Statistical Methods for Experimental Particle Physics

Tom Junk

Fermilab

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

Day 4:

- Density Estimation
  - Binning
  - Smoothing
- Model Validation

Density Estimation

- Sometimes the result of an experiment is a distribution, and not a number or small set of measured parameters.

- Even for simpler hypothesis tests and measurements, predicted distributions need to be compared with observed data.

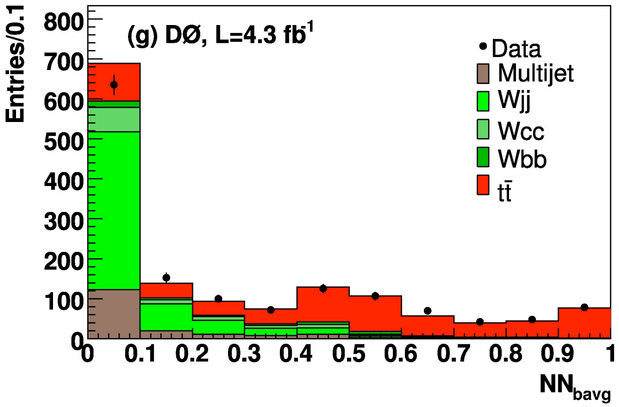
- We usually do not know *a priori* what the distribution is supposed to be, or even what the parameters are.

  - Underlying physics models may be "simple" – e.g. cosθ distribution of Z decay products at LEP: ~$(1+\cos^2\theta)$

  - Detector acceptance, trigger bias, analysis selection cuts sculpt simple distributions and make them complicated.

  - Some distributions we have even less *a priori* knowledge: MVA's for example. Or even just m_{jj} in W+jets events (thousands of diagrams in Madgraph).

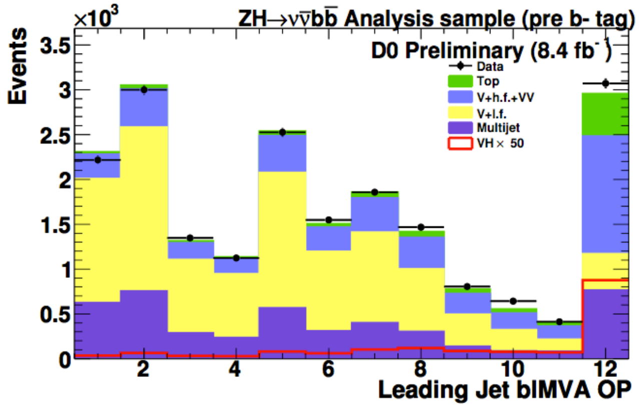
An Example Neural Network Output Distribution with an Odd Shape



D0 Collaboration, arXiv:1011.6549,
Submitted to Phys. Rev. D

Typical NN Software packages seek to rank outcomes in increasing s/b. NN output is usually very close to the s/b in the output bin.

If the selected data sample contains more than one category of events (even if they are not colored the same way in the stack), one can have bumps in the middle of the plot.



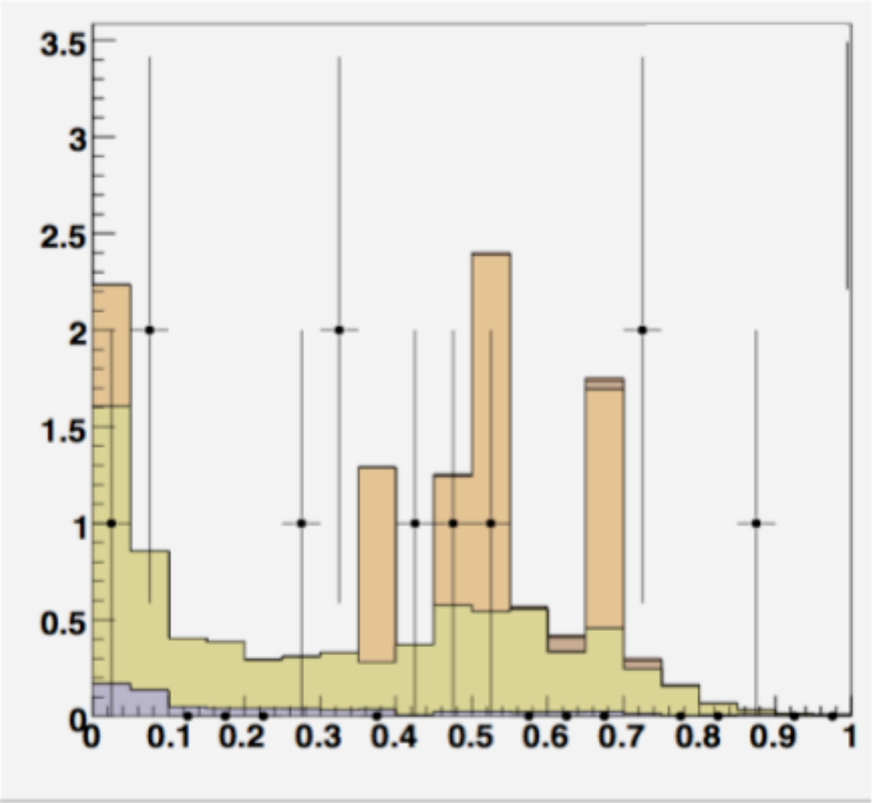
Usually these are investigated and explained a postiori. Usually it's okay – we care about modeling, but not about the distribution.

Many distributions (e.g., decision trees, binned likelihood functions) are not expected to have smooth distributions.

Normally we use Monte Carlo to predict the distributions of arbitrarily chosen reconstructed observables.

A Pitfall -- Not Enough MC (data) To Make Adequate Predictions

An Extreme Example (names removed)



Questions: What's the shape we are trying to estimate? What is the uncertainty on that shape?

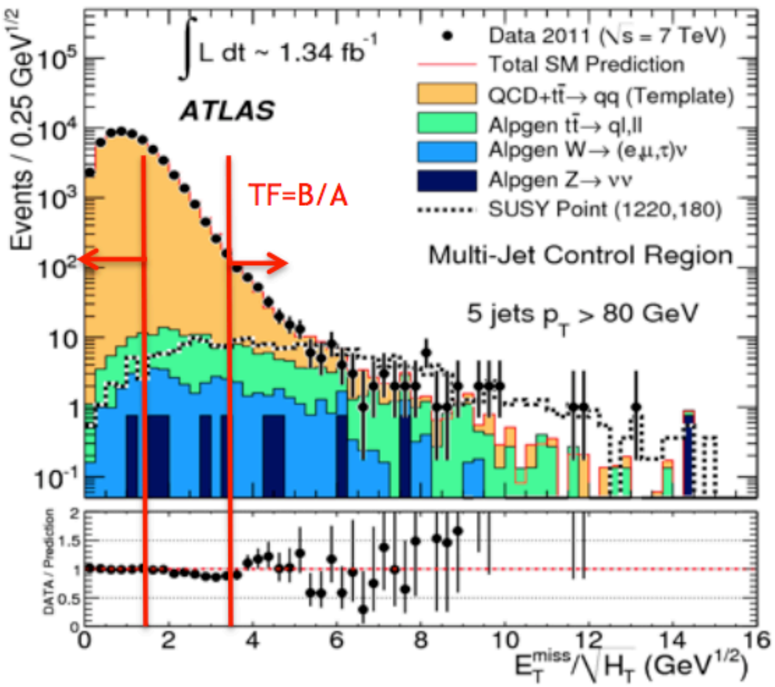
Cousins, Tucker and Linnemann tell us prior predictive p-values undercover with 0±0 events are predicted in a control sample.

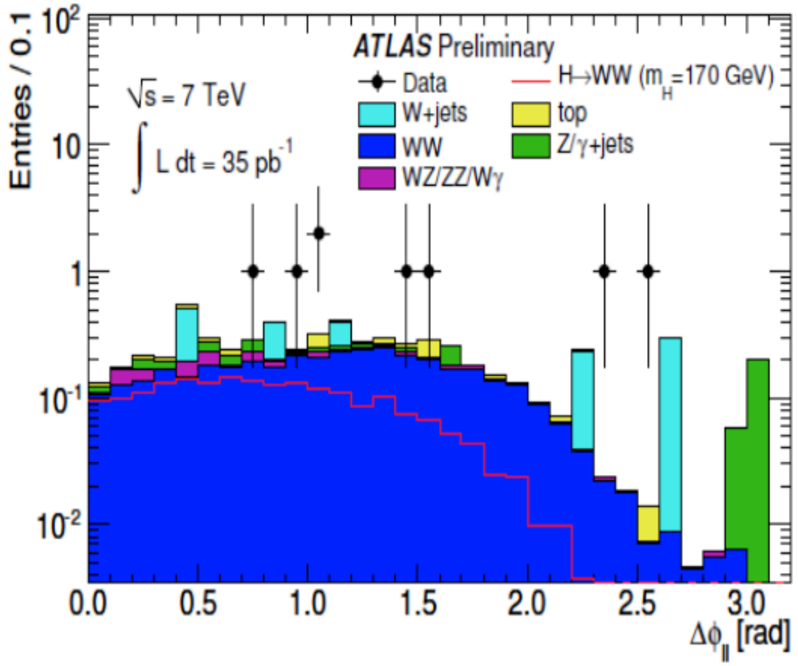
CTL Propose a flat prior in true rate, use joint LF in control and signal samples. Problem is, the mean expected event rate in the control sample is n_{obs}+1 in control sample. Fine binning → bias in background prediction.

Overcovers for discovery, undercovers for limits?

Some Very Early Plots from ATLAS

Suffer from limited sample sizes in control samples and Monte Carlo
Nearly all experiments are guilty of this, especially in the early days!





Data points’ error bars are not sqrt(n). What are they? I don’t know. How about the uncertainty on the prediction?

The left plot has adequate binning in the “uninteresting” region. Falls apart on the right-hand side, where the signal is expected.
Suggestions: More MC, Wider bins, transformation of the variable (e.g., take the logarithm).
Not sure what to do with the right-hand plot except get more modeling events.

An Extreme Example from Georgios Choudalakis

Ten MC events, used to estimate a background b, but with different weights.

τ_1=0.1
τ_2=0.2
τ_3=0.3
τ_4=0.4
τ_5=0.5
τ_6=0.6
τ_7=0.7
τ_8=0.8
τ_9=0.9
τ_{10}=1.0

The sum is 5.5 = b
But what to use for the prior on b?

Are there any possible (and possibly large) weights which are not represented here? Could we have gotten a MC event with weight=100?

Very little information about the distribution of the weights is present here.

Need acceptance as a function of weight.

General limit/discovery tools – do we need a histogram of weights for each bin of each signal and background contribution? What if this is insufficient anyway (as it is in this case).

Binned and Unbinned Analyses

- Binning events into histograms is necessarily a lossy procedure

- If we knew the distributions from which the events are drawn (for signal and background), we could construct likelihoods for the data sample without resort to binning. (Example Next page)

- Modeling issues: We have to make sure our parameterized shape is the right one or the uncertainty on it covers the right one at the stated C.L.

- Unfortunately there is no accepted unbinned goodness-of-fit test

A naive prescription: Let's compute L(data|prediction), and see where it falls on a distribution of possible outcomes –
compute the p-value for the likelihood.

Why this doesn't work: Suppose we expect a uniform distribution of events in some variable. Detector ϕ is a good variable. All outcomes have the same joint likelihood, even those for which all the data pile up at a specific value of phi. Chisquared catches this case much better.

Another example: Suppose we are measuring the lifetime of a particle, and we expect an exponential distribution of reconstructed times with no background contribution. The most likely

Kernel Estimation

Take the histogram, but replace "bin" function B with something else:

$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x - x_i; w),$$

where $k(x, w)$ is the "kernel function", normalized to unity:

$$\int_{-\infty}^{\infty} k(x; w)\, dx = 1.$$

Usually interested in kernels of the form

$$k(x - x_i; w) = \frac{1}{w} K\left(\frac{x - x_i}{w}\right),$$

indeed this may be used as the definition of "kernel". The kernel estimator for the PDF is then:

$$\widehat{p}(x) = \frac{1}{nw}\sum_{i=1}^{n} K\left(\frac{x - x_i}{w}\right),$$

The role of parameter w as a smoothing parameter is clearer.

The Problem With Smoothing (II)

For example, suppose we have a kernel estimator:

$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x - x_i; w),$$

Its expectation is:

$$\begin{aligned} E[\widehat{p}(x)] &= \frac{1}{n}\sum_{i=1}^{n}\int k(x - x_i; w)p(x_i)dx_i \\ &= \int k(x-y)p(y)dy. \end{aligned}$$

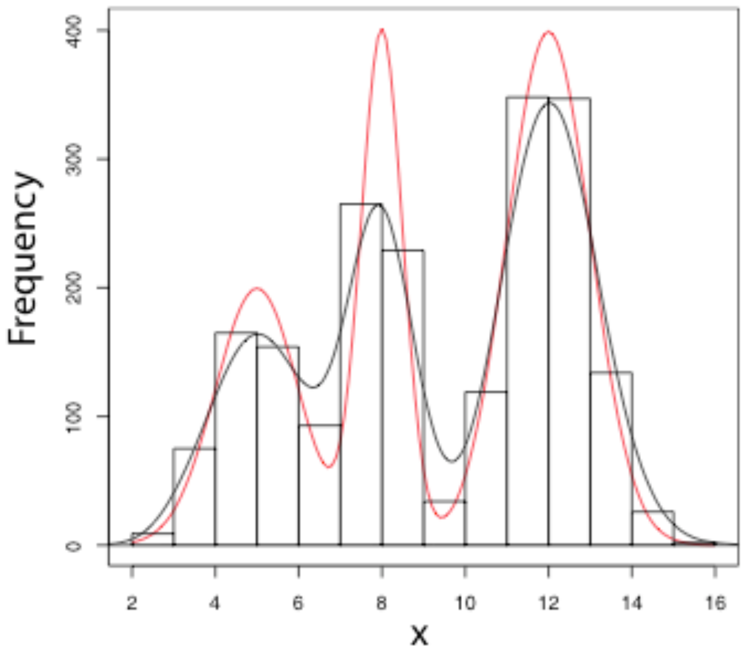
Unless $k(x-y) = \delta(x-y)$, $\widehat{p}(x)$ will be biased for some $p(x)$.

But $\delta(x-y)$ has infinite variance.

The Problem with Smoothing (III)

So the nice properties we strive for in parameter estimation (and sometimes achieve) are beyond reach.

Intuition: smoothing lowers peaks and fills in valleys.



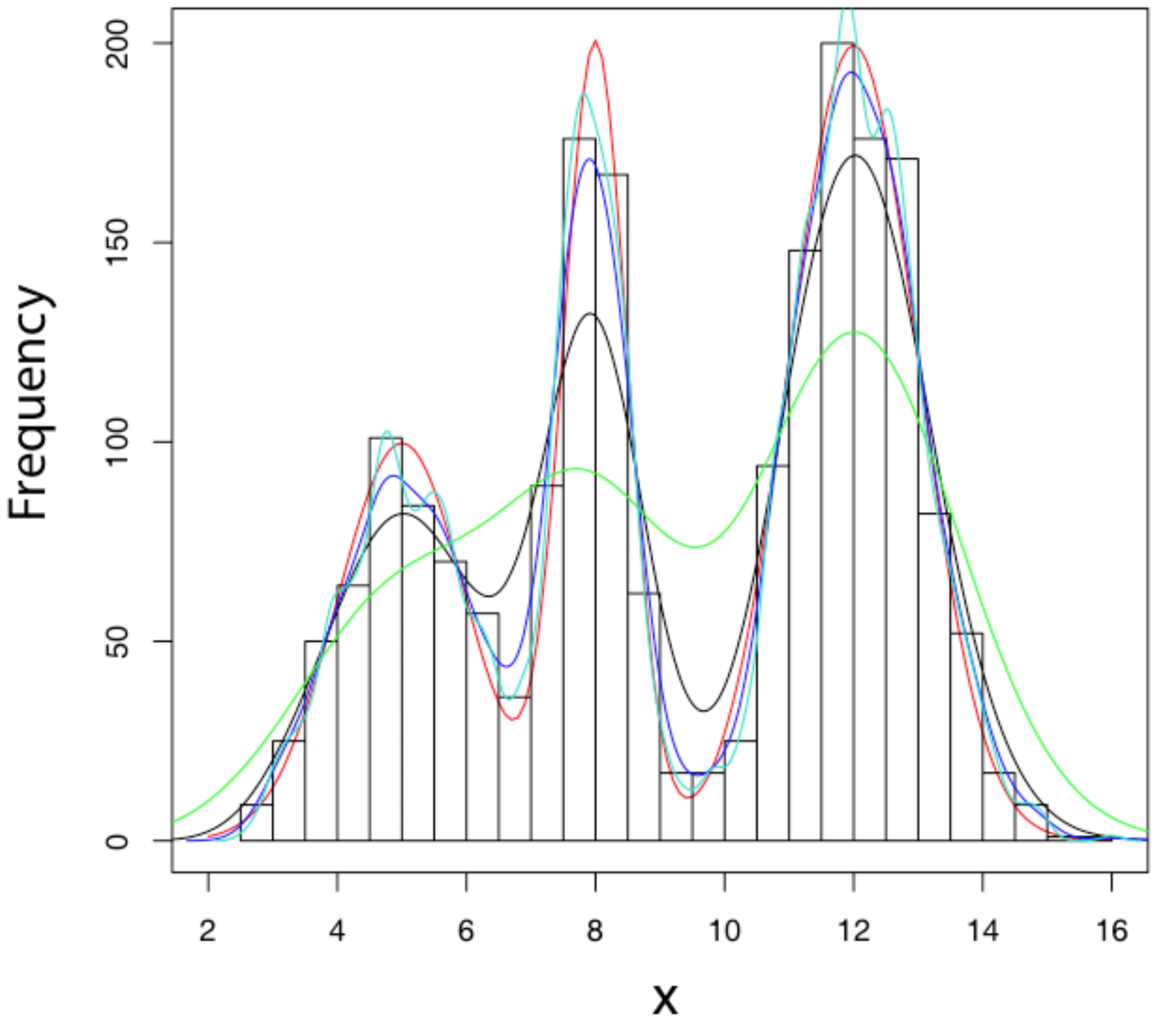
Red curve: PDF

Histogram: Sampling from PDF

Black curve: Gaussian kernel estimator for PDF

Dependence on Smoothing Parameter

Plot showing effect of choice of smoothing parameter”:



Red: Sampling PDF
Black: Default smoothing (w)
Blue: w/2 smoothing
Turquoise: w/4 smoothing
Green: 2w smoothing

Frank Porter, SLUO
lectures on statistics, 2006

Optimizing Histogram Binning

Two competing effects:

1) Separation of events into classes with different s/b improves the sensitivity of a search or a measurement. Adding events in categories with low s/b to events in categories with higher s/b dilutes information and reduces sensitivity.

→ Pushes towards more bins

2) Insufficient Monte Carlo can cause some bins to be empty, or nearly so. This only has to be true for one high-weight contribution.

Need reliable predictions of signals and backgrounds in each bin

→ Pushes towards fewer bins

Note: It doesn't matter that there are bins with zero data events – there's always a Poisson probability for observing zero.

The problem is inadequate prediction. Zero background expectation and nonzero signal expectation is a discovery!

Overbinning = Overlearning

A Common pitfall – Choosing selection criteria after seeing the data.
“Drawing small boxes around individual data events”

The same thing can happen with Monte Carlo Predictions –

Limiting case – each event in signal and background MC gets its own bin.
→Fake Perfect separation of signal and background!.

Statistical tools shouldn’t give a different answer if bins are shuffled/sorted.

Try sorting by s/b. And collect bins with similar s/b together. Can get arbitrarily good performance from an analysis just by overbinning it.

Note: Empty data bins are okay – just empty prediction is a problem. It is our
job however to properly assign s/b to data events that we did get (and all possible ones).

Model Validation

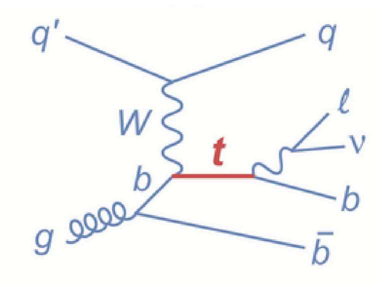
- Not normally a statistics issue, but something HEP experimentalists spend most of their time worrying about.

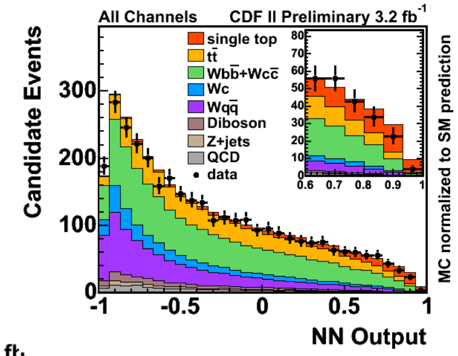
- Systematic Uncertainties on predictions are usually constrained by data predictions.

- Often discrepancies between data and prediction are the basis for estimating systematic uncertainty

Checking Input Distributions to an MVA

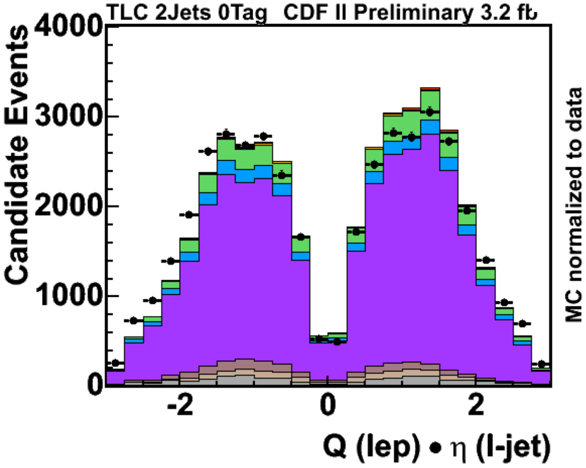
- Relax selection requirements – show modeling in an inclusive sample (example – no b-tag required for the check, but require it in the signal sample)
- Check the distributions in sidebands (require zero b-tags)
- Check the distribution in the signal sample for all selected events
- Check the distribution after a high-score cut on the MVA

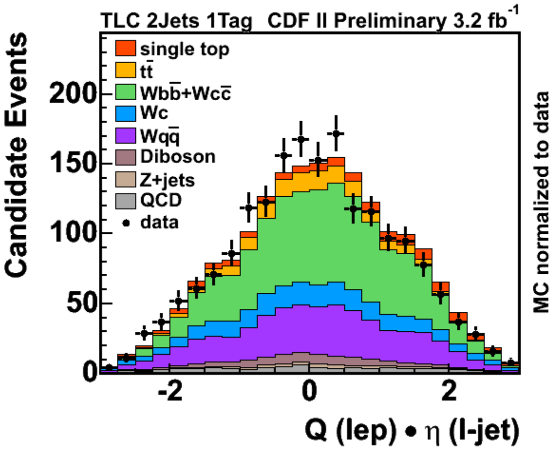


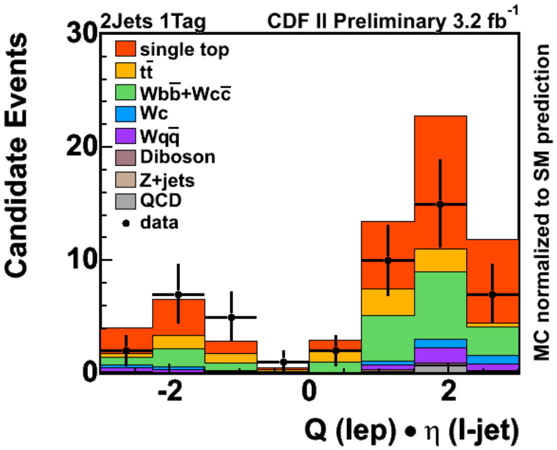


Example: $Q_{\text{lepton}}*\eta_{\text{untagged jet}}$ in CDF's single top analysis. Good separation power for t-channel signal.

Phys.Rev.D82:112005 (2010)





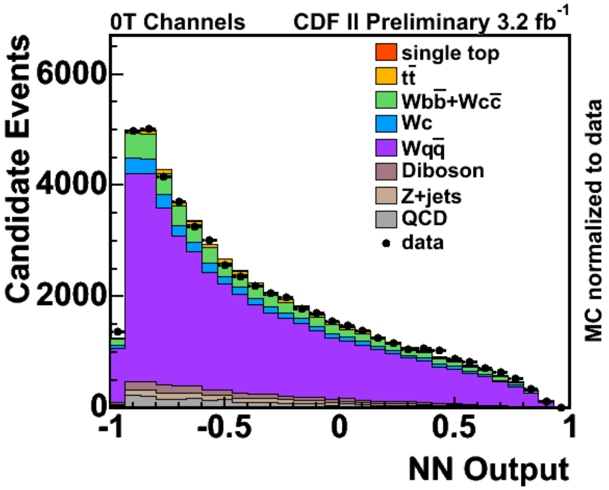


highest |η| jet as a well-chosen proxy

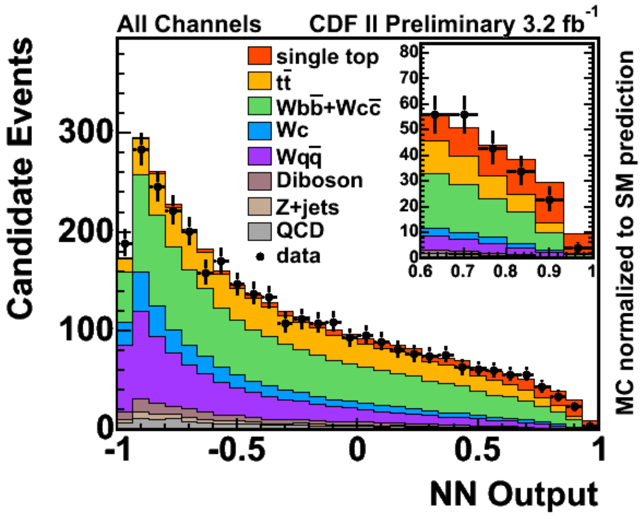
Checking MVA Output Distributions

- Calculate the same MVA function for events in sideband (control) regions
- For variables that are not defined outside of the signal regions, put in proxies. (sometimes just a zero for the input variable works well if the quantity really isn't defined at all – pick a typical value, not one way off on the edge of its distribution)
- Be sure to use the same MVA function as for analyzing the signal data.

Example: CDF NN single-top
NN validated using events with
zero b-tag



signal region

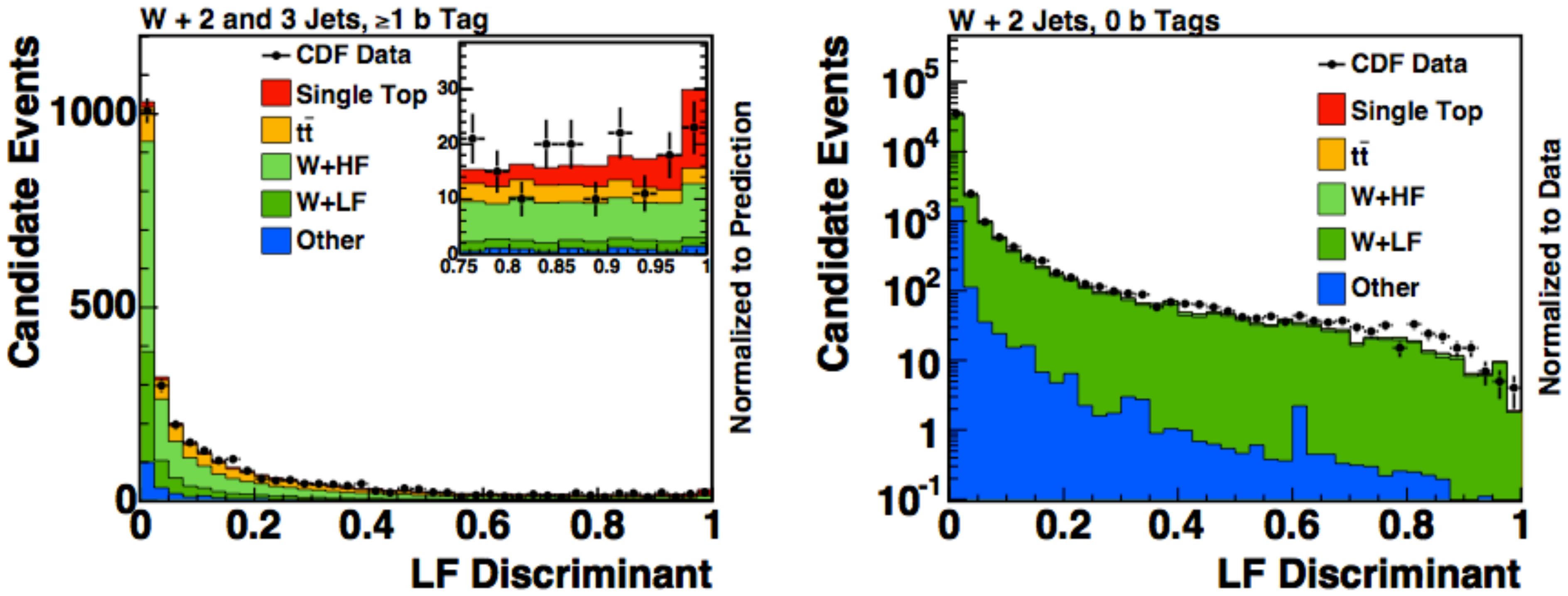


Phys.Rev.D82:112005 (2010)

A Comparison in a Control Sample that is Less than Perfect

CDF’s single top Likelihood Function discriminant checked in untagged events

(a) Phys.Rev.D82:112005 (2010) (b)

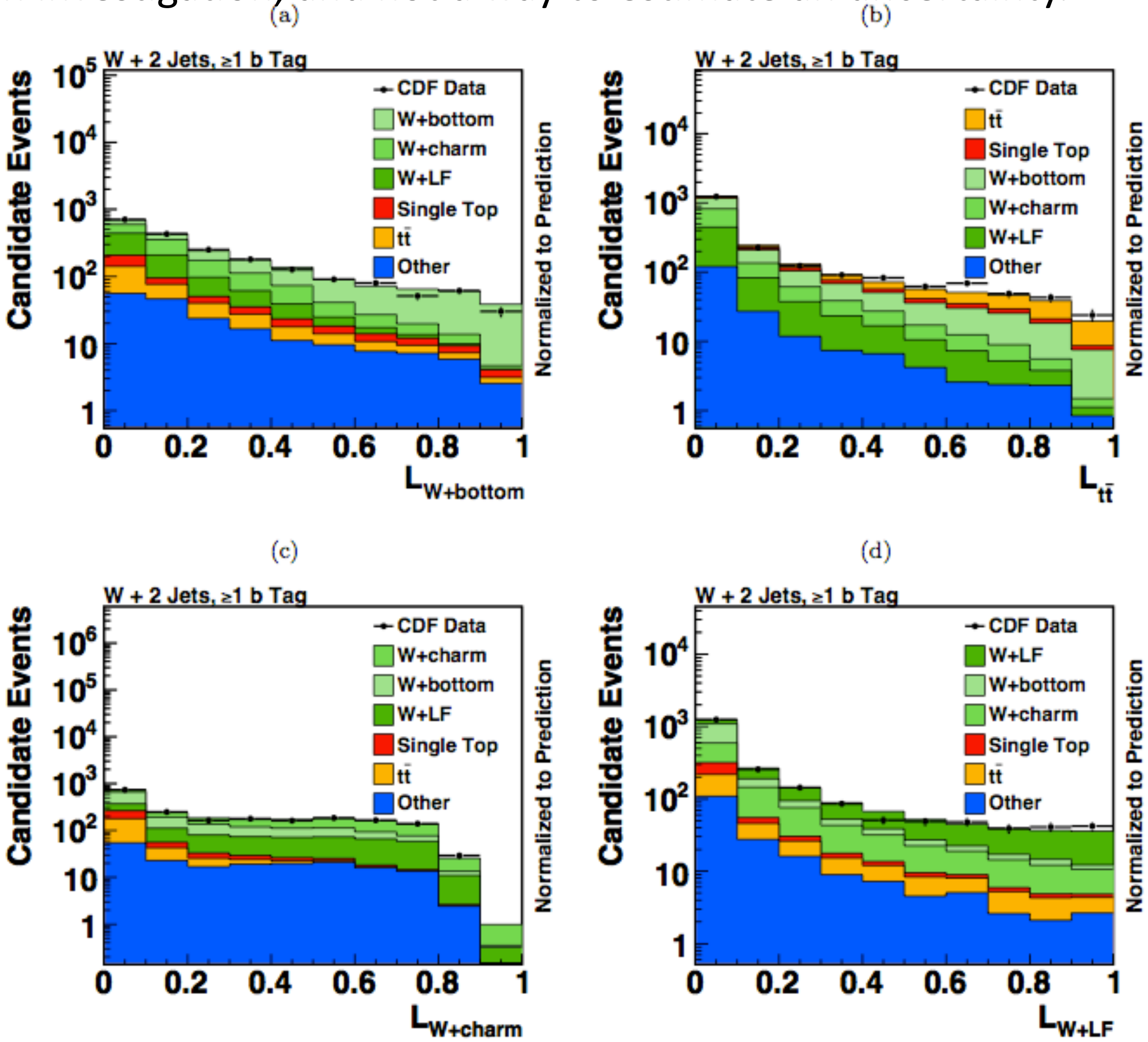


Strategy: Assess a shape systematic covering the difference between data and MC – extrapolate the uncertainty from the control sample to the signal sample.

If the comparison is okay within statistical precision, do not asses an additional uncertainty (even/especially if the precision is weak). Barlow, hep-ex/0207026 (2002).

Another Validation Possibility – Train Discriminants to Separate Each Background

Same input variables as signal LF. LF has the property that the sum of these plus the signal LF is 1.0 for each event. Gives confidence. If the check fails, it's a starting point for an investigation, and not a way to estimate an uncertainty.



Phys.Rev.D82:112005 (2010)

Model Validation with MVA's

- Even though input distributions can look well modeled, the MVA output could still be mismodeled.

  Possible cause – correlations between one or more variables could be mismodeled

- Checks in subsets of events can also be incomplete.

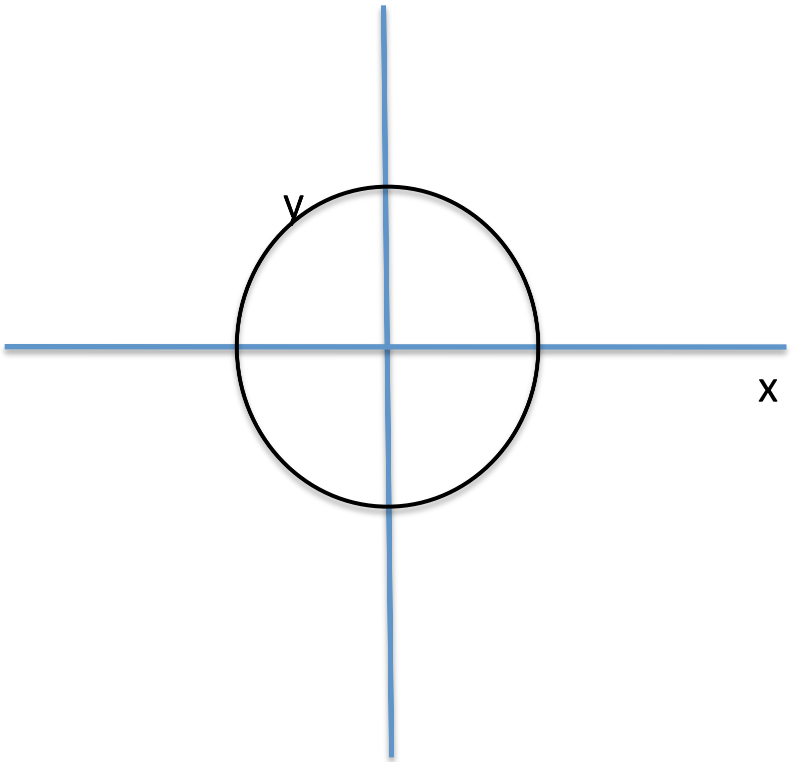
  A sum of distributions whose shapes are well reproduced by the theory can still be mismodeled if the relative normalizations of the components is mismodeled.

- Can check the correlations between variables pairwise between data and prediction
- Difficult to do if some of the prediction is a one-dimensional extrapolation from control regions (e.g., ABCD methods).

- My favorite: Check the MVA output distribution in bins of the input variables! We care more about the MVA output modeling than the input variable modeling anyway.
- Make sure to use the same normalization scheme as for the entire distribution – do not rescale to each bin's contents.

Ideally, we'd try to find a control sample depleted in signal that has exactly the same kind of background as the signal region (usually this is unavailable).

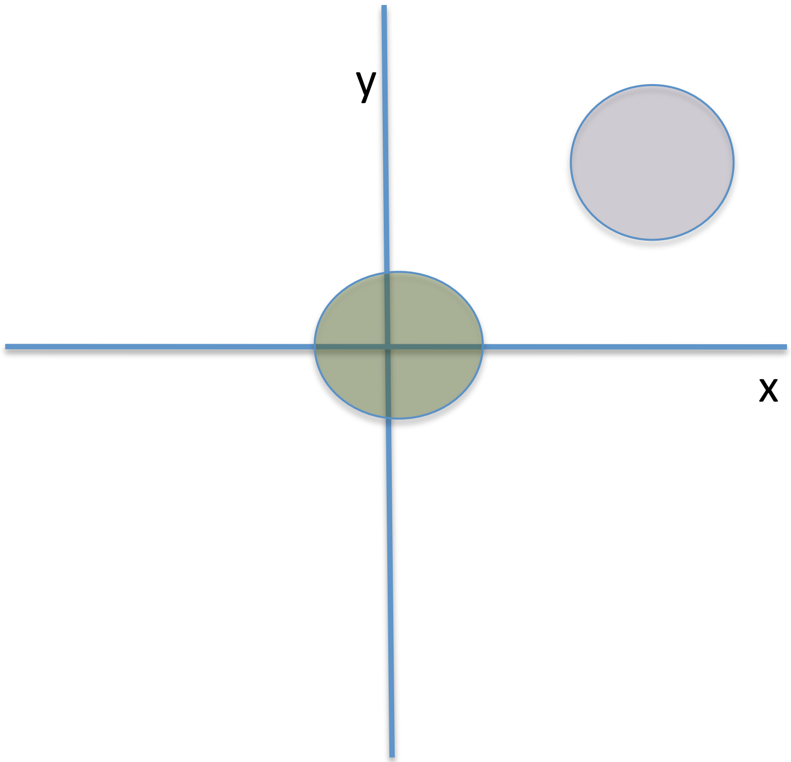
A Sample with Zero Covariance is Not Necessarily Uncorrelated



Example – perimeter of a circle. Knowledge of x provides knowledge of y up to a 2-fold ambiguity. But the covariance of the sample vanishes!

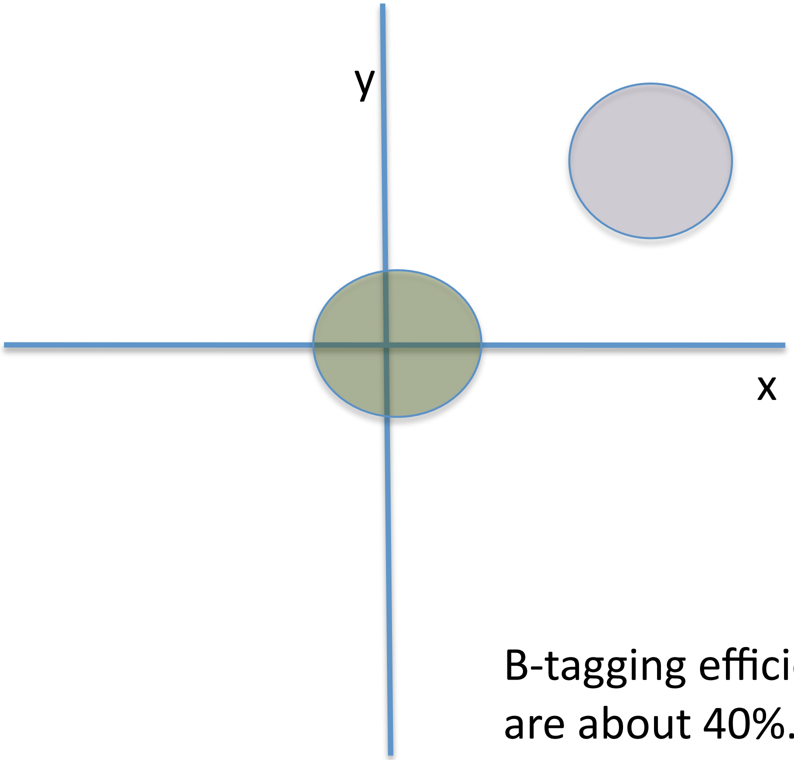
Something to watch out for with Principal Components Analysis – does not remove correlation, only covariance.

The Sum of Uncorrelated 2D Distributions may be Correlated



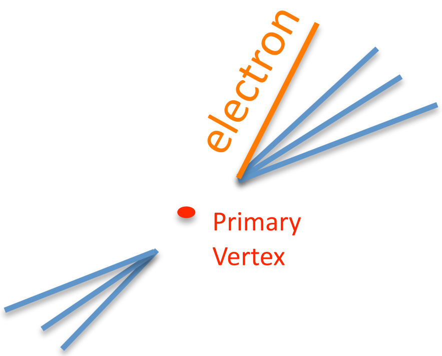
Knowledge of one variable helps identify which sample the event came from even if the individual samples have no covariance.

An Example: Double-Tag Methods



Dijet events at LEP1/SLD

Z→u,ubar
d,dbar
s,sbar
b,bbar
leptons
neutrinos

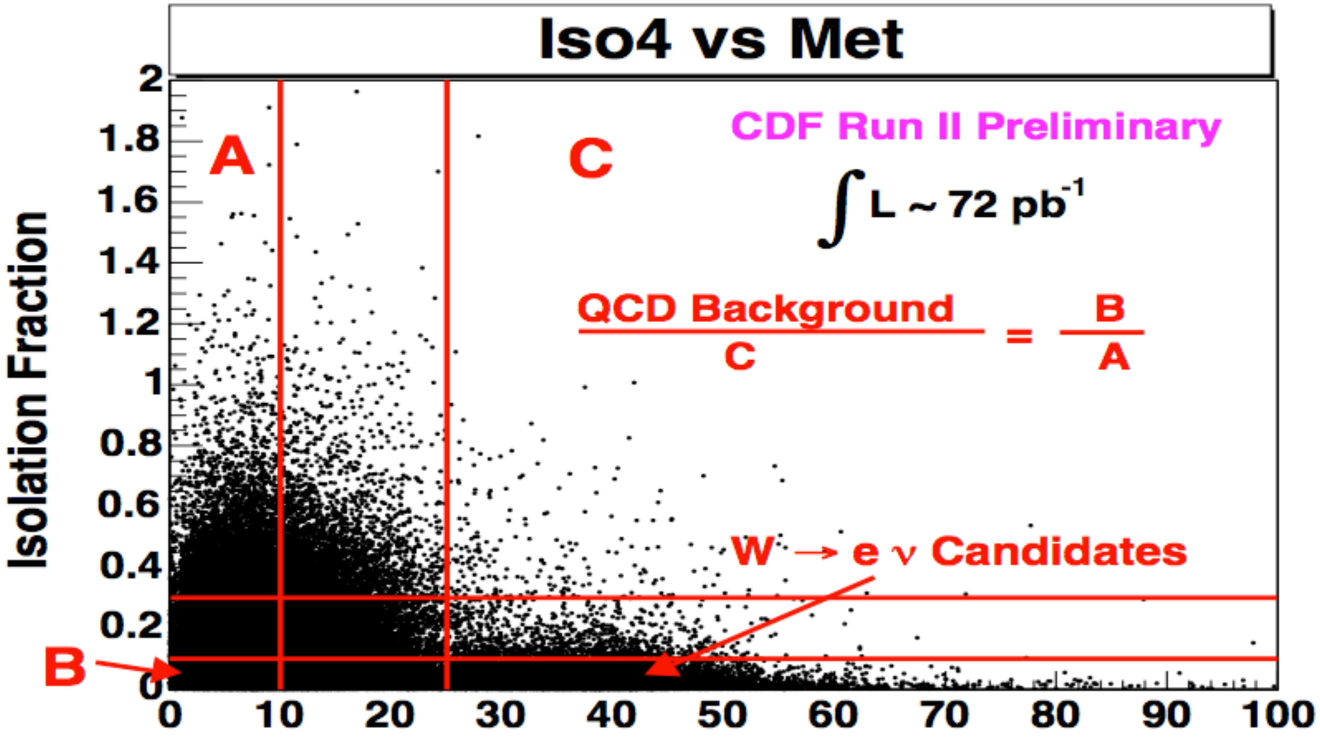


A double-vertex-B-tagged event with a semileptonic decay

B-tagging efficiencies (efficiency of finding the displaced vertex) are about 40%. We do not trust MC modeling of the b-tag efficiency. Would like to measure the B-tag efficiency and the Br(Z→b,bbar) branching fraction together in the same data. Count events with 0, 1, and 2 vertex tags. Enough information to solve for the Br and the efficiency.

x=b-tag of jet 1, y=b-tag of jet 2. Assume uncorrelated probabilities for tagging the jets. But the flavor of the jets is correlated! It is this flavor correlation that allows us to extract Br and Tag eff.

“ABCD” Methods



Missing Transverse Energy (MET)

Isolation fraction=

Energy in a cone of radius 0.4 around lepton candidate not including the lepton candidate / Energy of lepton candidate

Want QCD contribution to the “D” region where signal is selected.

CDF’s W Cross Section Measurement

Assumes: MET and ISO are uncorrelated sample by sample
Signal contribution to A,B, and C are small and subtractable

"ABCD" Methods

Advantages

- Purely data based, good if you don't trust the simulation
- Model assumptions are injected by hand and not in a complicated Monte Carlo program (mostly)
- Model assumptions are intuitive

Disadvantages

- The lack of correlation between MET and ISO assumption may be false. e.g., semileptonic B decays produce unisolated leptons and MET from the neutrinos.
- Even a two-component background can be correlated when the contributions aren't by themselves.
- Another way of saying that extrapolations are to be checked/assigned sufficient uncertainty
- Works best when there are many events in regions A,B, and C. Otherwise all the problems of low stats in the "Off" sample in the On/Off problem reappear here. Large numbers of events → Gaussian approximation to uncertainty in background in D
- Requires subtraction of signal from data in regions A, B, and C → introduces model dependence
- Worse, the signal subtraction from the sidebands depends on the signal rate being measured/tested.

→ A small effect if s/b in the sidebands is small

→ You can iterate the measurement and it will converge quickly

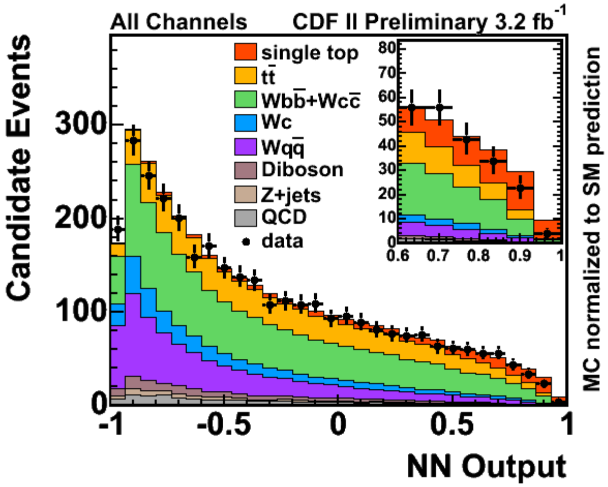
Examples of ABCD Methods

- MET vs. ISO
- Sideband calibration of background under a peak. (“what if the background peaks also where the signal peaks?)
- Upsilon polarization measurement from CDF
- The on-off problem with T=A/C. Very frequently samples A and C are in MC simulations, where we can be sure not to contaminate the background estimations wtih signal

Uncorrelated variable assumption == assumption that T is the same in the data and the MC. (check modeling of shape of distribution in the MC)

Equivalent of previous problem: Even if the background shapes are well modeled by the MC, if there are multiple background processes which contribute, they can have different fractional contributions, distorting the total shapes.

- Fitting an MVA shape to the data. Low-score MC = A, High-Score MC = C
  Low-score data = B, High-score Data=D.


All Channels
CDF II Preliminary 3.2 fb^{-1}
Candidate Events
MC normalized to SM prediction
NN Output
🟥 single top
🟧 $t\bar{t}$
🟩 $Wb\bar{b}+Wc\bar{c}$
🟦 Wc
🟪 $Wq\bar{q}$
▨ Diboson
▨ Z+jets
▨ QCD
● data