Statistical Methods for Experimental Particle Physics

Tom Junk

‡ Fermilab

Pauli Lectures on Physics
ETH Zürich
30 January — 3 February 2012

Day 5:   Statistical Tools and Examples

- Histogram Interpolation
- Integrating in Many Dimensions
- The "Punzi Effect"
- Optimizing MVA's

One Systematic at a Time?   (No! Need to simultaneously vary them all)

- Experimental uncertainties are typically evaluated one at a time
    - Additional MC and analysis burden to generate samples with varied parameters – must be done
    - Reweighting existing samples lessens the computational task. Sometimes this is more instructive anyhow (one can examine the weights to see if they make sense).
      Example – alternative PDF sets. Order 40 alternate samples but easy to reuse existing ones if we write the initial parton momenta into the event record.
    - Worse still, analyzers typically generate just ±1σ variations and extrapolate

- Varying more than one parameter at a time – Must be done to find the best fit in the nuisance parameter space or to integrate over the whole space.
    - Easy case: uncertain parameters affect the predictions multiplicatively
      Example: Luminosity, Lepton ID efficiency and B-tag Efficiency
       $R = R_0 * \Pi(1+\delta_i s_i)$ where δ_i is a fractional uncertainty due to the ith systematic uncertainty. s_i is the underlying uncertain parameter. It may affect several predictions with different impact. For example, the B-tat Efficiency affects single-tag events differently than double-tag events.
    - Nonlinearities must be estimated by analyzers – tools cannot know *a priori* about nonlinear effects or interactions between parameters. mclimit allows analyzers to specify a parameter as an arbitrary function of other parameters (say you care about the ratio of two parameters).
    - Need also to apply shape interpolations for multiple parameters at a time. Not totally trivial!

Histogram Interpolation

- Needed for several purposes

  - Finite grid of signal models are subject to Monte Carlo simulation
    example: Tevatron's m_H grid goes from 100 GeV to 200 GeV in 5 GeV steps
    What does a 117 GeV Higgs boson look like?
    How to fit for m_H with only a finite grid of MC?

  - Finite grid of nuisance parameter exploration.
    - Analyzers typically evaluate $\pm 1\sigma$ variations of their systematics
      We need to integrate over, or at least test all values
    - Need to figure out what happens when more than one nuisance
      parameter is varied at a time.

"Vertical" Interpolation

Linear interpolation within each bin.

Say the central value content of a bin is x (the zero-sigma variation)
For each nuisance parameter, we have a shape histogram with
varied contents v_i, where $i=1 \ldots N$ number of nuisance parameters.

Say the shape histogram corresponding to parameter i is a n_i-sigma variation.

We'd like a smeared shape corresponding to values of nuisance parameters
s_i, $i=1 \ldots N$

$$x_{var} = x + \sum_{i=1}^{N} s_i(v_i - x)/n_i$$

additive, so multiple sources of shape error are cumulative and commutative

- Not quite the formula – separate variations for s>0 and s<0, and the interpolator can start from a place other than s=0.
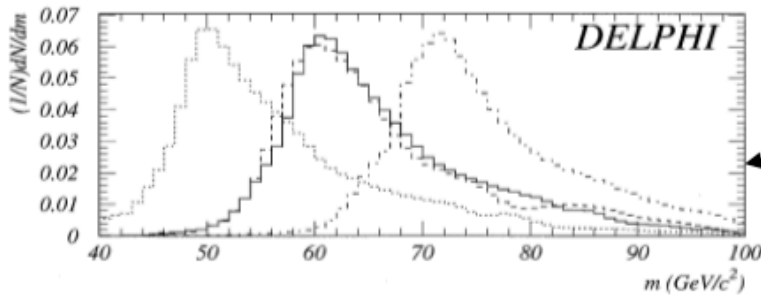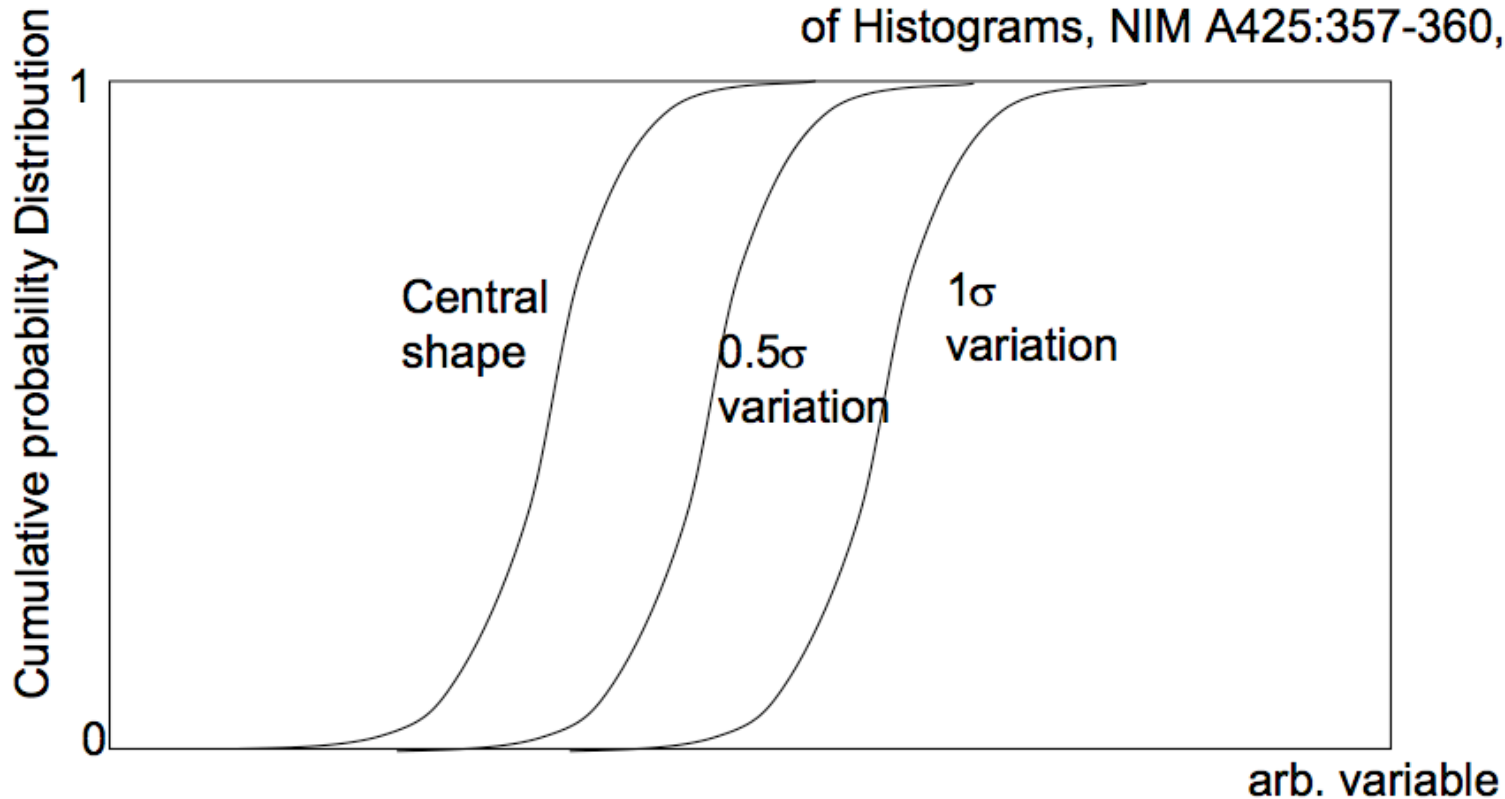
Good for: NN shape histograms, likelihood discriminants.

Not good for: Dijet invariant mass shapes under JES (move left and right).
Uncertainty in m_{top} when analyzing m_{lvb} for single top

"Horizontal" Interpolation

Template-morphing interpolation:

See A. Read: Linear Interpolation
of Histograms, NIM A425:357-360,1999.

Cumulative probability Distribution (y-axis, from 0 to 1)

Central
shape

0.5σ
variation

1σ
variation

arb. variable

DELPHI

$(1/N)dN/dm$ — 0.07, 0.06, 0.05, 0.04, 0.03, 0.02, 0.01, 0

40 50 60 70 80 90 100

$m \ (GeV/c^2)$

Differentiate to get histograms

Example: Interpolated with csm_interpolate_histogram

a component of mclimit_csm.C, .h
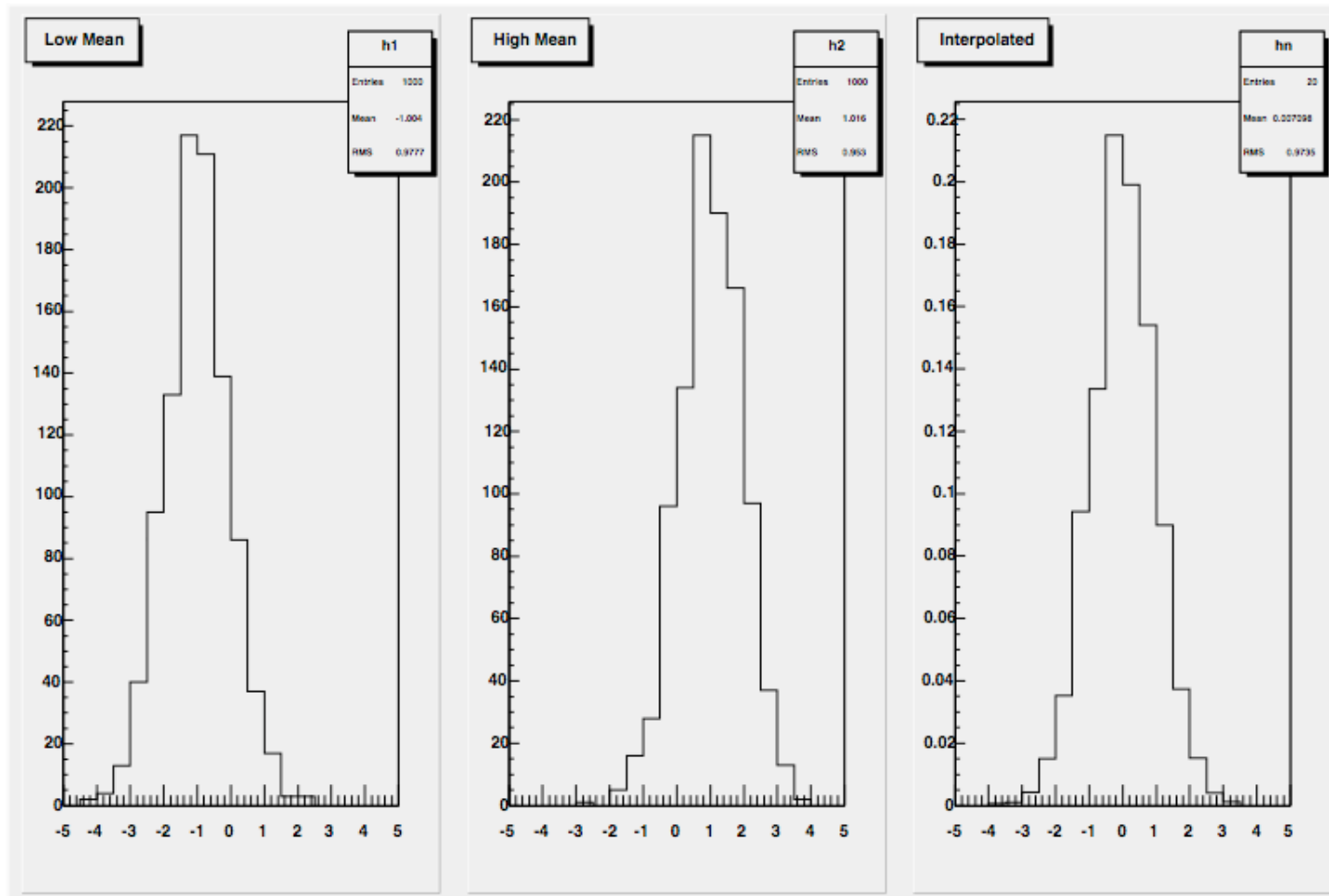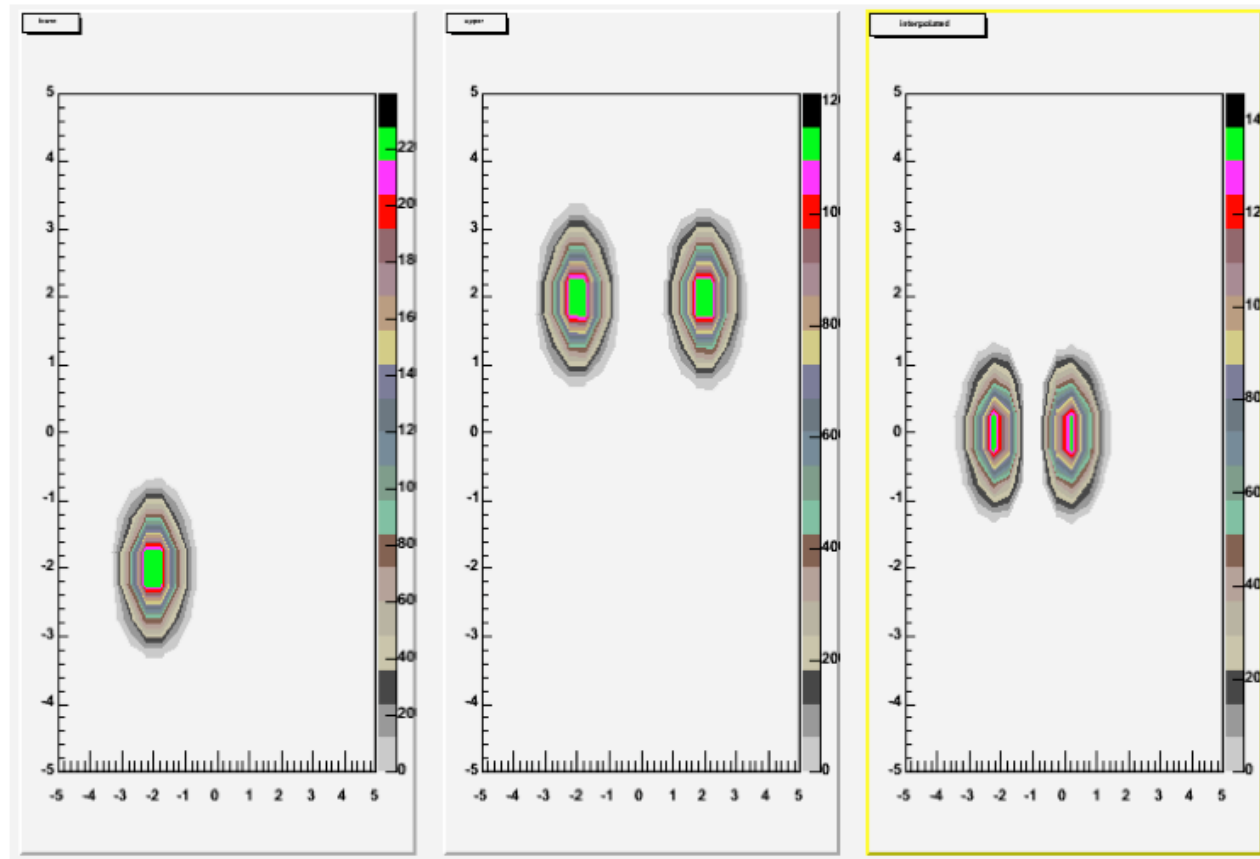

re-coded
in C++;
more robust
too.

Template: x=-1

Template x=+1

Interpolated Histogram
x=0

# csm_interpolate_histogram Works in 2D as well

Template x=-1

Template x=+1

Interpolated histogram
x=0

Many
thanks to
A. Read
for the
algorithm
example
d_pvmorph_2d

Shape Errors and Interpolation Issues

- Not all shapes are horizontally interpolatable. Example: NN outputs;
  two-peaked shapes, etc. Horizontal interpolation slides part of the PDF horizontally –
  you can get a third peak in the middle.

x=0   x=0.5   x=1
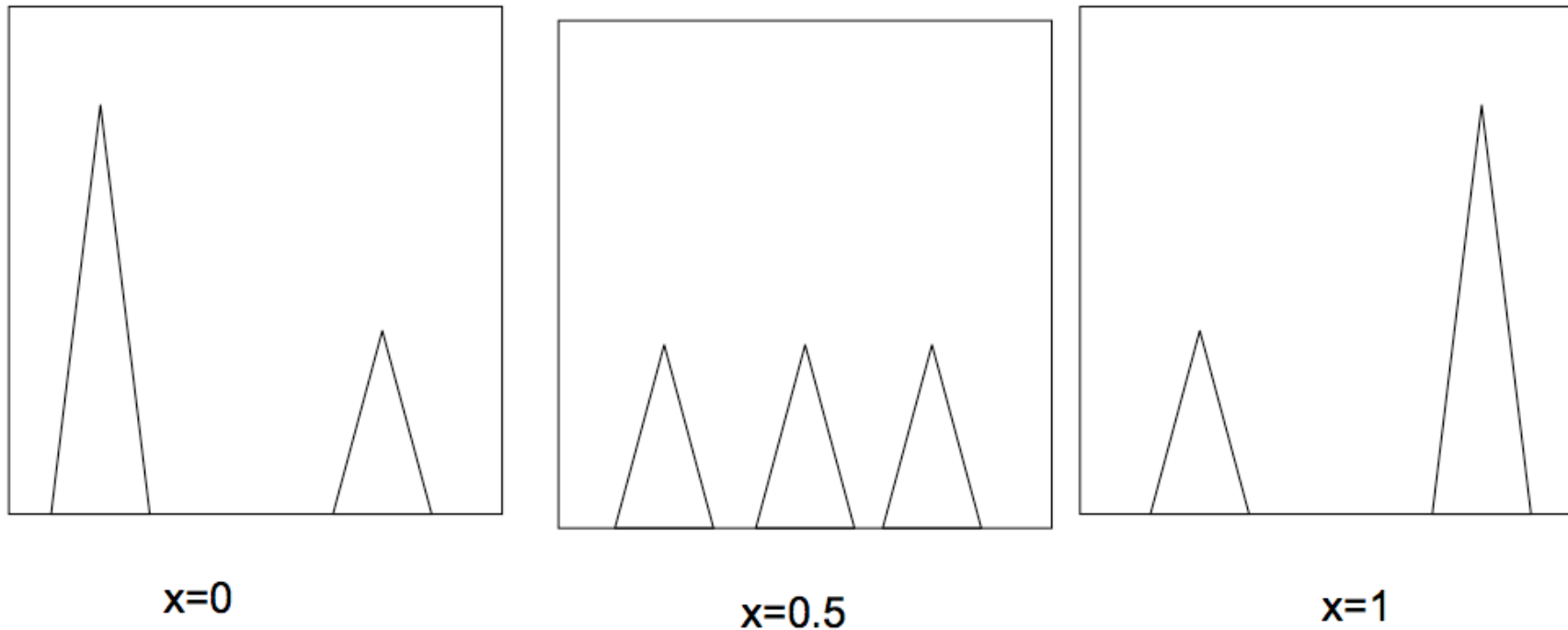
Can use "vertical" interpolation for these

Horizontal Interpolation Features and Warnings

You can extrapolate with this method too!

But watch out for histogram edges
min and max – the peak can wander
off the edge!

Also works for interpolating/extrapolating
the width of a peak.

But watch out – a peak can not have less
than zero width! Symptom of this –
the cumulative probability curve bends over
backwards.

"Horizontal" Interpolation

Template-morphing interpolation:

See A. Read: Linear Interpolation of Histograms, NIM A425:357-360,1999.

Differentiate to get histograms

Vertical interpolation by adding variations from different nuisance parameters was
commutative and cumulative. What is the equivalent for compounding several shape
distortions for horizontal interpolation? Say we want to distort the peak position and
the width independently.
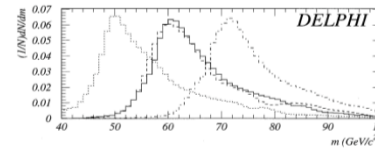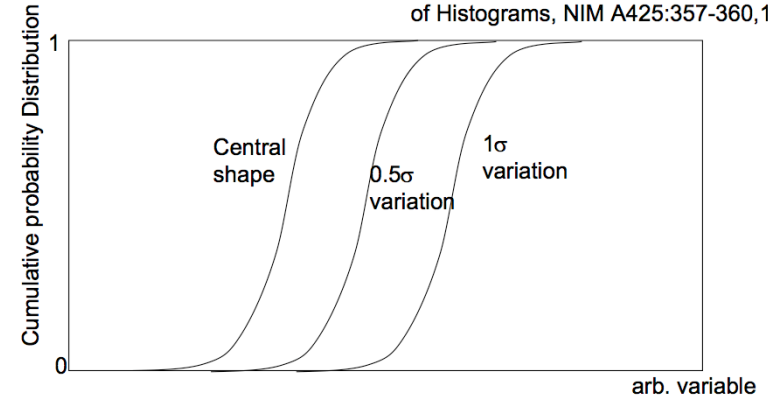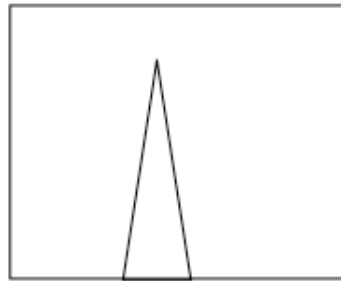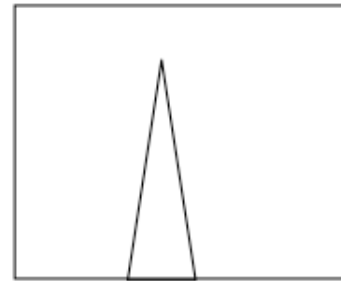
Problems of Compounded Shape Interpolations

Cartoon Example – two systematic shape variations giving similar shifted templates
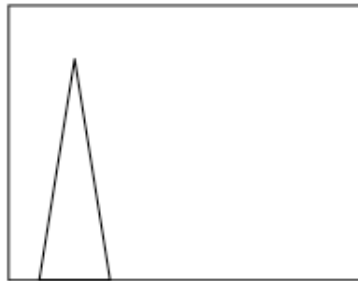
Central value histogram	+1σ systematic variation parameter #1 "JES$_1$"	+1σ systematic variation parameter #2 "JES$_2$"

Problem – What if you want a +1σ variation of *both* parameters, #1 and #2?

First interpolation – gets you to the varied template for parameter #1. Interpolate from there for parameter #2 and *nothing happens*.

What's really needed here is an extrapolation.

Shape Systematic Compounding Procedure in mclimit_csm.C

Old Strategy: Interpolate each shape systematic from the central value, and average the resulting interpolated histograms
- Commutative, at least
- Biased towards central shape: Add in a shape uncertainty which is small, and it averages in like the big shape uncertainties.

- Want: SAT-style "analogy" interpolation

A:B as C:D, where

A = central value
B = systematically distorted template
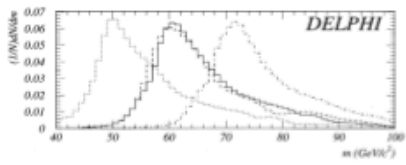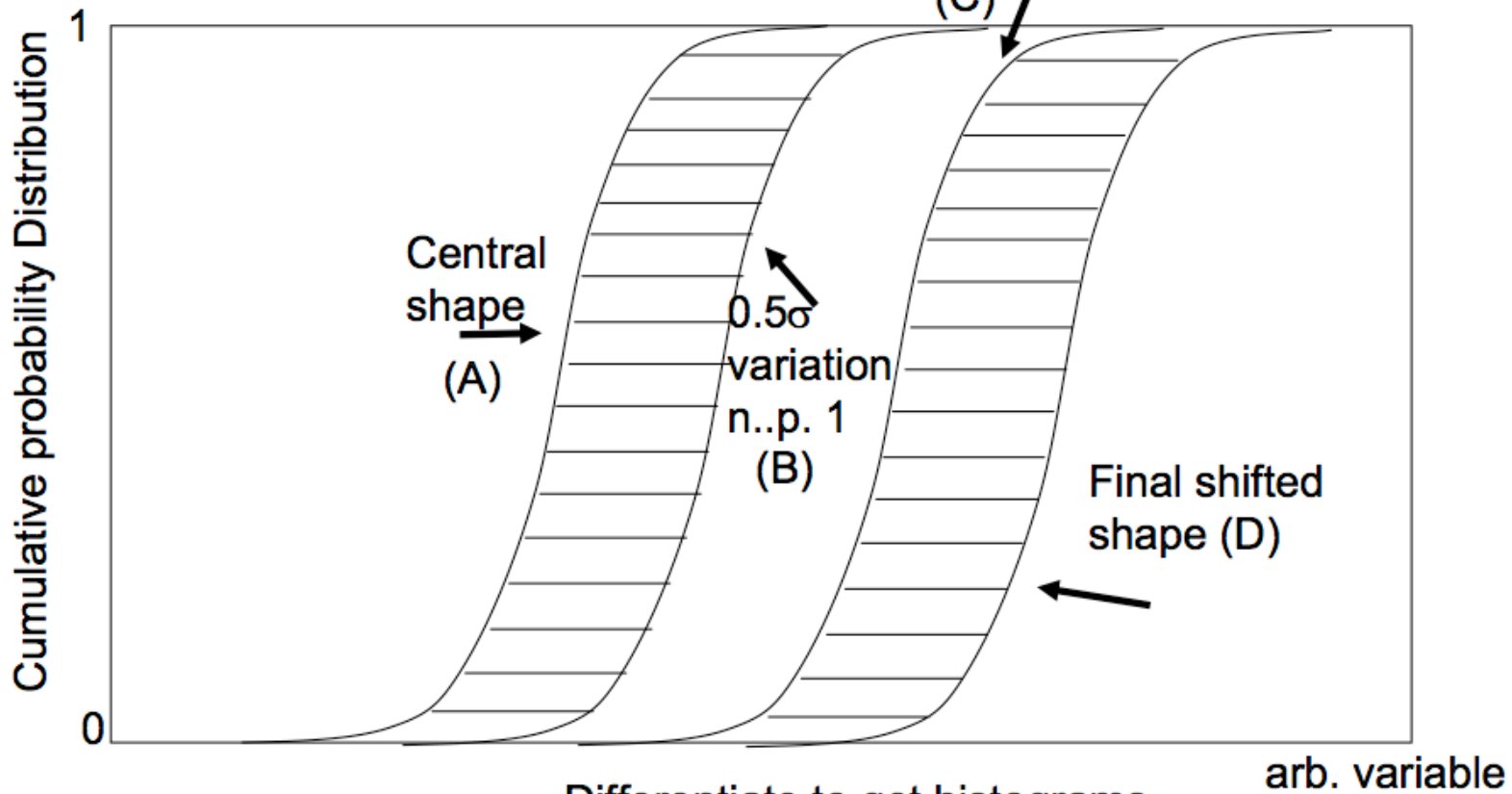C = result of previous distortions from other nuisance parameters
D = result of all distortions put together.

With one free parameter – how much to vary the nuisance parameter under consideration.

Shape Systematic Compounding Procedure in mclimit_csm.C

Template-morphing compounding:
(something like an extrapolation)

Result of
all other
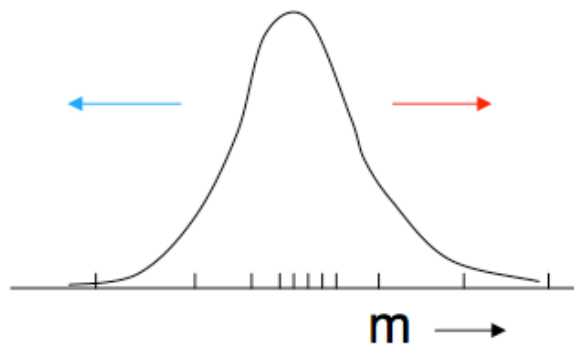n.p. interpolations
(C)

Cumulative probability Distribution

1

Central
shape
(A)

0.5σ
variation
n..p. 1
(B)

Final shifted
shape (D)

0

arb. variable

Differentiate to get histograms

DELPHI

Commutative

Commutative! Extrapolates!

Another Example – Resonance Peak Position and Width May Need Simultaneous Interpolation

$$y \sim \frac{\Gamma/2}{(m-M_0)^2 + (\Gamma/2)^2}$$

Vary M_0

Vary Γ
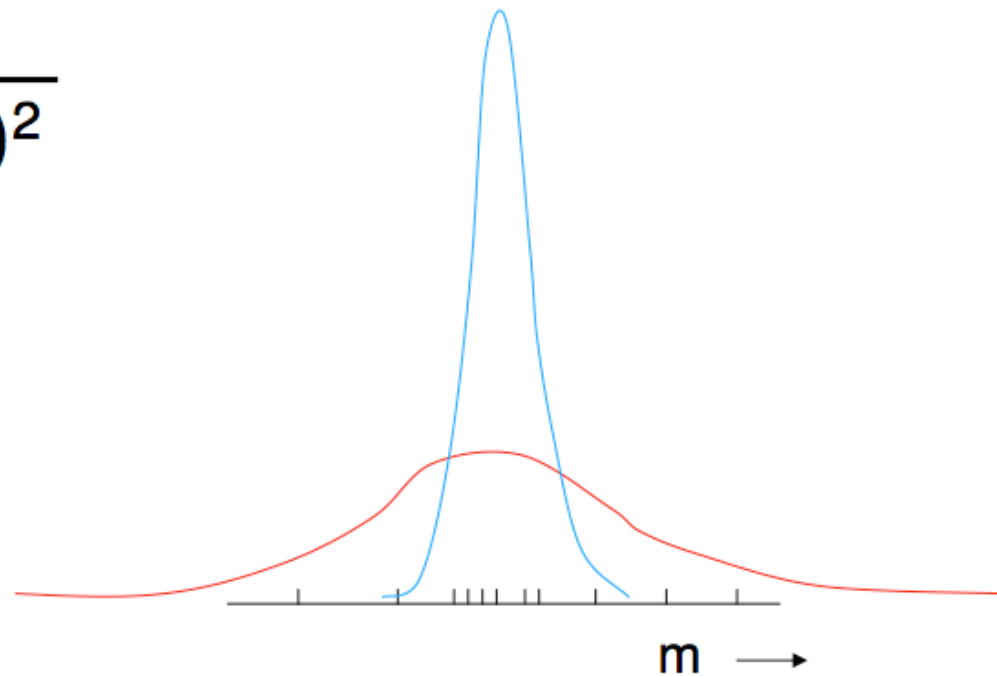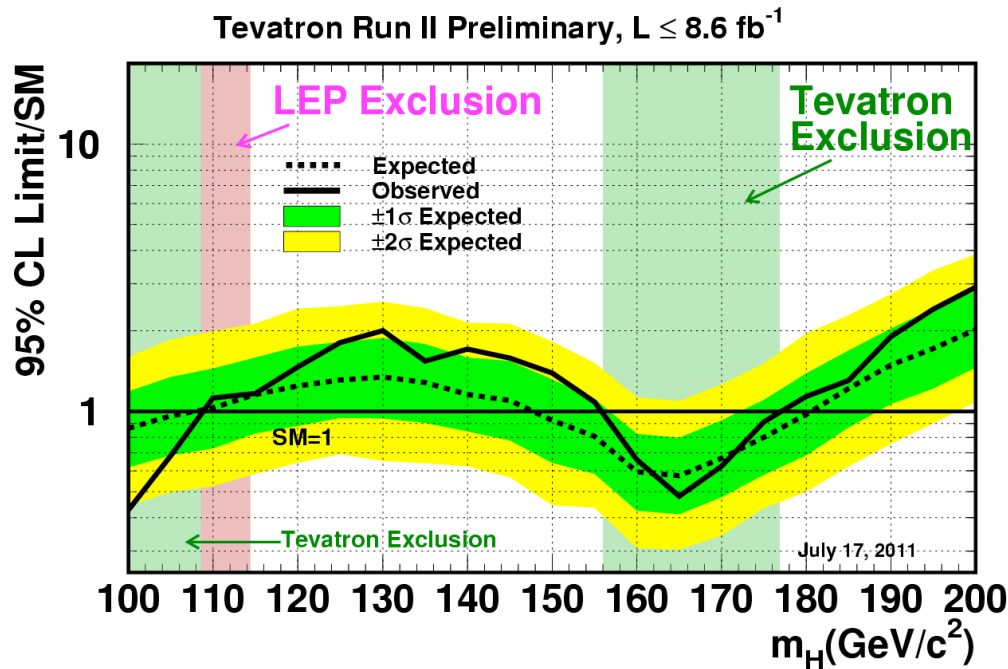
Frequently MC is generated with a fixed M_0 and several values of Γ, and with a fixed Γ and several values of M_0. Need a prediction for arbitrary M_0 and Γ. Compounded Horizontal Morphing is ideal for this case.

Question – Interpolating Searches with MVA's tuned up at each m_H?

Tevatron Run II Preliminary, L ≤ 8.6 fb⁻¹

A common question:
What allows us to draw straight lines on the observed limit plot?
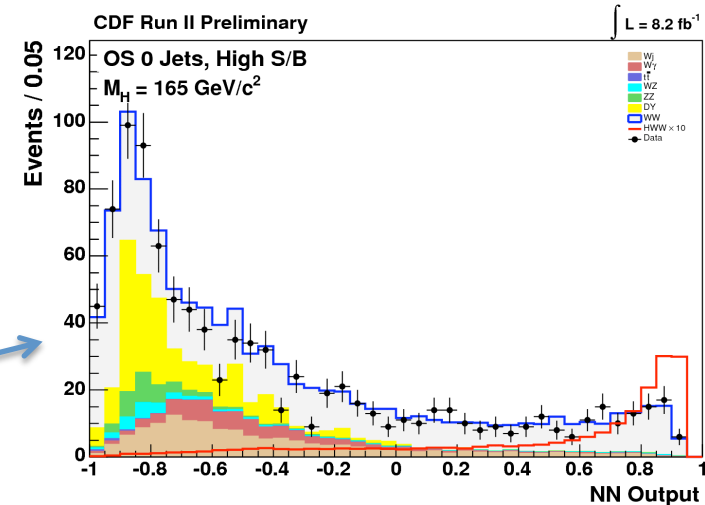
Why quote the m_H limits where these cross?

A typical MVA output. The MVA is trained to separate Higgs boson events from backgrounds at m_H=160 GeV. Similar shapes are obtained for m_H=155 and 165 GeV.

Interpolate signal and background prediction histograms to get, say, 157 GeV – dodgy, but you can test it by interpolating 155 and 165 to get 160's and compare.

The problem lies in interpolating the data. You can follow individual events' NN outputs vs. m_H but interpolating them on average biases the most significant ones down. Makes us uncomfortable – applying a different procedure to data and MC.

A Review of Setting Bayesian Limits and Measuring Quantities

Including uncertainties on nuisance parameters θ

$$L'(data \mid r) = \int L(data \mid r, \theta) \pi(\theta) d\theta$$

where $\pi(\theta)$ encodes our prior belief in the values of the uncertain parameters. Usually Gaussian centered on the best estimate and with a width given by the systematic. The integral is high-dimensional

Typically $\pi(r)$ is constant Other options possible.
Sensitivity to priors a concern.

Useful for a variety of results:

Limits:

$$0.95 = \frac{\int\limits_0^{r_{lim}} L'(data \mid r) \pi(r) dr}{\int\limits_0^{\infty} L'(data \mid r) \pi(r) dr}$$

CDF Run II Preliminary, L=3.6 fb^{-1}

Posterior Density = L'(r)×π(r)

Observed Limit

Mean 0.5284
RMS 0.4487

m_H=160

5% of integral

σ*BR/SM =r
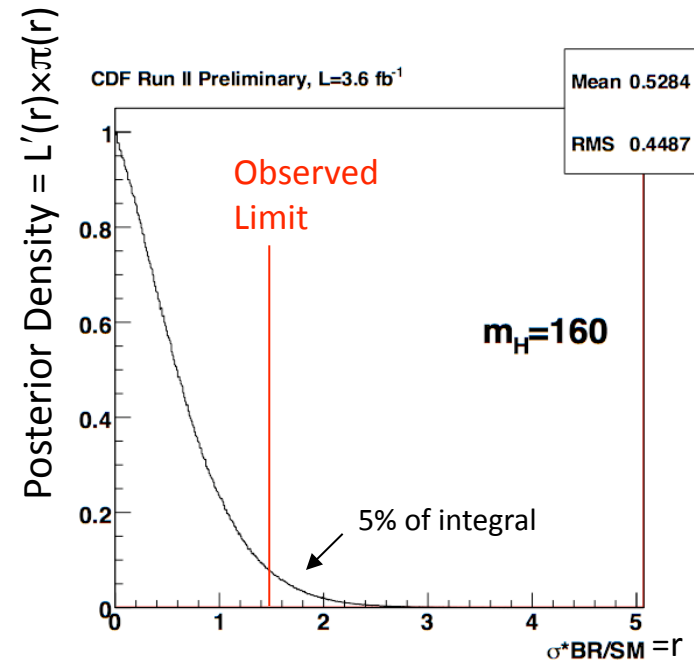
Reminder: Bayesian Cross Section Extraction

Same handling of nuisance parameters as for limits

$$L'(data \mid r) = \int L(data \mid r, \theta) \pi(\theta) d\theta$$

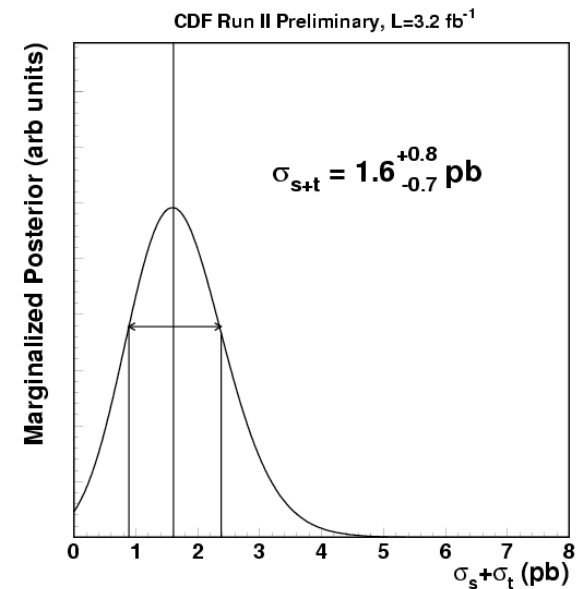$$0.68 = \frac{\int_{r_{low}}^{r_{high}} L'(data \mid r) \pi(r) dr}{\int_0^{\infty} L'(data \mid r) \pi(r) dr}$$

The measured cross section and its uncertainty

$$r = r_{\max}{}^{+(r_{high} - r_{\max})}_{-(r_{\max} - r_{low})}$$

Usually: shortest interval containing 68% of the posterior (other choices possible). Use the word "credibility" in place of "confidence"

If the 68% CL interval does not contain zero, then the posterior at the top and bottom are equal in magnitude.
The interval can also break up into smaller pieces! (example: WW TGC@LEP2

CDF Run II Preliminary, L=3.2 fb⁻¹

$\sigma_{s+t} = 1.6^{+0.8}_{-0.7}$ pb

Marginalized Posterior (arb units)

$\sigma_s + \sigma_t$ (pb)

Integration in Many Dimensions

Even calculation of p-values in the CL_s method (CL_{s+b} and $1-CL_b$) are high-dimensional integrations over the space of nuisance parameters and the space of possible experimental outcomes.

There are two techniques that I use that are easy to program:

- Scattershot – sample all variables to be integrated over from their (uncorrelated) priors. Impacts of nuisance parameters on predictions may be shared, correlating the predictions, but the parameters themselves should be designed to be independent from each other.

- Markov Chain – The Metropolis-Hastings algorithm (excellent article on Wikipedia)

 -- Pick a point in parameter space. Propose a next step in parameter space from a symmetric proposal function. Make the step if L(new)/L(old)>a randomly chosen number between 0 and 1 (not including 1). Make a histogram of parameter i's values as you go. This is the distribution of parameter i integrated over all the other parameters.

A Problem with Scattershot Integration

CDF Run II Preliminary, L=3.2 fb⁻¹
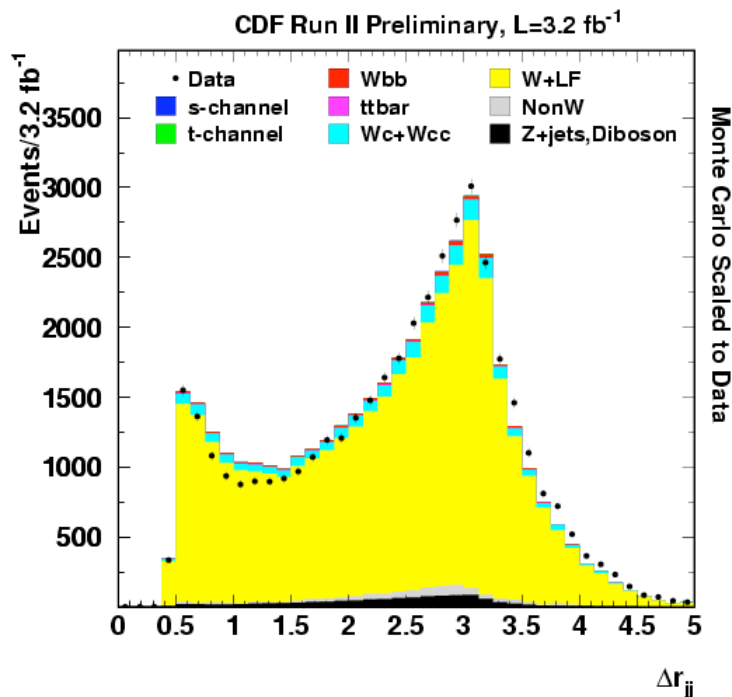
$$L'(data \mid r) = \int L(data \mid r, \theta)\pi(\theta)d\theta$$

Suppose we have a large *a priori* uncertainty on the normalization of the yellow background template (say ±30%) Most samplings from this prior distribution will "miss" the data by a lot, contributing a vanishingly small amount to the integral. You need many more samples to get the integral to converge well.

This problem becomes exponentially hard if there are more channels being combined in joint likelihood, and the sampling of the nuisance parameters must predict the data in all channels simultaneously well.

It would be nice to have a method that samples the peaks in L more than the large spaces where it (almost) vanishes.

A Metropolis-Hastings Example – Three Markov Chains Exploring the Same Space
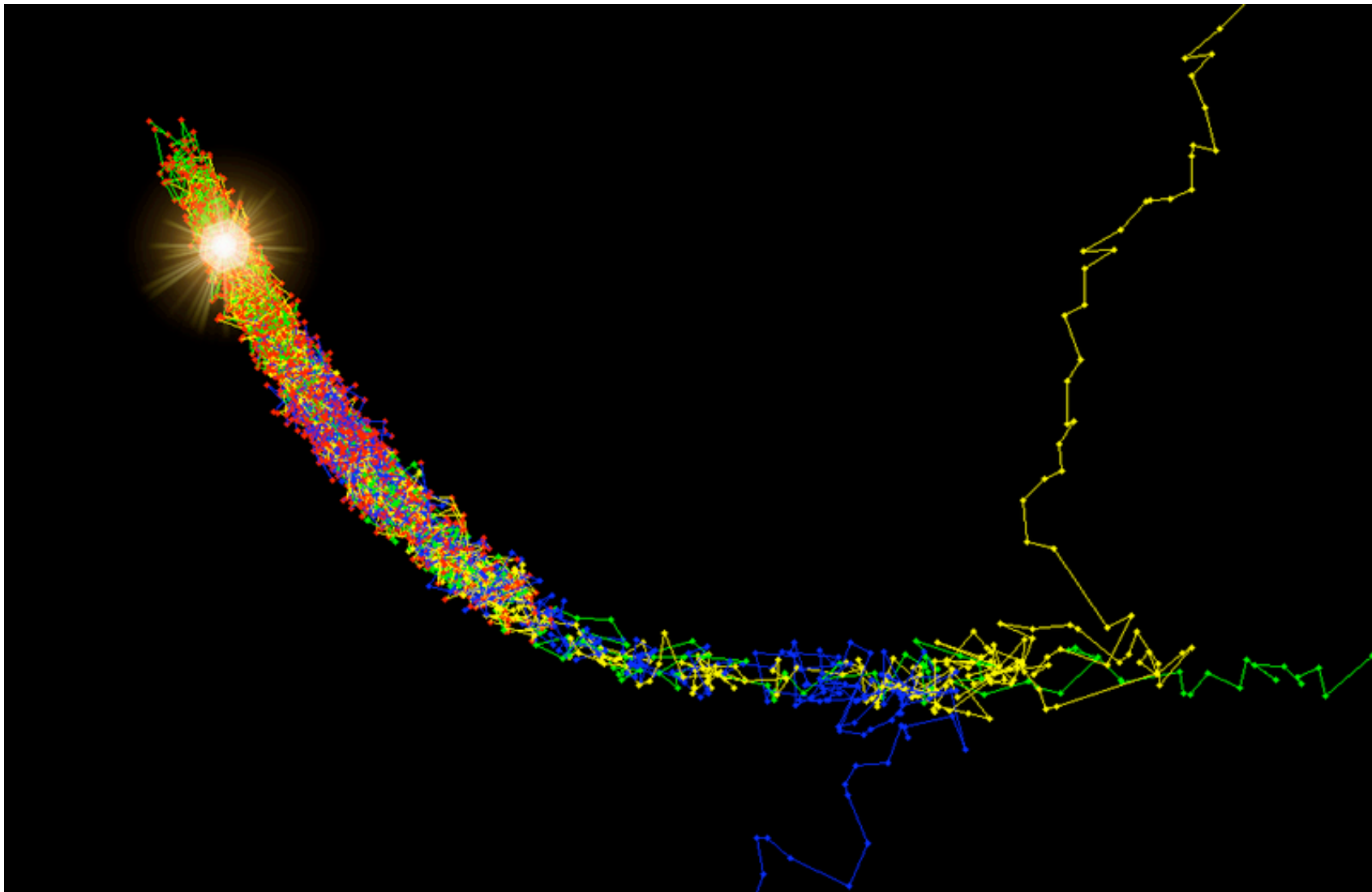

From Wikipedia

Checking the Consistency of Highly Correlated Analyses

- Frequently arises in large collaborations focusing on a small number of high-priority measurements.

- Analysis teams select highly overlapping data samples. Ideally, overlap should be zero or 100%, but if the teams do not communicate well, the overlap can be partial.

- In the case of zero overlap, to produce a combined result, just treat the results as if they were different channels (separate final states). Independent outcome probabilities, so consistency can be determined by computing $\Delta x / \sqrt{\sigma_1^2 + \sigma_2^2}$ taking out shared systematic uncertainties from the measurement uncertainty.

- In the case of 100% overlap, you can use BLUE, or a super-discriminant (an MVA built on MVA oututs). Example: CDF Single Top Observation, Phys. Rev. D **82**, 112005 (2010). But you should check consistency.
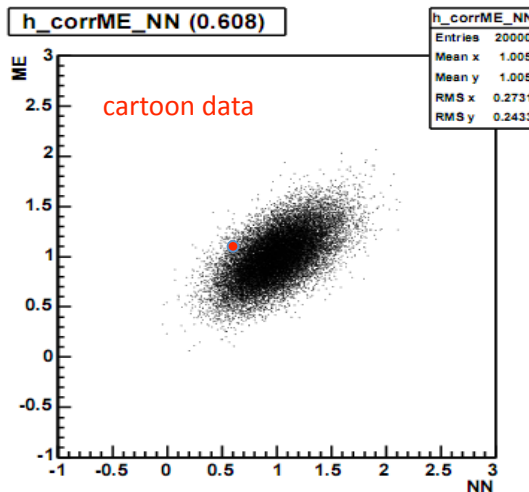
Checking the Consistency of Highly Correlated Analyses

Example – 100% selected event overlap, different analysis techniques.
CDF's 2.2 fb^{-1} single top analysis "Data" points are cartoons for illustration only.
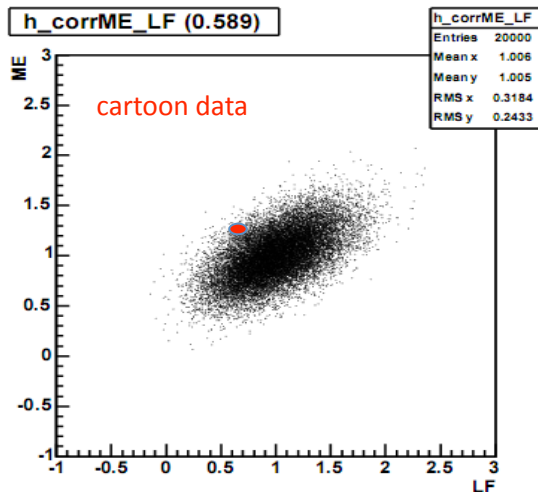
LF-ME 58.9%   ME-NN 60.8%   LF-NN 74.1%
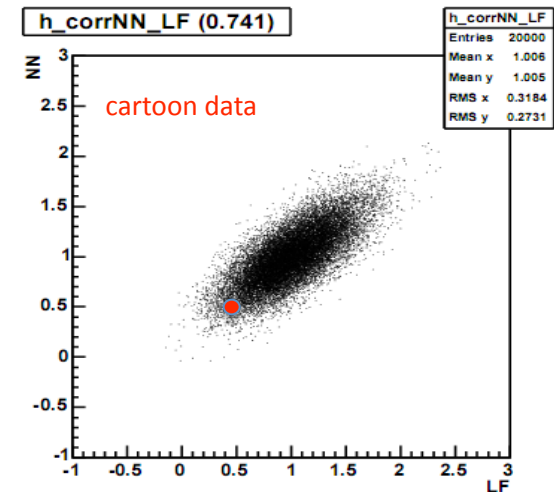
- Can define a p-value for how discrepant analysis results are.
- Above are distributions of possible cross section measurements in three highly correlated analyses considered pairwise.
- As usual, the ensemble depends on assumptions. In this case, we assumed the SM production cross section for the signal.
- Suggestion – use the maximum $|\Delta x|$ to reduce sensitivity to the model assumptions. (they won't entirely go away thourgh).

Measured Uncertainties and the Punzi Effect

- Reconstruction algorithms typically supply also an "uncertainty" on reconstructed parameters.

- At some point, the uncertainties should be checked for proper pulls:
  $(q-q_{inject})$/uncertainty^2 should be a unit-width Gaussian centered on zero.

- But not all distributions are Gaussian. Some have tails that carry physics information.

The tails may be a mixture of physics and and misreconstruction.

d0 from light quark jets (other taus in evts with taus matching MC)

jet_d0	
Entries	11988
Mean	0.006378
RMS	0.01347
Underflow	0
Overflow	558
Integral	1.143e+04
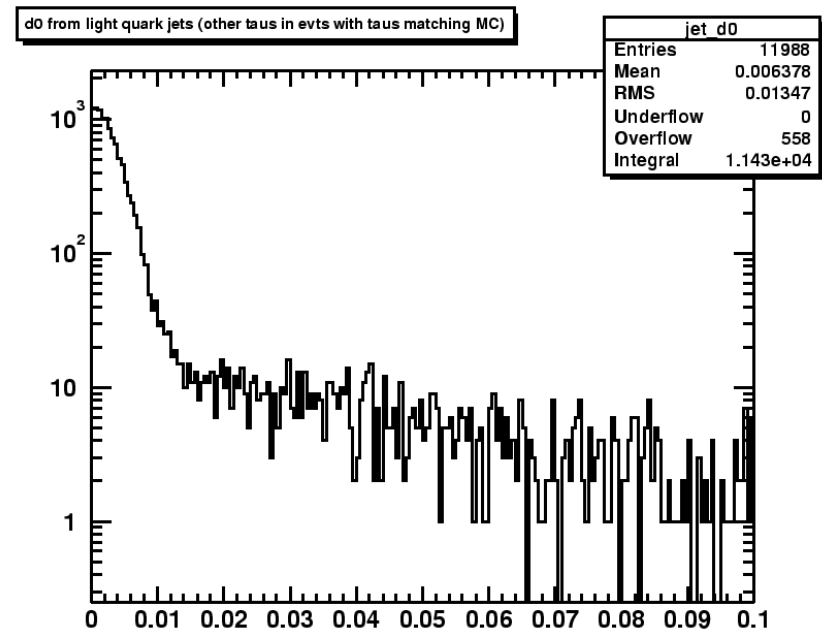
Getting $\mathcal{L}$ wrong: Punzi effect

Giovanni Punzi @ PHYSTAT2003
"Comments on $\mathcal{L}$ fits with variable resolution"

Separate two close signals, when resolution σ varies event by event, and is different for 2 signals

e.g. 1) Signal 1 $1+\cos^2\theta$
       Signal 2   Isotropic
       and different parts of detector give different σ

    2) M (or τ)
       Different numbers of tracks $\rightarrow$ different σ_M (or σ_τ)

L. Lyons

Events characterised by x_i and σ_i

A events centred on $x = 0$

B events centred on $x = 1$

$$\mathcal{L}(f)_{wrong} = \Pi\,[f * G(x_i,0,\sigma_i) + (1-f) * G(x_i,1,\sigma_i)]$$

$$\mathcal{L}(f)_{right} = \Pi\,[f*p(x_i,\sigma_i;A) + (1-f) * p(x_i,\sigma_i;B)]$$

$$p(S,T) = p(S|T) * p(T)$$

$$p(x_i,\sigma_i|A) = p(x_i|\sigma_i,A) * p(\sigma_i|A)$$

$$= G(x_i,0,\sigma_i) * p(\sigma_i|A)$$

So

$$\mathcal{L}(f)_{right} = \Pi[f * G(x_i,0,\sigma_i) * p(\sigma_i|A) + (1-f) * G(x_i,1,\sigma_i) * p(\sigma_i|B)]$$

If $p(\sigma|A) = p(\sigma|B)$, $\mathcal{L}_{right} = \mathcal{L}_{wrong}$

but NOT otherwise

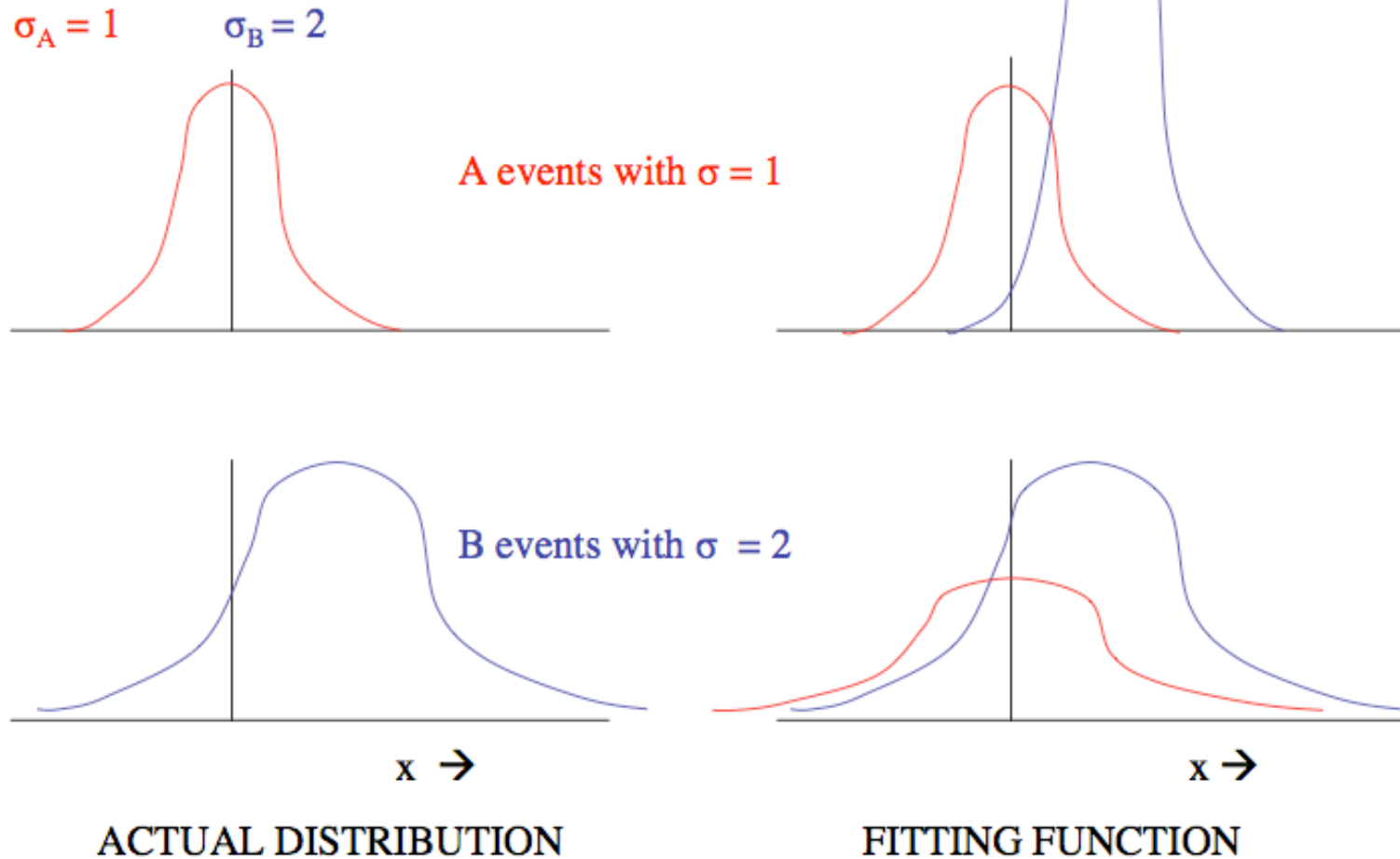L. Lyons

Giovanni's Monte Carlo for   A : $G(x, 0, \sigma_A)$

B : $G(x, 1, \sigma_B)$

$f_A = 1/3$

		$\mathcal{L}_{wrong}$		$\mathcal{L}_{right}$	
σ_A	σ_B	f_A	σ_f	f_A	σ_f
1·0	1·0	0·336(3)	0·08	Same	
1·0	1·1	0·374(4)	0·08	0·333(0)	0
1·0	2·0	0·645(6)	0·12	0·333(0)	0
1 → 2	1.5 → 3	0·514(7)	0·14	0·335(2)	0·03
1.0	1 → 2	0.482(9)	0.09	0.333(0)	0

1) $\mathcal{L}_{wrong}$ OK for $p(\sigma_A) = p(\sigma_B)$, but otherwise BIASSED

2) $\mathcal{L}_{right}$ unbiassed, but $\mathcal{L}_{wrong}$ biassed (enormously)!

3) $\mathcal{L}_{right}$ gives smaller σ_f than $\mathcal{L}_{wrong}$

L. Lyons, G. Punzi

Explanation of Punzi bias

$\sigma_A = 1$ $\sigma_B = 2$

A events with $\sigma = 1$

B events with $\sigma = 2$

x →

x →

ACTUAL DISTRIBUTION

FITTING FUNCTION

[N_A/N_B variable, but same for A and B events]

Fit gives upward bias for N_A/N_B because (i) that is much better for A events; and

(ii) it does not hurt too much for B events

L. Lyons

Avoiding Punzi Bias

A reconstructed "uncertainty" is an observable!
Treat it like any other reconstructed quantity.

BASIC RULE:
Write pdf for ALL observables, in terms of parameters

- Include $p(\sigma|A)$ and $p(\sigma|B)$ in fit
  (But then, for example, particle identification may be determined more by momentum distribution than by PID)

OR

- Fit each range of σ_i separately, and add $(N_A)_i \rightarrow (N_A)_{total}$, and similarly for B

Incorrect method using $\mathcal{L}_{wrong}$ uses weighted average of $(f_A)_j$, assumed to be independent of j

Talk by Catastini at PHYSTAT05

L. Lyons

The ALEPH Tau Neutrino Mass Constraint

- Tau lepton decays involve at least one neutrino – sometimes two in the case of a charged lepton in the decay.
- Constraint on m_ν is better if the invariant mass of the visible decay products is large. Initial tau momentum is not perfectly known (the recoiling tau in Z decay has its own neutrino(s)).
- All four experiments at LEP attempted this constraint, but ALEPH got very lucky and observed a tau lepton decay with five charged tracks in it with very little invisible mass. Did not expect on average to get this lucky.

We learned more from the lucky data than we would have on average. Seems okay this time.

**Eur.Phys.J. C2 (1998) 395-406**
Phys.Lett. B349 (1995) 585-596

Banff Challenge 2: Parametric and Nonparametric Discovery Issues

A simple "mock data challenge"

http://www-cdf.fnal.gov/~trj

And also the associated presentations and writeup for PHYSTAT2011.

Several groups supplied solutions to the task of detecting small signals on large, uncertain backgrounds, with varying degrees of success.
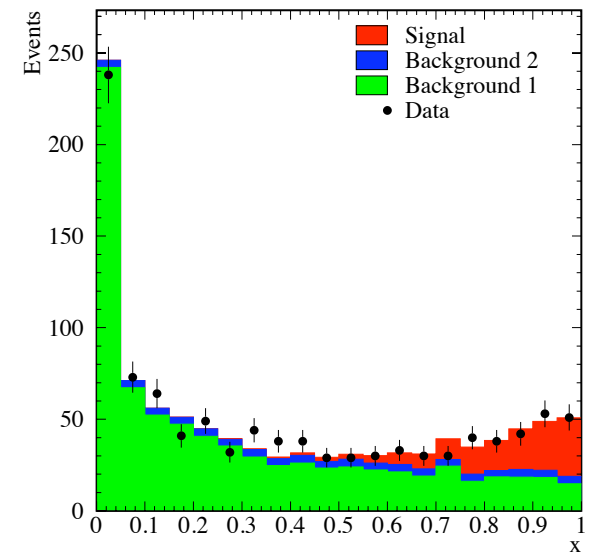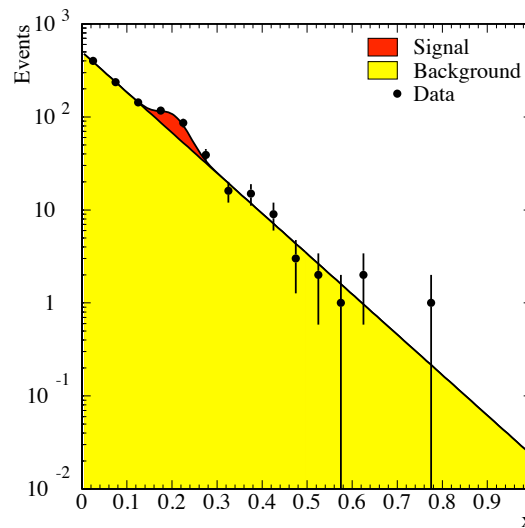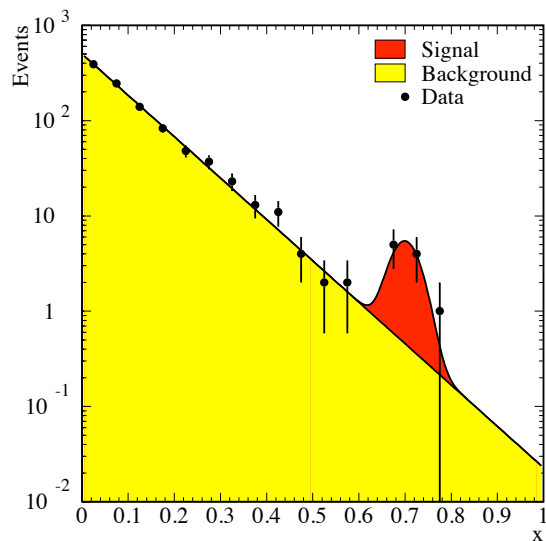
Optimizing MultiVariate Analyses

- Many choices of MVA's available:
  - Neural Networks
  - Bayesian Neural Networks (Radford Neal's work and others)
  - Matrix Element-based discriminants
  - Decision Trees (J. Friedman), and Boosted Decision Trees (common now)
  - Support Vector Machines
  - Likelihood Functions ("naive Bayes")
  - K-Nearest-Neighbors
  - Pick the right variable(s)! Input variable choice is usually broader than the choice of MVA method. An MVA output variable really is just another reconstructed quantity for each event, whose modeling has to be checked. Maybe there really is only one variable with all the s/b separation power. If you happen to know what it is, then there's no need for machine learning!

    Frequently though, even though signal and background may differ in a theoretically tractable way, the detector, trigger, and data selection requirements, and "instrumental" backgrounds usually mean multiple variables will still carry useful information.

- I won't describe these in any detail (subject of next week's lectures)

Optimizing MultiVariate Analyses

- Just what is the figure of merit? What is optimized?
    - Neural networks: Sum of squares of classification errors.
       Chosen for easy back-propagation to compute derivatives with respect to weights
    - Boosted Decision Trees: The Gini Coefficient purity*(1-purity).

- Which one is the best? Answer: We don't care about the sum of squares of classification errors, or the Gini Coefficient! We care about
    - Median Expected Limit if a signal is absent
    - Median Expected p-value if a signal is present
    - Median Expected Measurement Uncertainty if we are making a measurement

- These are the things we should optimize – in fact, they should drive most of the choices we make as experimentalists
    - Which MVA to use
    - Which variables to put in it
    - Which analysis and trigger requirements to place
    - Which accelerator and experiment to build

Optimizing MultiVariate Analyses

- Optimization of these things is a bit tedious, but worth it.
- I take issue with a statement in the D0 single top evidence PRD:
  D0 Collab., Phys. Rev. D **78**, 12005 (2008).

Optimal event discrimination is a well-defined problem with a well-defined and unique solution. Given the probability

$$p(S|\mathbf{x}) = \frac{p(\mathbf{x}|S)p(S)}{p(\mathbf{x}|S)p(S) + p(\mathbf{x}|B)p(B)} \quad (5)$$

that an event described by the variables $\mathbf{x}$ is of the signal class, S, the signal can be extracted optimally, that is, with the smallest possible uncertainty [53], by weighting events with $p(S|\mathbf{x})$, or, as we have done, by fitting the sum of distributions of $p(S|\mathbf{x})$ for signal and background to data, as described in Sec. XVIII. In practice, since *any* one-to-one function of $p(S|\mathbf{x})$ is equivalent to $p(S|\mathbf{x})$, it is sufficient to construct an approximation to the discriminant

$$D(\mathbf{x}) = \frac{p(\mathbf{x}|S)}{p(\mathbf{x}|S) + p(\mathbf{x}|B)} \quad (6)$$

built using equal numbers of signal and background events, that is, with $p(S) = p(B)$. Each of the three

Two problems with this overly optimistic appraisal:

1) Systematic Uncertainties on the signal and background rates and shapes

2) Binning (which really is just case 1 for finite MC or data sidebands)

Clearly not the case if a measurement is systematics dominated!

An Example MVA – CDF's Single Top Analysis with Matrix Elements

Main backgrounds: Wbb, Wcc+Wcj, W+LF, ttbar, Z+jets, diboson, multijets
Discriminant does a great job separating single top signal from the backgrounds.
It is not optimized to separate one background from another, however!
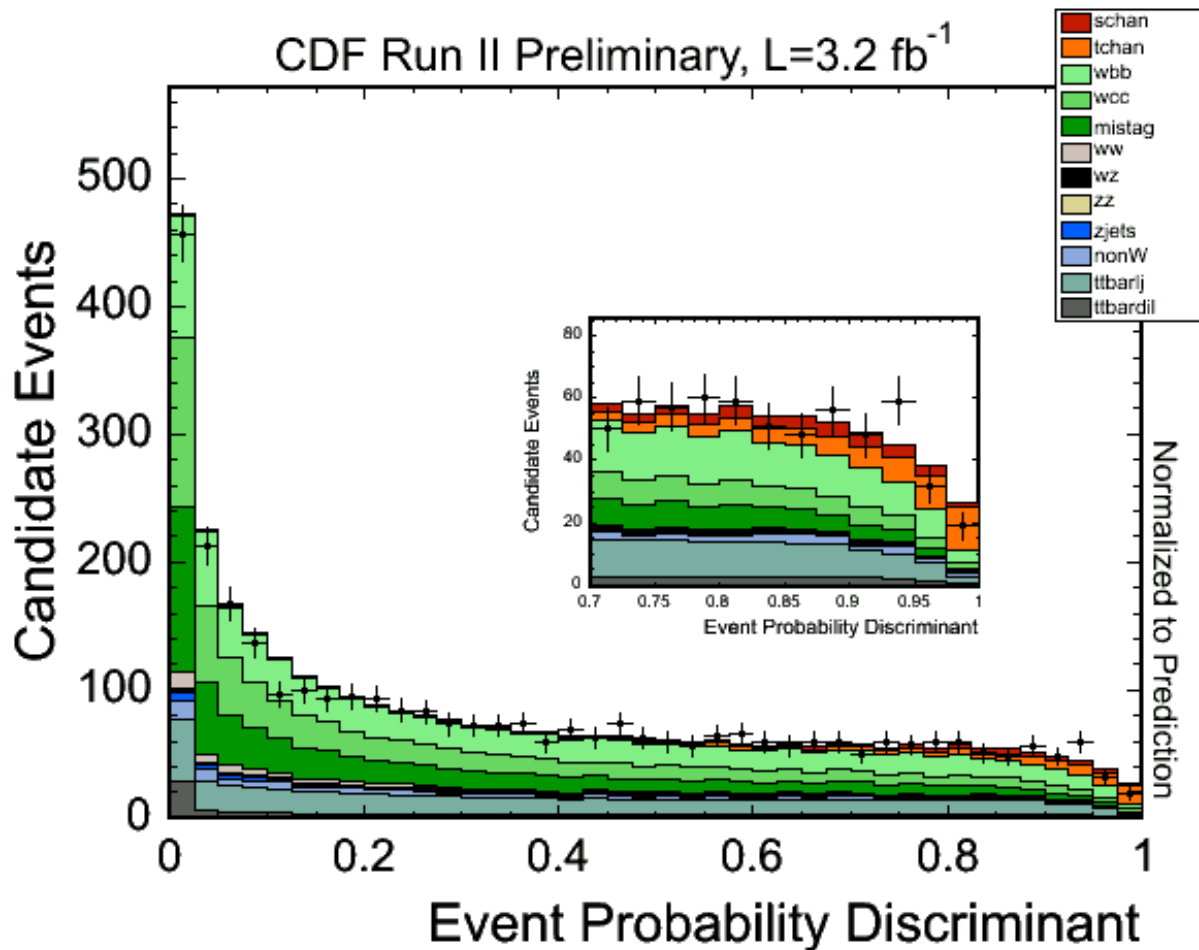

High-score bins provide sensitivity to test for the signal.

Low-score bins help constrain backgrounds

Extrapolation of background constraints to the signal region requires knowledge of shapes (and inclusion of shape uncertainties!)

Different backgrounds have different shapes – analysis is more optimal if these can be fit separately!

CDF's ZH→llbb Analysis Strategy
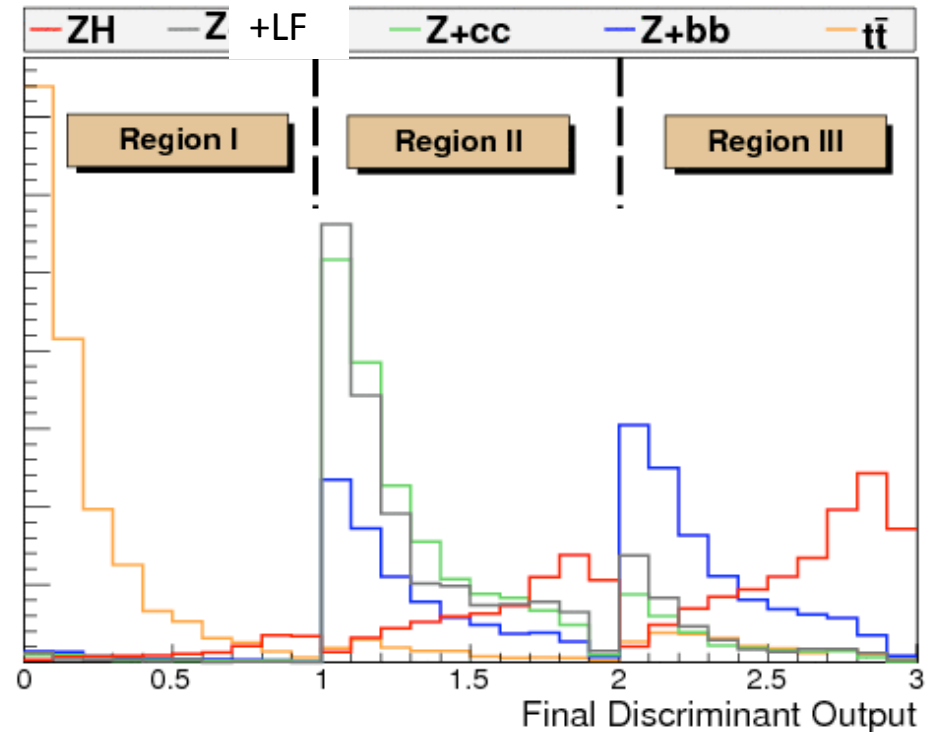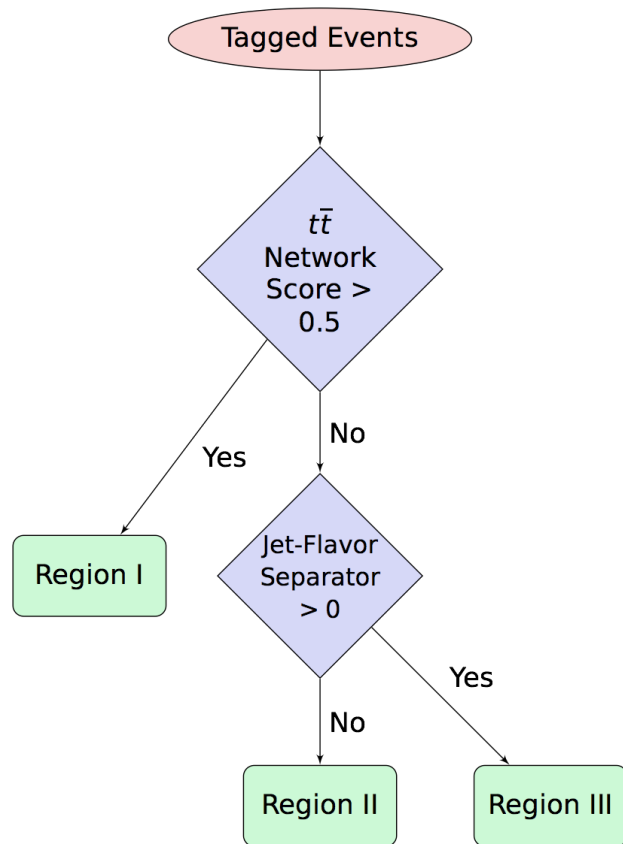
- Select events with Z→ll + jets with as loose a lepton selection as possible – still quite pure in Z decays.
- Train NN's first against ttbar, then to separate out the different flavor Z+jet samples (Zcc, Zbb, Z+LF (mistag)).

CDF's ZH→llbb Output Disriminants – Electron Channels
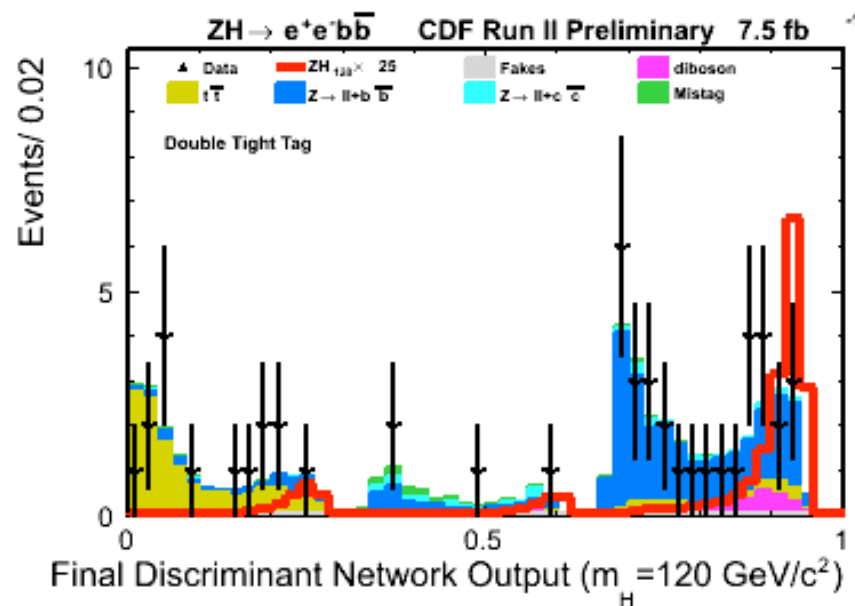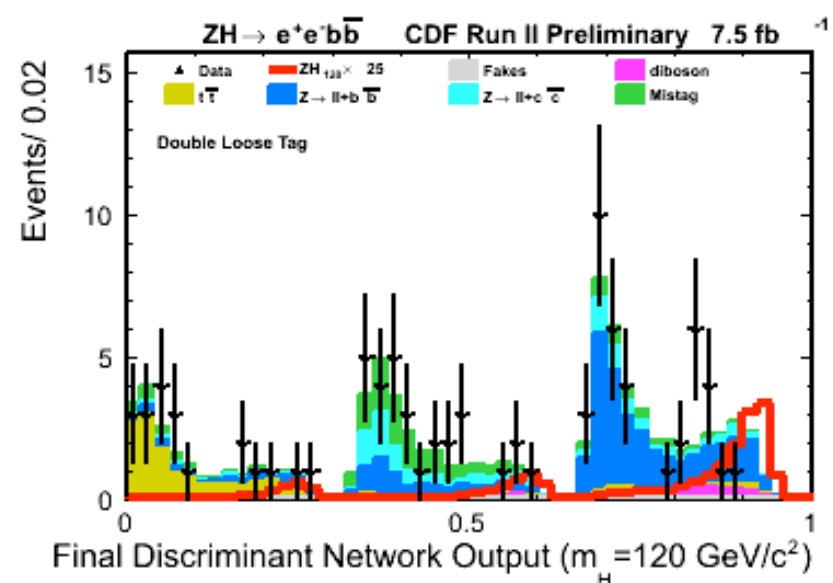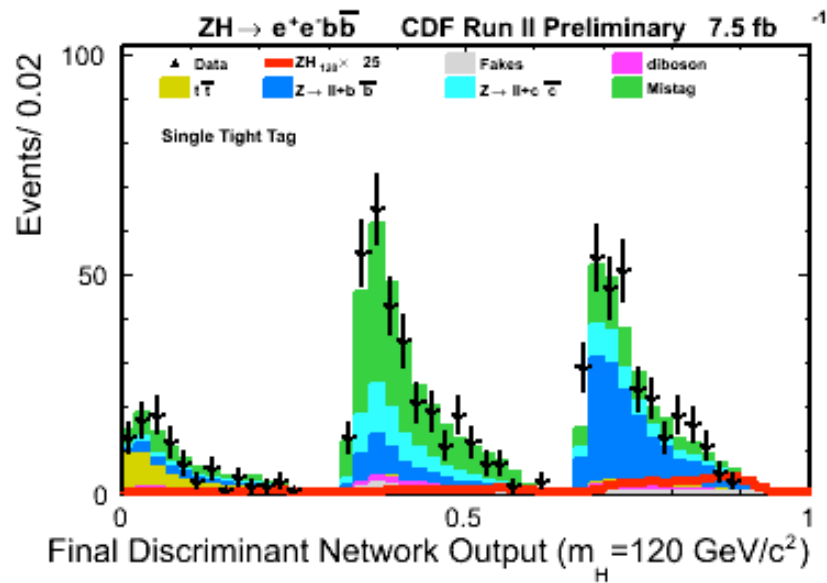
You Can Make a Discovery with Just One Observed Event

But it takes 3 expected signal events to exclude!

H_{null}= Bear rate=0. H_{test}= Bear rate > 0.   p-value is *almost* zero.

Some contributions to the expected background rate:
• People dressing as grizzly bears   (good selection requirements can reduce this background)
• Cardboard cutout pictures of grizzly bears
• Digital photograph manipulation

Each background source needs some kind of prior, or auxiliary measurement if possible.
There is also not much skepticism about the discovery claim.

From a CNN story on January 14, 2012

Okay, it was written by a comedian, and yes, it's a joke, but the statistics are obviously messed up.

Yes, some will scoff at Colbert running ahead of Huntsman -- a candidate running below the margin of error in some polls, meaning he may have zero support or may actually owe votes -- but keep in mind that in the recent Iowa caucus, Huntsman received 745 votes.

Dean Obedeillah, 2012

"Polling below the margin of error" usually just means scornfully low ratings for a politician. But it illustrates the lack of usefulness of

measured value/error

as a significance guess. There of course are positive Huntsman supporters.

Extensions of Banff Challenge 1

$n_{off} * \tau$ as an estimate of b

n_{on} is the measurement in the signal region, with an estimated signal acceptance of ε. Given n_{off}, τ, ε, n_{on}, set a limit on the signal rate s (where $s\varepsilon$ is the expected signal yield and b is the background yield)

1) Usually there are multiple background sources b_1 ... b_n
2) Often there's more than one kind of signal, too. And they don't have to scale together (multidimensional signal parameter space). Grizzlies, brown bears, black bears, sun bears, ...
3) Usually there's more than one signal region ($n_{on\_1=}$, ... n_{on_n}), each with its own sets of ε's and τ's. Direct sightings of bears, observation of disturbed garbage cans, eyewitness accounts, auditory-only incidents, etc.
4) The ε's are uncertain. Sometimes they are just ratios of Poisson distributed numbers, but often there are more sources of uncertainty than just that. Same with the τ's. How to convert grizzlies/day to an expected number of pictures of grizzlies/day?
5) Often we have two or more "off-signal" auxiliary experiments used to evaluate b, each with its τ. What to do when they disagree?

Banff Challenge 2 samples 1, 3, and 4 above. 2 isn't so important as long as we can understand how to deal with the 1-signal problem, although problems occur in high-dimensional models that are not present in 1D models.

A Comment on low s and low b

Bins with tiny s and tiny b can have large s/b   (Louis Lyons: large s/sqrt(b) is suspicious)

Naturally occurring in HEP and others seeking discovery:

1) Each beam crossing has very small s and b but has the same s/b as
     neighboring beam crossings. Can make a histogram of the search for new
      physics separately for each beam crossing. Same s and b predictions, just
      scaled down very small.

      Adding is the same as a more elaborate combination if the histograms were
       accumulated under identical conditions (all rates, shapes, and systematics are
       the same)

2) Surveillance video catching a bear – each frame has a small s, b, but still
      worthwhile to collect each frame (and analyze them separately)

Available Software, Tools, Documentation

CDF Statistics Committee

http://www-cdf.fnal.gov/physics/statistics/statistics_home.html
  Useful for documentation. Provides advice for common, thorny questions

BaBar Statistics Working Group

http://www.slac.stanford.edu/BFROOT/www/Statistics/

ROOSTATS

https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome
  A very complete toolset. I haven't used it (but I should have). It's in
  common use at the LHC

MCLIMIT

http://www-cdf.fnal.gov/~trj/mclimit/production/mclimit.html
  Used on CDF, some use on D0 and LHC. Limits, cross sections, p-values,
  both Frequentist and Bayesian tools

PHYSTAT.ORG

http://www.phystat.org
  Maintained by Jim Linnemann. We toolsmiths really
    should keep it up to date…

Available Software, Tools, Documentation

PDG Probability and Statistics Reviews (ed. Glen Cowan)

http://pdg.lbl.gov/2011/reviews/rpp2011-rev-probability.pdf
http://pdg.lbl.gov/2011/reviews/rpp2011-rev-statistics.pdf

If these links get out of date, just search pdg.lbl.gov for the mathematical reviews
Excellent brief reference, but maybe a little too brief to learn the material.

Good Reads:

Frederick James, "Statistical Methods in Experimental
 Physics", 2nd edition, World Scientific, 2006

Louis Lyons, "Statistics for Nuclear and Particle Physicists"
 Cambridge U. Press, 1989

Glen Cowan, "Statistical Data Analysis" Oxford Science Publishing, 1998

Roger Barlow, "Statistics, A guide to the Use of Statistical
Methods in the Physical Sciences", (Manchester Physics Series) 2008.

Bob Cousins, "Why Isn't Every Physicist a Bayesian"
Am. J. Phys **63**, 398 (1995).

Available Software, Tools, Documentation

A simple web-based limit calculator based on a one-dimensional event count

http://www-d0.fnal.gov/Run2Physics/limit_calc/limit_calc.html

Summary

Statistics, like physics, is a lot of fun!

It's central to our job as scientists, and about how human knowledge is obtained from observation.

Lots of ways to address the same problems.

Many questions do not have a single answer. Room for uncertainty.   Probability and uncertainty are different but related.

Think about how your final result will be extracted from the data before you design your experiment/analysis -- keep thinking about it as you improve and optimize it.