Pick your favorite particle* $>$ was found?
  - Can I determine the spin and/or parity of $<$ *Pick your favorite particle* $>$