Regression with Gaussian Processes

Saarik Kalia
Professor Daniel Whiteson

CERN

August 9, 2016

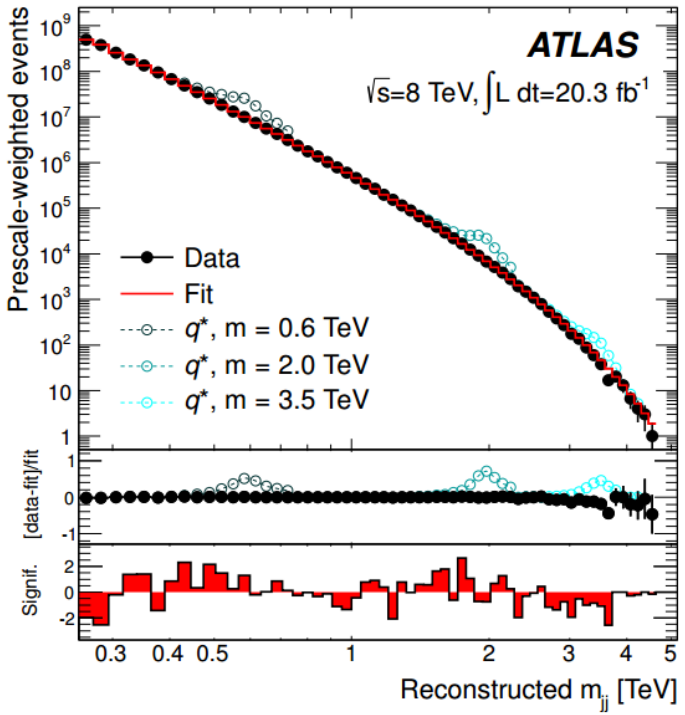
Fitting with Fixed Functional Forms

- Many plots in experimental physics use functional forms to fit background curves

Fitting with Fixed Functional Forms

- Many plots in experimental physics use functional forms to fit background curves
- Dijet mass spectrum with $\sqrt{s} = 8$ TeV uses form

$$f(x) = p_1(1-x)^{p_2} x^{p_3 + p_4 \ln x}$$



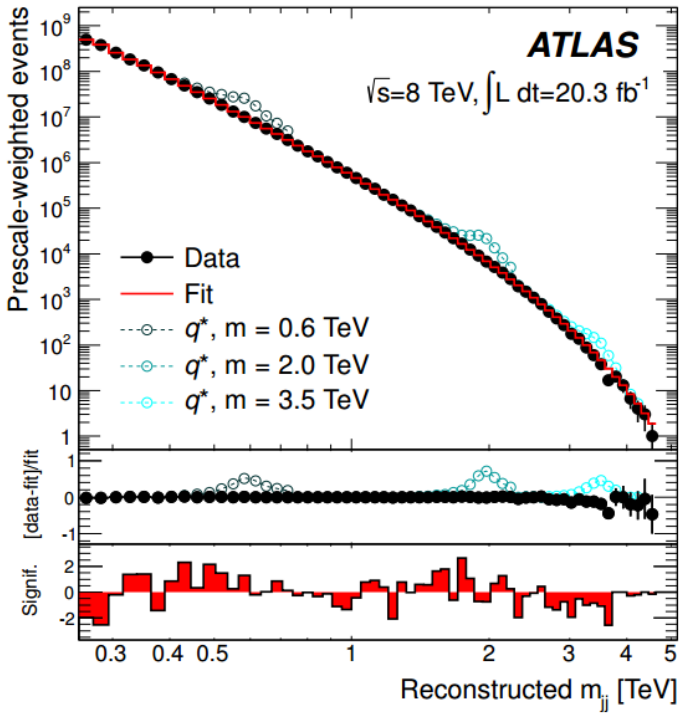
Taken from "Search for new phenomena in the dijet mass distribution using pp collision data at $\sqrt{s} = 8$ TeV with the ATLAS detector"

Fitting with Fixed Functional Forms

- Many plots in experimental physics use functional forms to fit background curves
- Dijet mass spectrum with $\sqrt{s} = 8$ TeV uses form

$$f(x) = p_1(1-x)^{p_2} x^{p_3 + p_4 \ln x}$$

- Sometimes just guessed, not theoretically well-motivated



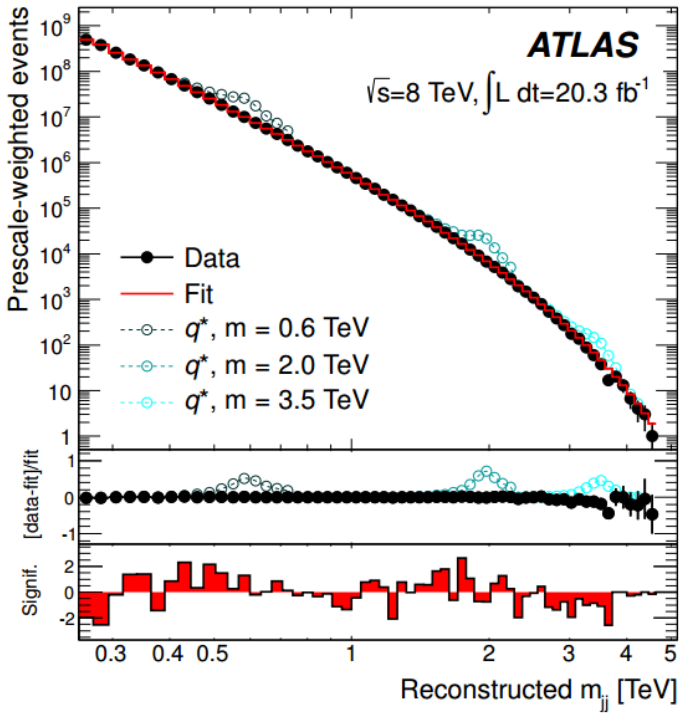
Taken from "Search for new phenomena in the dijet mass distribution using pp collision data at $\sqrt{s} = 8$ TeV with the ATLAS detector"

Fitting with Fixed Functional Forms

- Many plots in experimental physics use functional forms to fit background curves
- Dijet mass spectrum with $\sqrt{s} = 8$ TeV uses form

$$f(x) = p_1(1-x)^{p_2} x^{p_3 + p_4 \ln x}$$

- Sometimes just guessed, not theoretically well-motivated
- Doesn't work well at high luminosity



Taken from "Search for new phenomena in the dijet mass distribution using pp collision data at $\sqrt{s} = 8$ TeV with the ATLAS detector"

Gaussian Processes

- Want functionless regression technique

Gaussian Processes

- Want functionless regression technique
- Assume values at nearby data points correlate

Gaussian Processes

- Want functionless regression technique
- Assume values at nearby data points correlate
- For simplicity, assume mean is zero and joint distribution is multivariate normal

Gaussian Processes

- Want functionless regression technique
- Assume values at nearby data points correlate
- For simplicity, assume mean is zero and joint distribution is multivariate normal
- Only need to choose covariances between values at pairs of points

Kernels

- Covariances between points are given by covariance function, or kernel

Kernels

- Covariances between points are given by covariance function, or kernel
- Common choice is the exponential squared kernel:

$$K(x, y) = Ae^{-\frac{(x-y)^2}{2\ell^2}}$$

Kernels

- Covariances between points are given by covariance function, or kernel
- Common choice is the exponential squared kernel:

$$K(x, y) = Ae^{-\frac{(x-y)^2}{2\ell^2}}$$

- Want amplitude and length scale to vary:

$$K(x, y) = A(x) \cdot A(y) \cdot \sqrt{\frac{2\ell(x)\ell(y)}{\ell(x)^2 + \ell(y)^2}} \cdot e^{-\frac{(x-y)^2}{\ell(x)^2+\ell(y)^2}}$$

Kernels

- Covariances between points are given by covariance function, or kernel
- Common choice is the exponential squared kernel:

$$K(x, y) = Ae^{-\frac{(x-y)^2}{2\ell^2}}$$

- Want amplitude and length scale to vary:

$$K(x, y) = A(x) \cdot A(y) \cdot \sqrt{\frac{2\ell(x)\ell(y)}{\ell(x)^2 + \ell(y)^2}} \cdot e^{-\frac{(x-y)^2}{\ell(x)^2+\ell(y)^2}}$$

- Choose $A(x) = Ce^{-\frac{x}{d}}$ and $\ell(x) = Lx^p$

Kernels

- Covariances between points are given by covariance function, or kernel
- Common choice is the exponential squared kernel:

$$K(x, y) = Ae^{-\frac{(x-y)^2}{2\ell^2}}$$

- Want amplitude and length scale to vary:

$$K(x, y) = A(x) \cdot A(y) \cdot \sqrt{\frac{2\ell(x)\ell(y)}{\ell(x)^2 + \ell(y)^2}} \cdot e^{-\frac{(x-y)^2}{\ell(x)^2+\ell(y)^2}}$$

- Choose $A(x) = Ce^{-\frac{x}{d}}$ and $\ell(x) = Lx^p$
- Four parameters C, d, L, p

Choosing Parameters

- When fitting for background:
  - Set limits on parameter space
  - Choose allowed parameters that maximize the likelihood of observing data

Choosing Parameters

- When fitting for background:
  - Set limits on parameter space
  - Choose allowed parameters that maximize the likelihood of observing data
- When fitting for background + signal:
  - For simplicity, assume signals are Gaussian (in practice, use other shapes)

Choosing Parameters

- When fitting for background:
  - Set limits on parameter space
  - Choose allowed parameters that maximize the likelihood of observing data
- When fitting for background + signal:
  - For simplicity, assume signals are Gaussian (in practice, use other shapes)
  - Allowed to subtract off Gaussian before fitting
  - Parameters of Gaussian (height, width, mean) are optimized with kernel parameters

Quality Metrics

- We need a method that is flexible enough to fit the background, but rigid enough not to fit a signal
- Two measures of effectiveness:

Quality Metrics

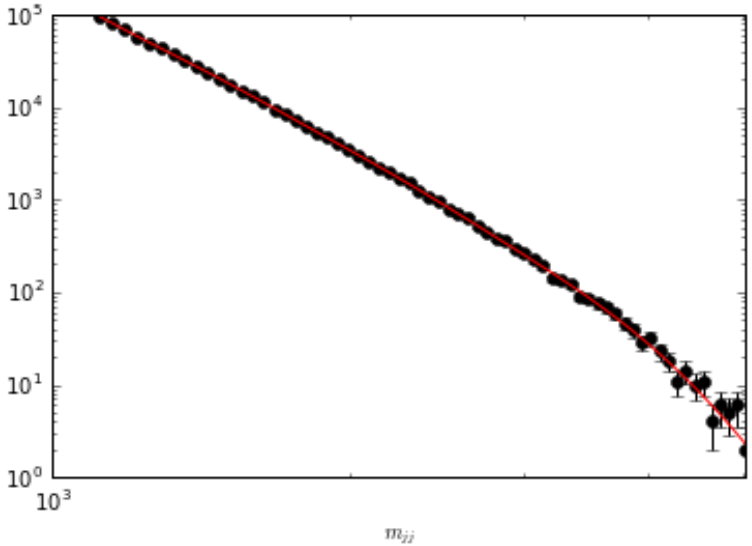
- We need a method that is flexible enough to fit the background, but rigid enough not to fit a signal
- Two measures of effectiveness:
  - χ^2 value = sum of squares of statistical fluctuations
    - Good value is approximately equal to degrees of freedom
    - Can also look at fluctuations at each point individually

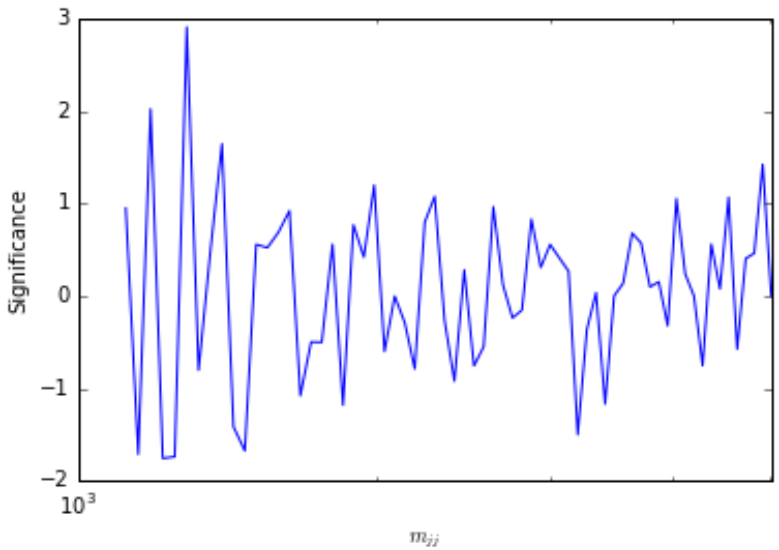
Quality Metrics

- We need a method that is flexible enough to fit the background, but rigid enough not to fit a signal
- Two measures of effectiveness:
  - χ^2 value = sum of squares of statistical fluctuations
    - Good value is approximately equal to degrees of freedom
    - Can also look at fluctuations at each point individually
  - Inject signals of varying heights, and compare to fitted signals
    - Should exhibit linear relationship

Results

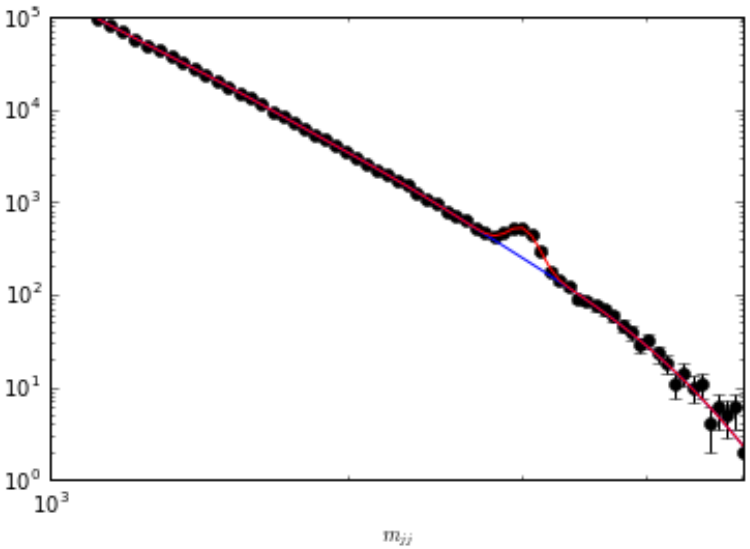
Background fit and significances ($\chi^2 = 60.16$ for 62 d.o.f.)

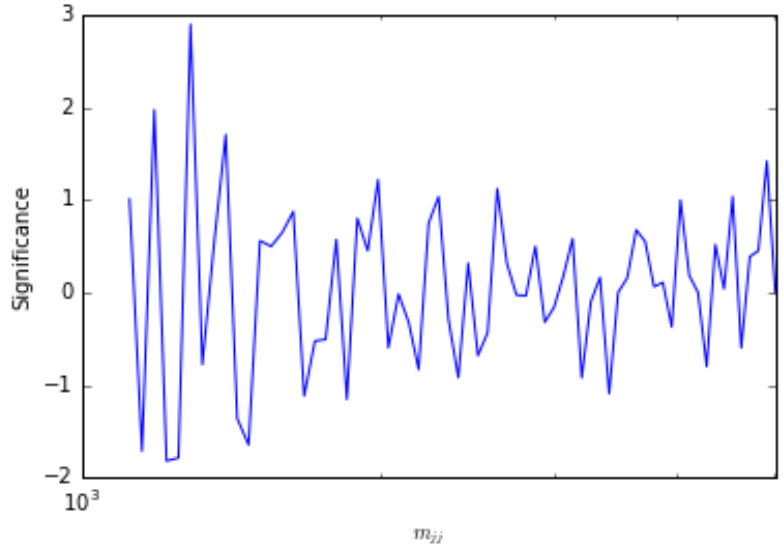




Results

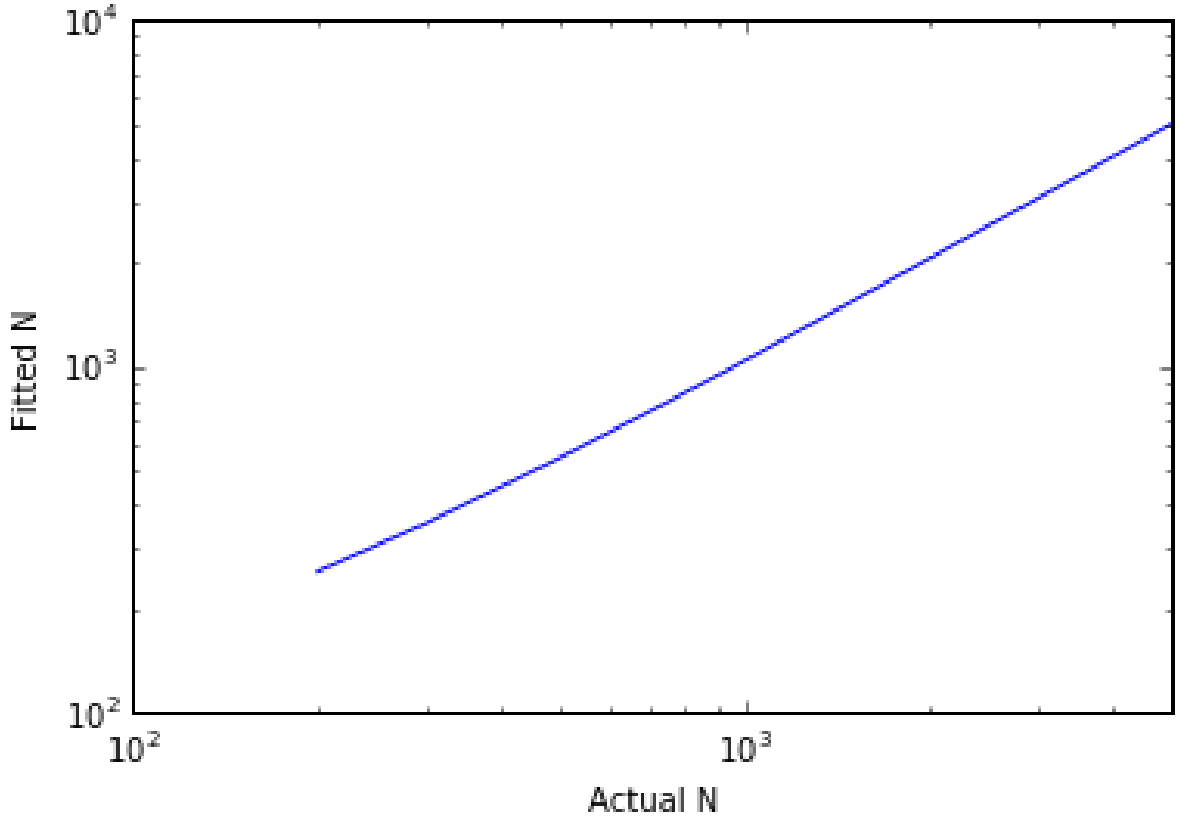
Fit and significances with $N = 1000$ signal at 3000 GeV
(Fitted signal of $N = 1042$, $\chi^2 = 57.89$ for 59 d.o.f.)





Results

Comparison of injected and fitted signals at 3000 GeV



Future Research

- More tests of effectiveness:
  - Test how well background-only fit will fit background+signal
  - Test on perturbations of data/other data sets
- Test against existing methods

Acknowledgements

I'd like to thank:

- Daniel Whiteson and University of California, Irvine for providing me with this project and mentorship
- University of Michigan for organizing this program
- National Science Foundation for funding my work

Getting Close with Fellow Summer Students