Statistical Methods in Particle Physics
Lecture 1: Bayesian methods

SUSSP65

St Andrews

16–29 August 2009

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Lecture #1: An introduction to Bayesian statistical methods

Role of probability in data analysis (Frequentist, Bayesian)

A simple fitting problem : Frequentist vs. Bayesian solution

Bayesian computation, Markov Chain Monte Carlo

Lecture #2: Setting limits, making a discovery

Frequentist vs Bayesian approach,

treatment of systematic uncertainties

Lecture #3: Multivariate methods for HEP

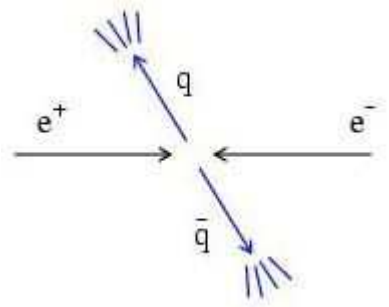
Event selection as a statistical test

Neyman-Pearson lemma and likelihood ratio test

Some multivariate classifiers:

NN, BDT, SVM, ...

Data analysis in particle physics



Observe events of a certain type

Measure characteristics of each event (particle momenta, number of muons, energy of jets,...)

Theories (e.g. SM) predict distributions of these properties up to free parameters, e.g., α, G_F, M_Z, α_s, m_H, ...

Some tasks of data analysis:

Estimate (measure) the parameters;

Quantify the uncertainty of the parameter estimates;

Test the extent to which the predictions of a theory are in agreement with the data ($\rightarrow$ presence of New Physics?)

Dealing with uncertainty

In particle physics there are various elements of uncertainty:

theory is not deterministic

quantum mechanics

random measurement errors

present even without quantum effects

things we could know in principle but don't

e.g. from limitations of cost, time, ...

We can quantify the uncertainty using PROBABILITY

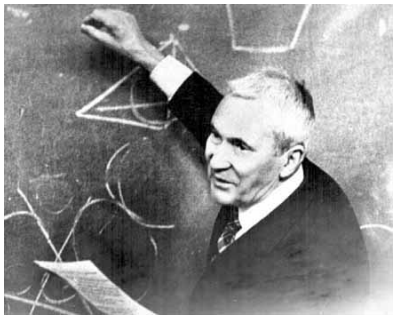
A definition of probability

Consider a set S with subsets A, B, ...

For all $A \subset S, P(A) \geq 0$

$$P(S) = 1$$

If $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$



Kolmogorov axioms (1933)

Also define conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Interpretation of probability

I. Relative frequency

A, B, ... are outcomes of a repeatable experiment

$$P(A) = \lim_{n\to\infty} \frac{\text{times outcome is } A}{n}$$

cf. quantum mechanics, particle scattering, radioactive decay...

II. Subjective probability

A, B, ... are hypotheses (statements that are true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

- Both interpretations consistent with Kolmogorov axioms.
- In particle physics frequency interpretation often most useful, but subjective probability can provide more natural treatment of non-repeatable phenomena:

  systematic uncertainties, probability that Higgs boson exists,...

Bayes' theorem

From the definition of conditional probability we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

but $P(A \cap B) = P(B \cap A)$, so

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem

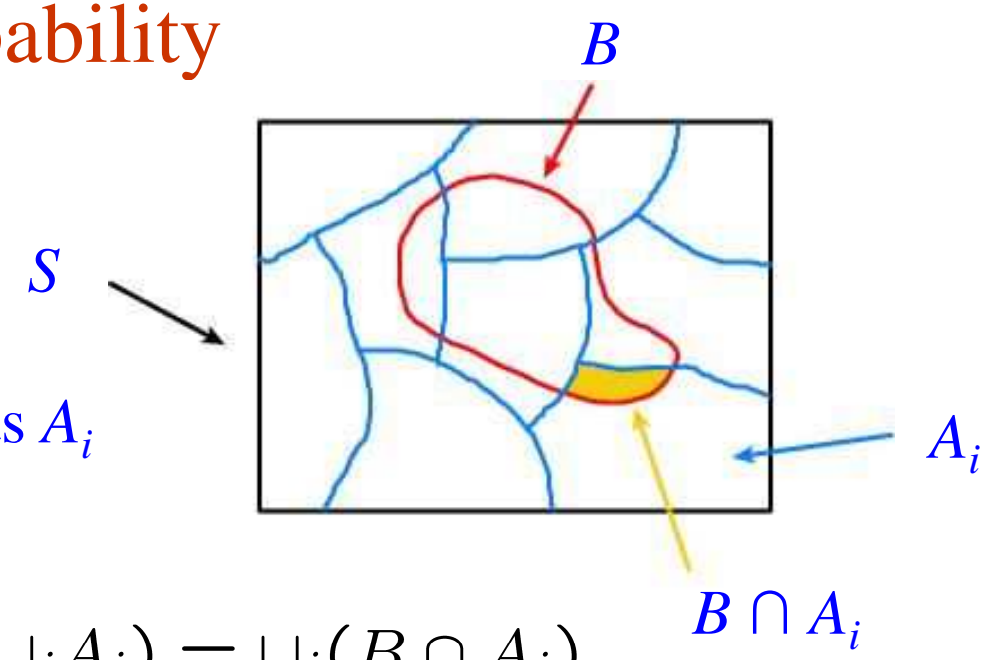
First published (posthumously) by the Reverend Thomas Bayes (1702–1761)

An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc. **53** (1763) 370; reprinted in Biometrika, **45** (1958) 293.

The law of total probability

Consider a subset B of the sample space S,

divided into disjoint subsets A_i such that $\cup_i A_i = S$,



$$\rightarrow \quad B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i),$$

$$\rightarrow \quad P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

$$\rightarrow \quad P(B) = \sum_i P(B|A_i) P(A_i)$$ law of total probability

Bayes' theorem becomes

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
P ($0.117 < \alpha_s < 0.121$),

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian Statistics – general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:

probability of the data assuming hypothesis H (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors ("if-then" character of Bayes' thm.)

Statistical vs. systematic errors

Statistical errors:

How much would the result fluctuate upon repetition of the measurement?

Implies some set of assumptions to define probability of outcome of the measurement.

Systematic errors:

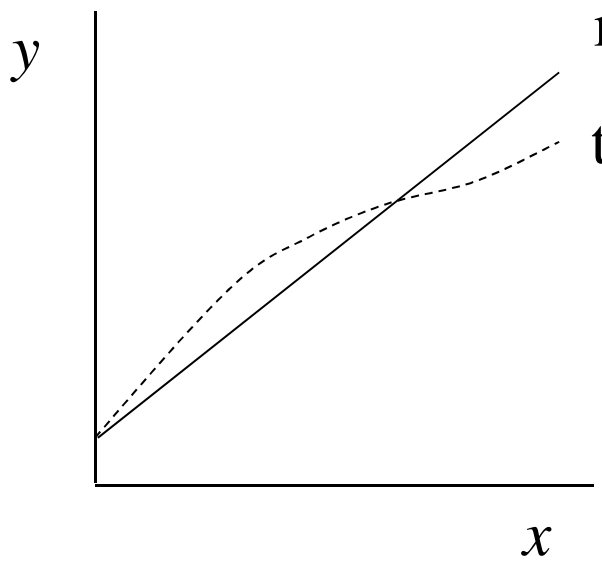
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modeling of measurement apparatus.

Usually taken to mean the sources of error do not vary upon repetition of the measurement. Often result from uncertain value of calibration constants, efficiencies, etc.

Systematic errors and nuisance parameters

Model prediction (including e.g. detector effects)
never same as "true prediction" of the theory:



model: $y = \alpha + \beta x$

truth: $y = \alpha + \beta x + \gamma x^2 + \varepsilon x^3 + \ldots$

Model can be made to approximate better the truth by including more free parameters.

systematic uncertainty ↔ nuisance parameters

Example: fitting a straight line

Data: (x_i, y_i, σ_i) , $i = 1, \ldots, n$.

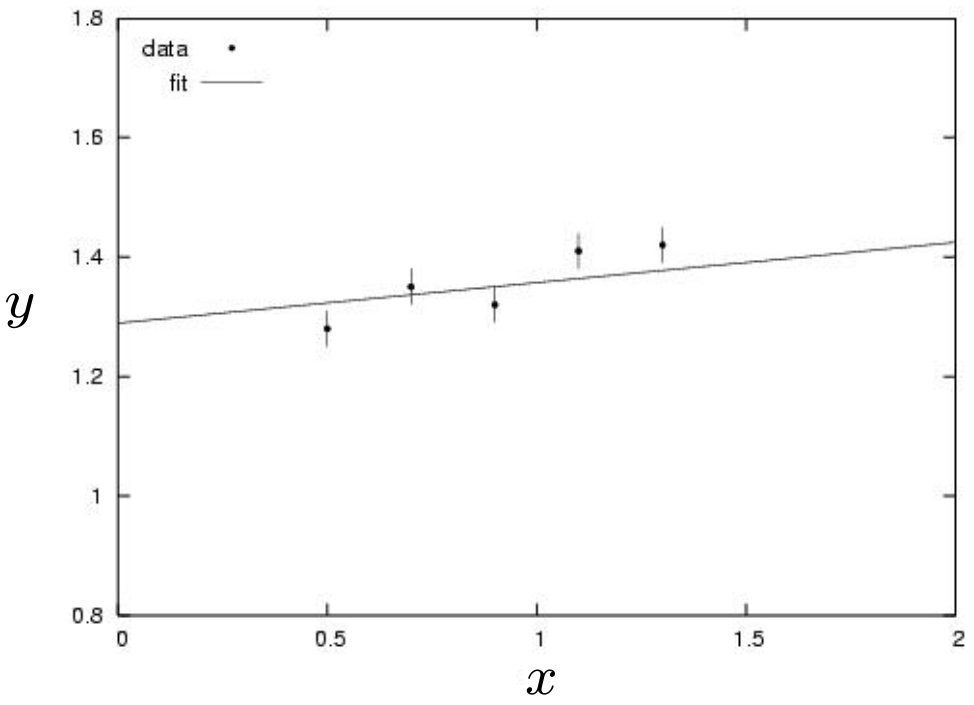
Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x$,

assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).



Frequentist approach

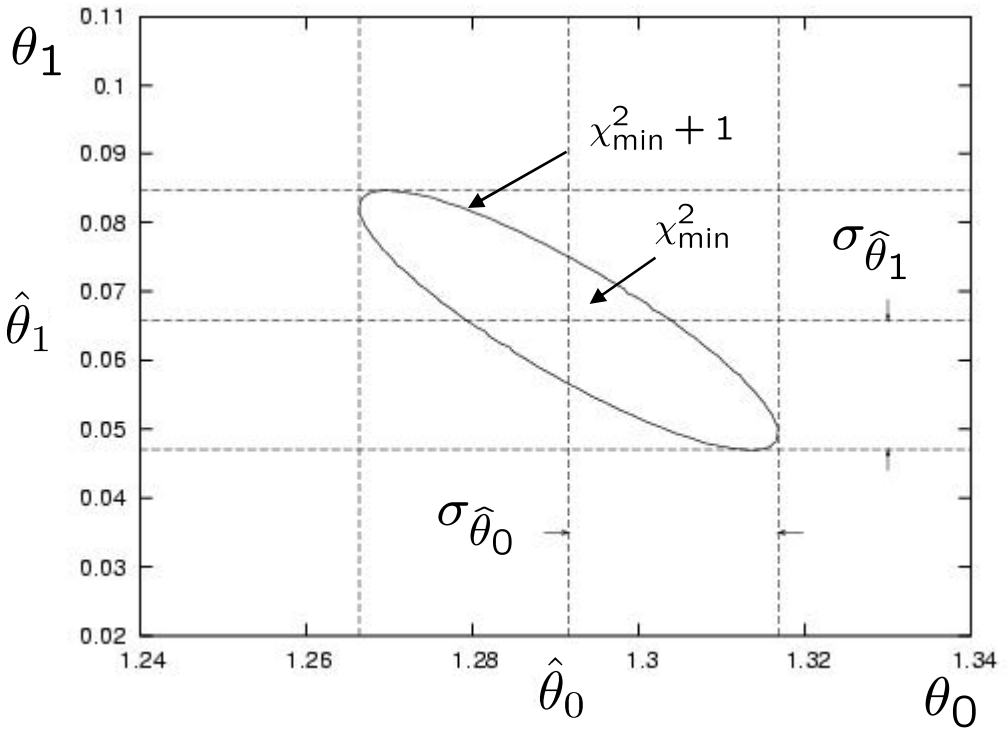
$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] ,$$

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from tangent lines to contour

$$\chi^2 = \chi^2_{\text{min}} + 1 .$$

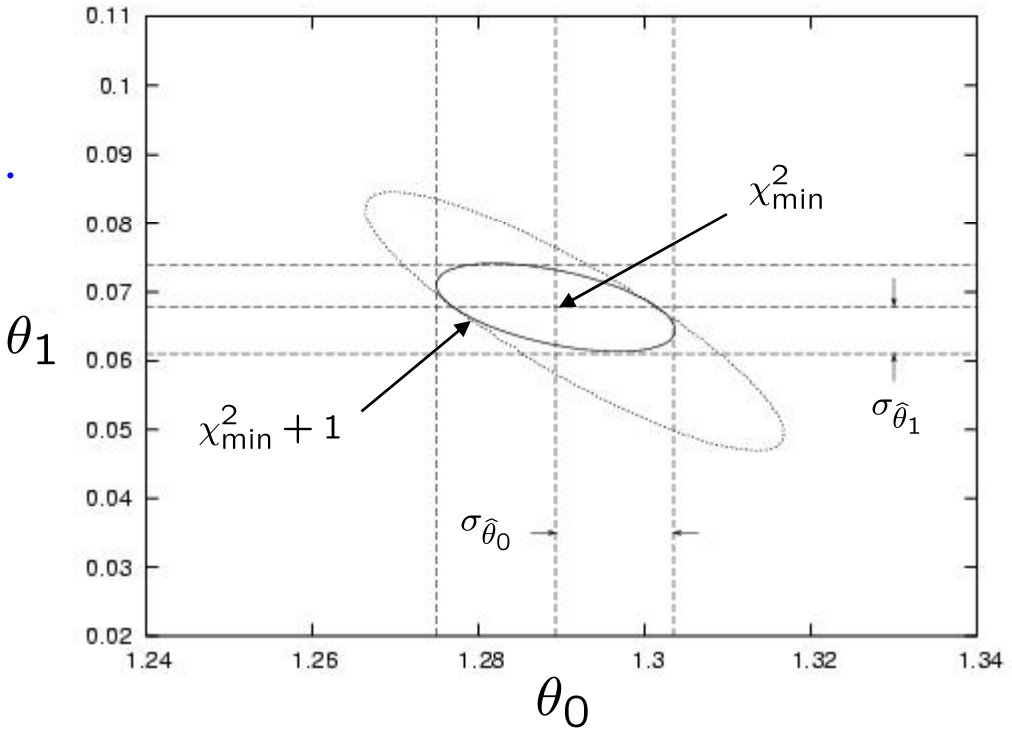
Correlation between $\hat{\theta}_0$, $\hat{\theta}_1$ causes errors to increase.



Frequentist case with a measurement t_1 of θ_1

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2} \, .$$

The information on θ_1 improves accuracy of $\hat{\theta}_0$.



Bayesian method

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\,\pi_1(\theta_1)$$

$$\pi_0(\theta_0) = \text{const.}$$

reflects 'prior ignorance', in any case much broader than $L(\theta_0)$

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

← based on previous measurement

Putting this into Bayes' theorem gives:

$$p(\theta_0, \theta_1|\vec{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2} \; \pi_0 \; \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

posterior $\propto$ likelihood $\times$ prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 \,|\, x)$ to find $p(\theta_0 \,|\, x)$:

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

In this example we can do the integral (rare). We find

$$\begin{aligned} p(\theta_0|x) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0-\hat{\theta}_0)^2/2\sigma^2_{\theta_0}} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \text{ (same as before)} \end{aligned}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form,
also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

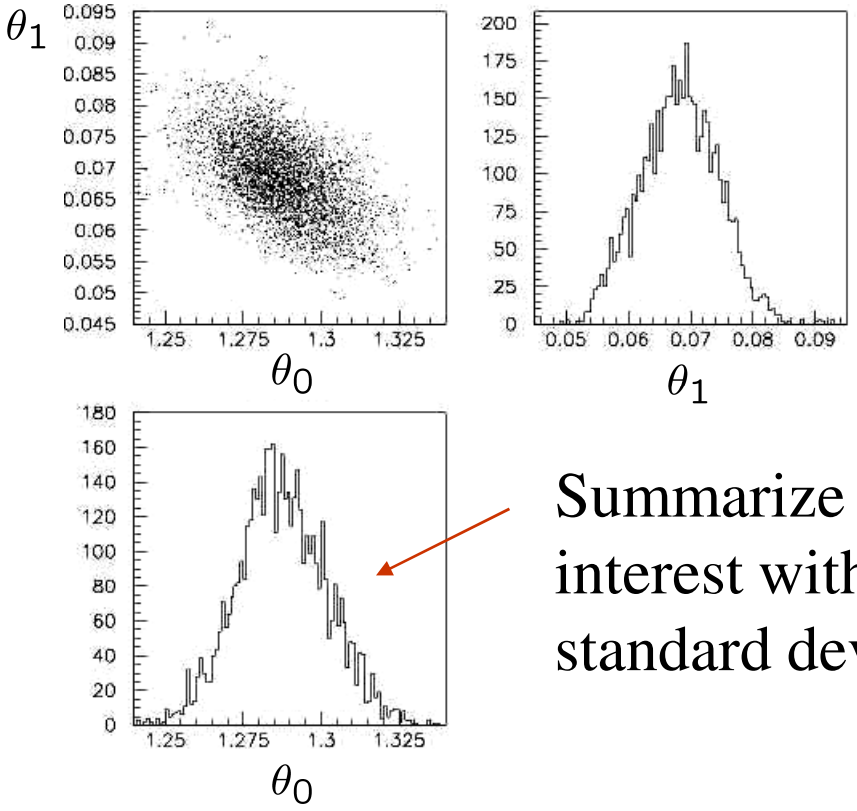
MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than if uncorrelated .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred
about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathsf{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than it would be with uncorrelated points.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it;
if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.
If proposed step rejected, hop in place.

Metropolis-Hastings caveats

Actually one can only prove that the sequence of points follows the desired pdf in the limit where it runs forever.

There may be a "burn-in" period where the sequence does not initially follow $p(\vec{\theta})$.

Unfortunately there are few useful theorems to tell us when the sequence has converged.

Look at trace plots, autocorrelation.

Check result with different proposal density.

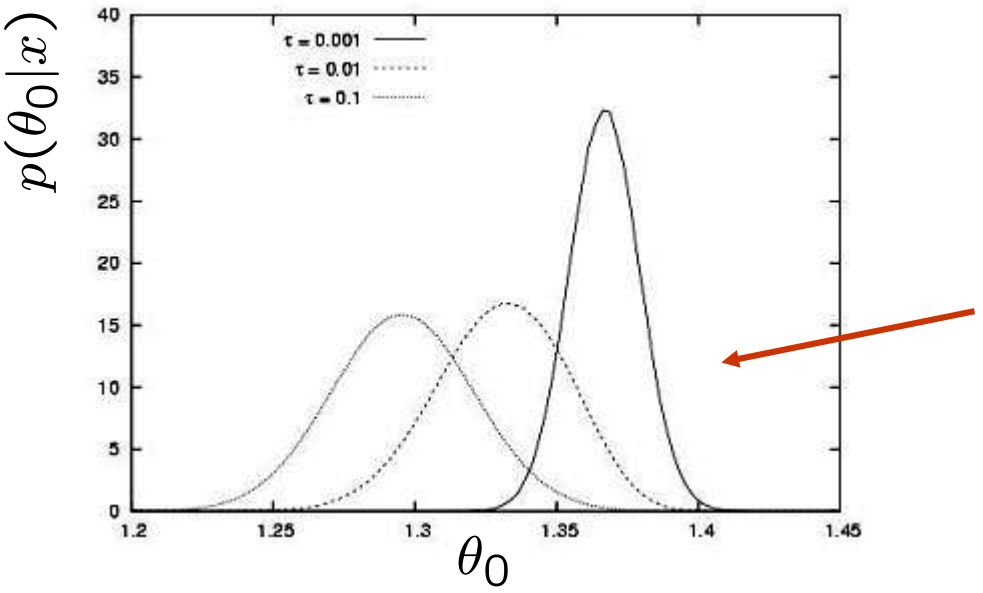
If you think it's converged, try starting from a different point and see if the result is similar.

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau} , \quad \theta_1 \geq 0 , \quad \tau = 0.1 .$$

From this we obtain (numerically) the posterior pdf for θ_0:



This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

A more general fit (symbolic)

Given measurements: $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$,

and (usually) covariances: V_{ij}^{stat}, V_{ij}^{sys} .

Predicted value: $\mu(x_i; \theta)$, expectation value $E[y_i] = \mu(x_i; \theta) + b_i$

control variable — parameters — bias

Often take: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

Minimize $\chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing $L(\theta) \sim e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Its Bayesian equivalent

Take $L(\vec{y}|\vec{\theta},\vec{b}) \sim \exp\left[-\frac{1}{2}(\vec{y}-\vec{\mu}(\theta)-\vec{b})^T\, V_{\text{stat}}^{-1}\,(\vec{y}-\vec{\mu}(\theta)-\vec{b})\right]$

$\pi_b(\vec{b}) \sim \exp\left[-\frac{1}{2}\,\vec{b}^T\, V_{\text{sys}}^{-1}\,\vec{b}\right]$

$\pi_\theta(\theta) \sim \text{const.}$

Joint probability for all parameters

and use Bayes' theorem: $p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$

To get desired probability for θ, integrate (marginalize) over $\boldsymbol{b}$:

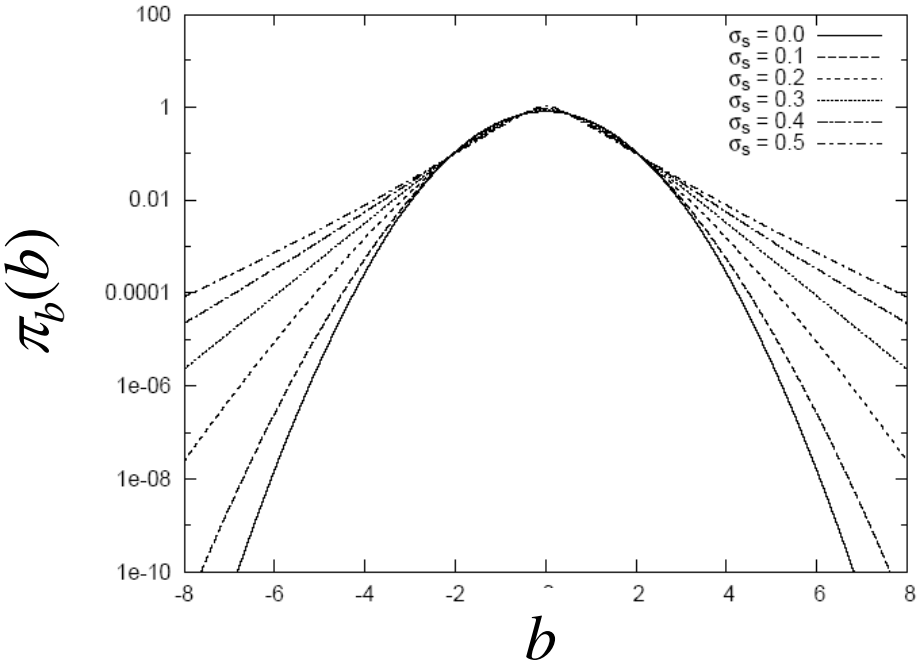
$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\, d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\text{min}} + 1$. (Back where we started!)

Alternative priors for systematic errors

Gaussian prior for the bias b often not realistic, especially if one considers the "error on the error". Incorporating this can give a prior with longer tails:

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i\sigma_i^{\text{sys}})^2}\right] \pi_s(s_i)\, ds_i$$



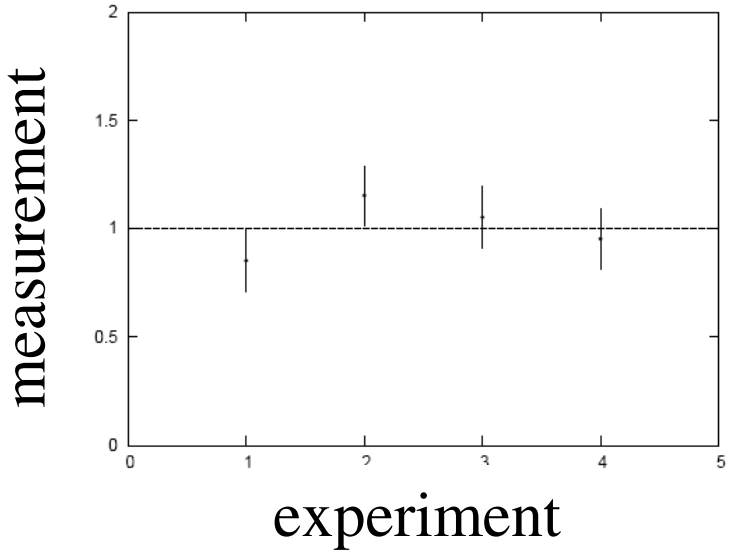
Represents 'error on the error'; standard deviation of $\pi_s(s)$ is σ_s.

A simple test

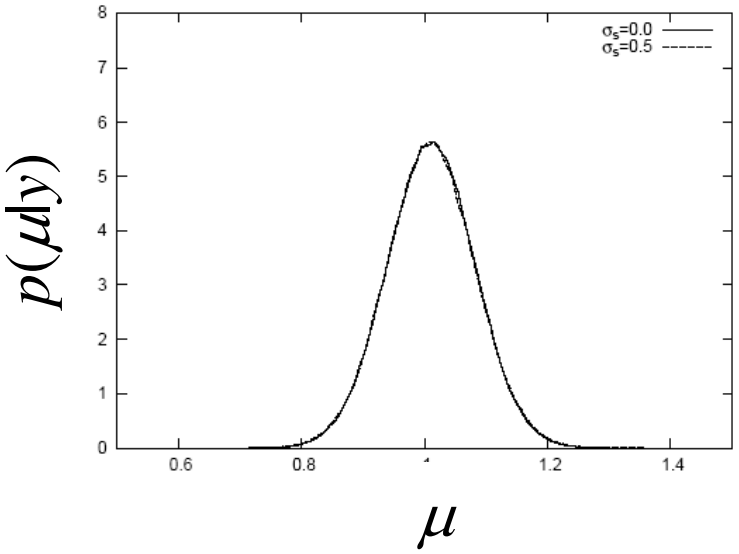
Suppose fit effectively averages four measurements.

Take $\sigma_{\rm sys} = \sigma_{\rm stat} = 0.1$, uncorrelated.

Case #1: data appear compatible



Posterior $p(\mu|y)$:



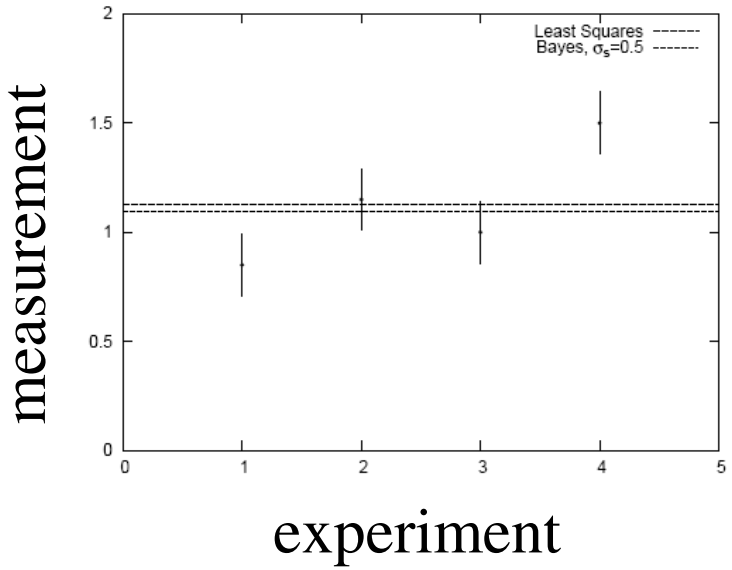
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$\sigma_s = 0.0: \quad \hat{\mu} = 1.000 \pm 0.071$

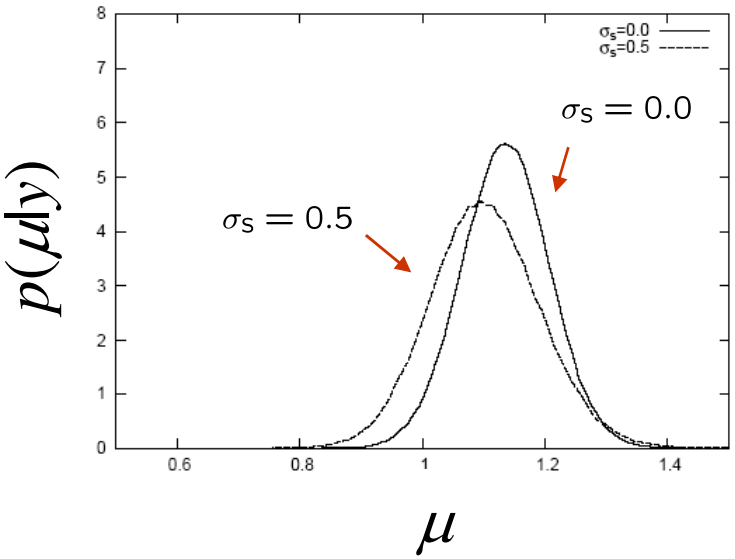
$\sigma_s = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$$\sigma_s = 0.0: \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_s = 0.5: \quad \hat{\mu} = 1.093 \pm 0.089$$

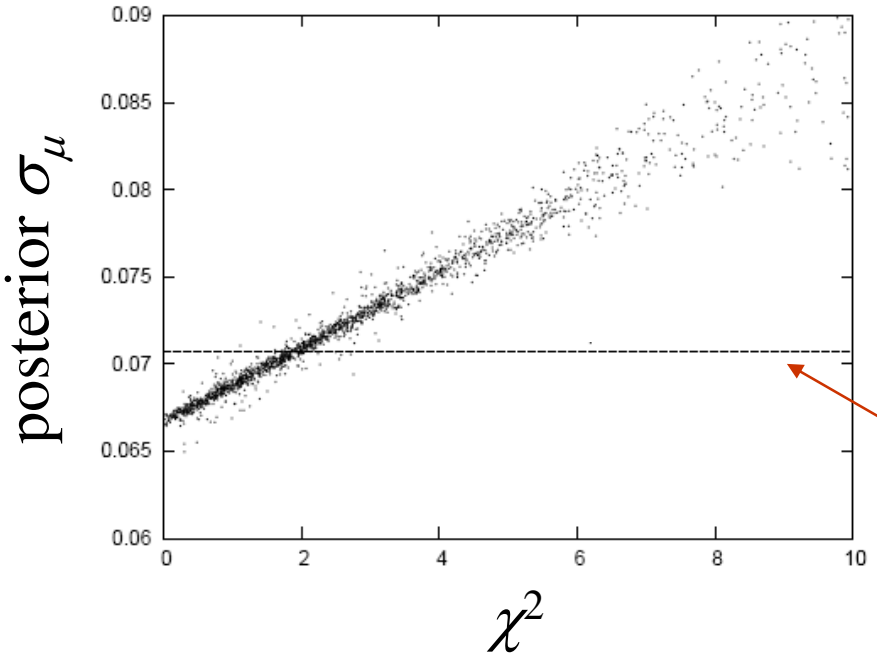
→ Bayesian fit less sensitive to outlier.

(See also D'Agostini 1999; Dose & von der Linden 1999)

Goodness-of-fit vs. size of error

In LS fit, value of minimized χ^2 does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high χ^2 corresponds to a larger error (and vice versa).



2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

σ_μ from least squares

Summary of lecture 1

The distinctive features of Bayesian statistics are:

Subjective probability used for hypotheses (e.g. a parameter).

Bayes' theorem relates the probability of data given H (the likelihood) to the posterior probability of H given data:

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

Requires prior probability for H

Bayesian methods often yield answers that are close (or identical) to those of frequentist statistics, albeit with different interpretation.

This is not the case when the prior information is important relative to that contained in the data.

Extra slides

Some Bayesian references

P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, CUP, 2005

D. Sivia, *Data Analysis: a Bayesian Tutorial*, OUP, 2006

S. Press, *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*, 2nd ed., Wiley, 2003

A. O'Hagan, Kendall's, *Advanced Theory of Statistics, Vol. 2B, Bayesian Inference*, Arnold Publishers, 1994

A. Gelman et al., *Bayesian Data Analysis*, 2nd ed., CRC, 2004

W. Bolstad, *Introduction to Bayesian Statistics*, Wiley, 2004

E.T. Jaynes, *Probability Theory: the Logic of Science*, CUP, 2003

Statistical Methods in Particle Physics
Lecture 2: Limits and Discovery

SUSSP65

St Andrews

16–29 August 2009

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Lecture #1: An introduction to Bayesian statistical methods

- Role of probability in data analysis (Frequentist, Bayesian)
- A simple fitting problem : Frequentist vs. Bayesian solution
- Bayesian computation, Markov Chain Monte Carlo

Lecture #2: Setting limits, making a discovery

- Frequentist vs Bayesian approach,
- treatment of systematic uncertainties

Lecture #3: Multivariate methods for HEP

- Event selection as a statistical test
- Neyman-Pearson lemma and likelihood ratio test
- Some multivariate classifiers:
  - NN, BDT, SVM, ...

Setting limits: Poisson data with background

Count n events, e.g., in fixed time or integrated luminosity.

s = expected number of signal events

b = expected number of background events

n ~ Poisson(s+b): $$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Suppose the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{\text{obs}} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

→ set upper limit on the parameter s, taking into consideration any uncertainty in b.

Setting limits

Frequentist intervals (limits) for a parameter s can be found by defining a test of the hypothesized value s (do this for all s):

> Specify values of the data n that are 'disfavoured' by s (critical region) such that $P(n \text{ in critical region}) \leq \gamma$ for a prespecified γ, e.g., 0.05 or 0.1.
>
> If n is observed in the critical region, reject the value s.

Now invert the test to define a confidence interval as:

> set of s values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

The interval will cover the true value of s with probability $\geq 1 - \gamma$.

Frequentist upper limit for Poisson parameter

First suppose that the expected background b is known.

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\text{obs}}; s, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\text{up}}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\text{obs}} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\text{up}} = -\ln \gamma \approx 3.00$$

$[0, s_{\text{up}}]$ is an example of a confidence interval. It is designed to include the true value of s with probability at least $1-\gamma$ for any s.

Calculating Poisson parameter limits

Analogous procedure for lower limit $s_{\rm lo}$.

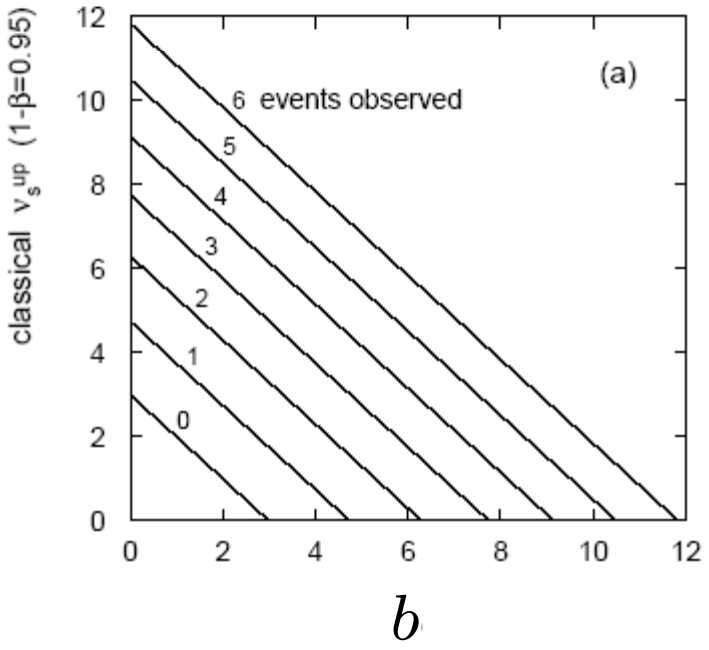
To solve for $s_{\rm lo}$, $s_{\rm up}$, can exploit relation to χ^2 distribution:

$$s_{\rm lo} = \frac{1}{2} F^{-1}_{\chi^2}(\alpha; 2n) - b$$

Quantile of χ^2 distribution

$$s_{\rm up} = \frac{1}{2} F^{-1}_{\chi^2}(1-\beta; 2(n+1)) - b$$

For low fluctuation of n this can give negative result for $s_{\rm up}$; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (\text{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when limit of parameter is close to a physical boundary, cf. m_ν estimated using $E^2 - p^2$.

Expected limit for on s if $s = 0$

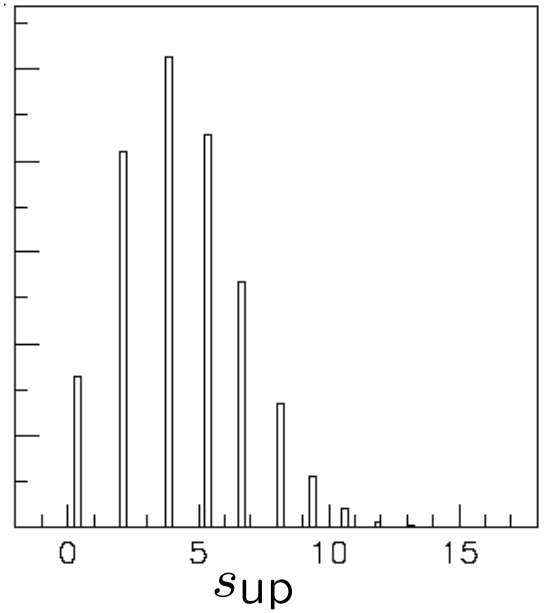
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

Even better: for CL = 0.917923 we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



Likelihood ratio limits (Feldman-Cousins)

Define likelihood ratio for hypothesized parameter value s:

$$l(s) = \frac{L(n|s,b)}{L(n|\hat{s},b)} \quad \text{where} \quad \hat{s} = \begin{cases} n-b & n \geq b, \\ 0 & \text{otherwise} \end{cases}$$

Here $\hat{s}$ is the ML estimator, note $0 \leq l(s) \leq 1$.

Define a statistical test for a hypothetical value of s:

Rejection region defined by low values of likelihood ratio.

Reject s if p-value = $P(l(s) \leq l_{\text{obs}} \mid s)$ is less than γ (e.g. $\gamma = 0.05$).

Confidence interval at CL = $1-\gamma$ is the set of s values not rejected.

Resulting intervals can be one- or two-sided (depending on n).

(Re)discovered for HEP by Feldman and Cousins,
Phys. Rev. D 57 (1998) 3873.

More on intervals from LR test (Feldman-Cousins)

Caveat with coverage: suppose we find $n >> b$.
Usually one then quotes a measurement: $\hat{s} = n - b\,, \quad \hat{\sigma}_{\hat{s}} = \sqrt{n}$

If, however, n isn't large enough to claim discovery, one sets a limit on s.

FC pointed out that if this decision is made based on n, then the actual coverage probability of the interval can be less than the stated confidence level ('flip-flopping').

FC intervals remove this, providing a smooth transition from 1- to 2-sided intervals, depending on n.

But, suppose FC gives e.g. $0.1 < s < 5$ at 90% CL,
p-value of s=0 still substantial. Part of upper-limit 'wasted'?

Nuisance parameters and limits

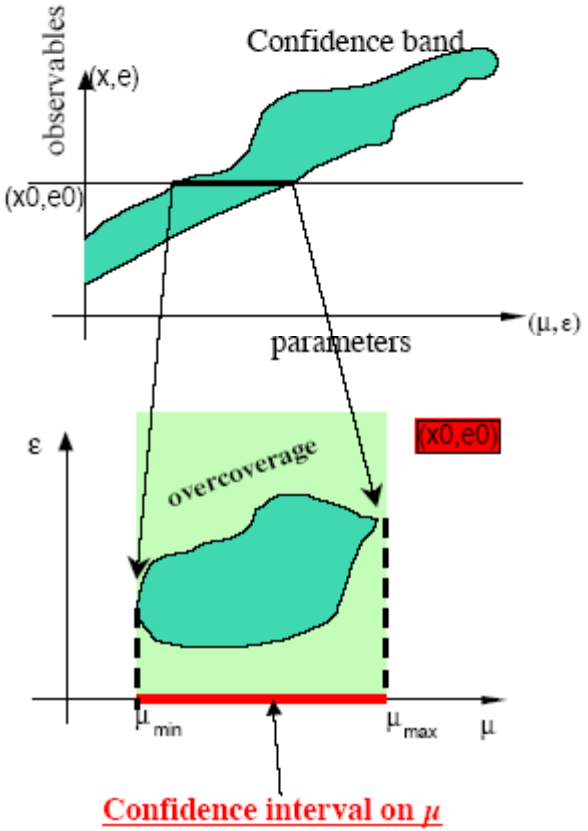
In general we don't know the background b perfectly.

Suppose we have a measurement of b, e.g., $b_{\text{meas}} \sim N(b, \sigma_b)$

So the data are really: n events and the value b_{meas}.

In principle the confidence interval recipe can be generalized to two measurements and two parameters.

Difficult and rarely attempted, but see e.g. talks by K. Cranmer at PHYSTAT03 and by G. Punzi at PHYSTAT05.



G. Punzi, PHYSTAT05

Nuisance parameters and profile likelihood

Suppose model has likelihood function

$$L(\mu, \nu) = P(\vec{x}|\mu, \nu)$$

Parameters of interest Nuisance parameters

Define the profile likelihood ratio as

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\nu}})}{L(\hat{\mu}, \hat{\nu})}$$

Maximizes L for given value of μ

Maximizes L

$\lambda(\mu)$ reflects level of agreement between data and μ $(0 \leq \lambda(\mu) \leq 1)$

Equivalently use $q_\mu = -2 \ln \lambda(\mu)$

p-value from profile likelihood ratio

Large q_μ means worse agreement between data and μ

p-value = Prob(data with $\leq$ compatibility with μ when compared to the data we got | μ)

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu \approx 1 - F_{\chi^2_n}(q_{\mu,\mathrm{obs}})$$

rapidly approaches chi-square pdf (Wilks' theorem)

chi-square cumulative distribution, degrees of freedom = dimension of μ

Reject μ if $p_\mu < \gamma = 1 - \mathrm{CL}$

(Approx.) confidence interval for μ = set of μ values not rejected.

Coverage not exact for all ν but very good if $\nu \approx \hat{\nu}$.

The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf $p(\theta\,|\,x)$ to give interval with any desired probability content.

For e.g. Poisson parameter 95% CL upper limit from

$$0.95 = \int_{-\infty}^{s_{\mathrm{up}}} p(s|n)\,ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s<0$.

Often try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian interval with flat prior for s

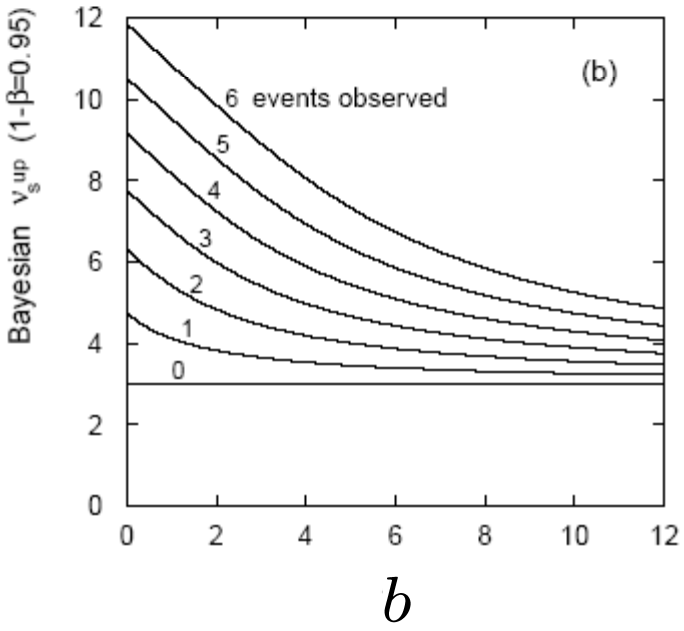
Solve numerically to find limit s_{up}.

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as classical case ('coincidence').

Otherwise Bayesian limit is everywhere greater than classical ('conservative').

Never goes negative.

Doesn't depend on b if $n = 0$.



Bayesian limits with uncertainty on b

Uncertainty on b goes into the prior, e.g.,

$$\pi(s,b) = \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)}$$

$$\pi_s(s) = \text{const.}, \quad \sim 1/s, \ldots$$

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}$$

Put this into Bayes' theorem,

$$p(s,b|n) \propto L(n|s,b)\pi(s,b)$$

Marginalize over b, then use $p(s|n)$ to find intervals for s with any desired probability content.

Controversial part here is prior for signal $\pi_s(s)$
(treatment of nuisance parameters is easy).

Frequentist discovery, p-values

To discover e.g. the Higgs, try to reject the background-only (null) hypothesis (H_0).

Define a statistic t whose value reflects compatibility of data with H_0.

p-value = Prob(data with $\leq$ compatibility with H_0 when compared to the data we got | H_0)

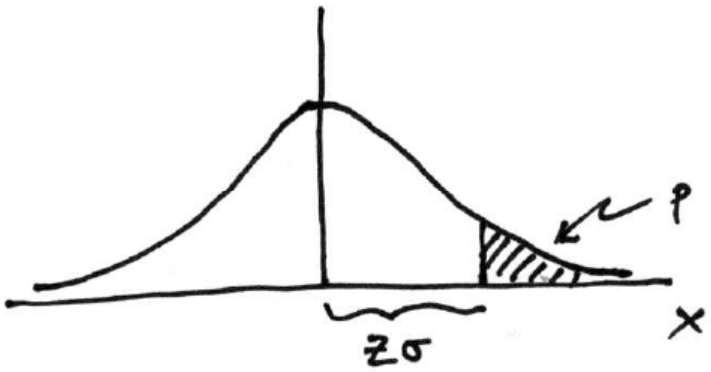
For example, if high values of t mean less compatibility,

$$p = \int_t^\infty f(t'|H_0)\, dt' \ .$$

If p-value comes out small, then this is evidence against the background-only hypothesis → discovery made!

Significance from p-value

Define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ **TMath::Prob**

$$Z = \Phi^{-1}(1-p)$$ **TMath::NormQuantile**

When to publish

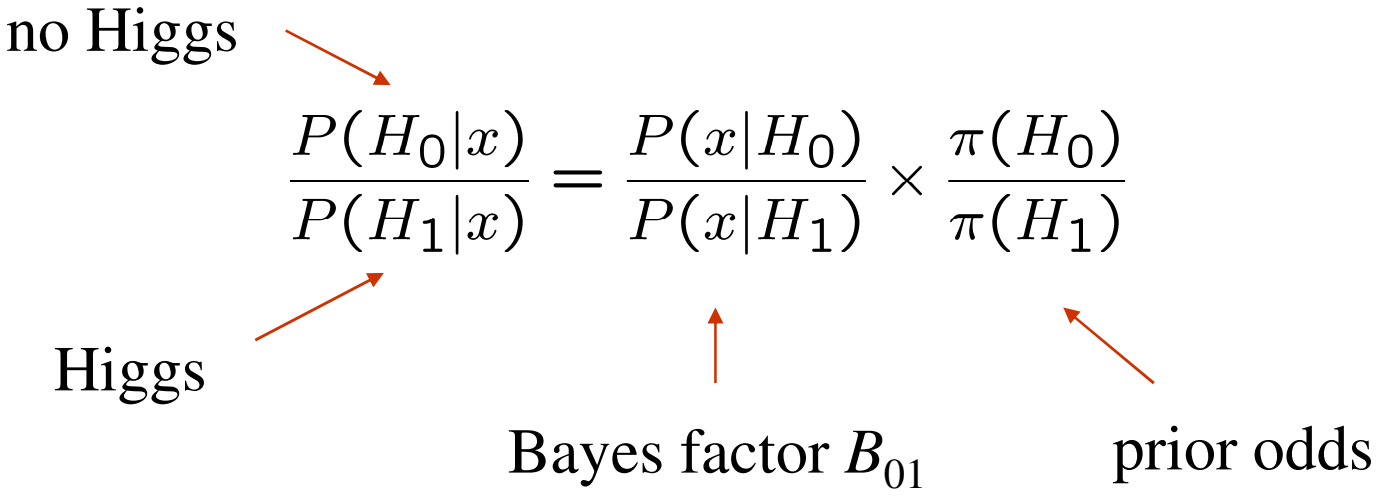
HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable p-value for discovery
$D^0\bar{D}^0$ mixing	~0.05
Higgs	~ 10^{-7} (?)
Life on Mars	~10^{-10}
Astrology	~10^{-20}

Bayesian model selection ('discovery')

The probability of hypothesis H_0 relative to its complementary alternative H_1 is often given by the posterior odds:



The Bayes factor is regarded as measuring the weight of evidence of the data in support of H_0 over H_1.

Interchangeably use $B_{10} = 1/B_{01}$

Assessing Bayes factors

One can use the Bayes factor much like a p-value (or Z value).

There is an "established" scale, analogous to our 5σ rule:

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

Will this be adopted in HEP?

Rewriting the Bayes factor

Suppose we have models H_i, $i = 0, 1, \ldots,$

each with a likelihood $p(x|H_i, \vec{\theta}_i)$

and a prior pdf for its internal parameters $\pi_i(\vec{\theta}_i)$

so that the full prior is $\pi(H_i, \vec{\theta}_i) = p_i \pi_i(\vec{\theta}_i)$

where $p_i = P(H_i)$ is the overall prior probability for H_i.

The Bayes factor comparing H_i and H_j can be written

$$B_{ij} = \frac{P(H_i|\vec{x})}{P(H_i)} \Big/ \frac{P(H_j|\vec{x})}{P(H_j)}$$

Bayes factors independent of $P(H_i)$

For B_{ij} we need the posterior probabilities marginalized over all of the internal parameters of the models:

$$P(H_i|\vec{x}) = \int P(H_i, \vec{\theta}_i|\vec{x})\, d\vec{\theta}_i$$

Use Bayes theorem

$$= \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) p_i \pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{P(x)}$$

So therefore the Bayes factor is

Ratio of marginal likelihoods

$$B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) \pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j) \pi_j(\vec{\theta}_j)\, d\vec{\theta}_j}$$

The prior probabilities $p_i = P(H_i)$ cancel.

Numerical determination of Bayes factors

Both numerator and denominator of B_{ij} are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\, d\vec{\theta} \quad \longleftarrow \quad \text{'marginal likelihood'}$$

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

Harmonic Mean (and improvements)
Importance sampling
Parallel tempering (~thermodynamic integration)
Nested sampling
...

See e.g. Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

Example of systematics in a search

Combination of Higgs boson search channels (ATLAS)

Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics, arXiv:0901.0512, CERN-OPEN-2008-20.

Standard Model Higgs channels considered (more to be used later):

$H \to \gamma\gamma$

$H \to WW^{(*)} \to e\nu\mu\nu$

$H \to ZZ^{(*)} \to 4l$ (l = e, μ)

$H \to \tau^+\tau^- \to ll, lh$

Used profile likelihood method for systematic uncertainties:

background rates, signal & background shapes.

Statistical model for Higgs search

Bin i of a given channel has n_i events, expectation value is

$$E[n_i] = \mu L \varepsilon_i \sigma_i \mathcal{B} + b_i \equiv \mu s_i + b_i$$

μ is global strength parameter, common to all channels.
$\mu = 0$ means background only, $\mu = 1$ is SM hypothesis.

Expected signal and background are:

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s)\, dx \ ,$$

$$b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b)\, dx$$

b_{tot}, $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$ are nuisance parameters

The likelihood function

The single-channel likelihood function uses Poisson model for events in signal and control histograms:

data in signal histogram

data in control histogram

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

here signal rate is only parameter of interest

$\boldsymbol{\theta}$ represents all nuisance parameters, e.g., background rate, shapes

There is a likelihood $L_i(\mu, \boldsymbol{\theta}_i)$ for each channel, $i = 1, \ldots, N$.

The full likelihood function is $L(\mu, \boldsymbol{\theta}) = \prod_i L_i(\mu, \boldsymbol{\theta}_i)$

Profile likelihood ratio

To test hypothesized value of μ, construct profile likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

$\hat{\hat{\boldsymbol{\theta}}}$: Maximized L for given μ

$\hat{\mu}, \hat{\boldsymbol{\theta}}$: Maximized L

Equivalently use $q_\mu = -2 \ln \lambda(\mu)$:

data agree well with hypothesized $\mu \rightarrow q_\mu$ small

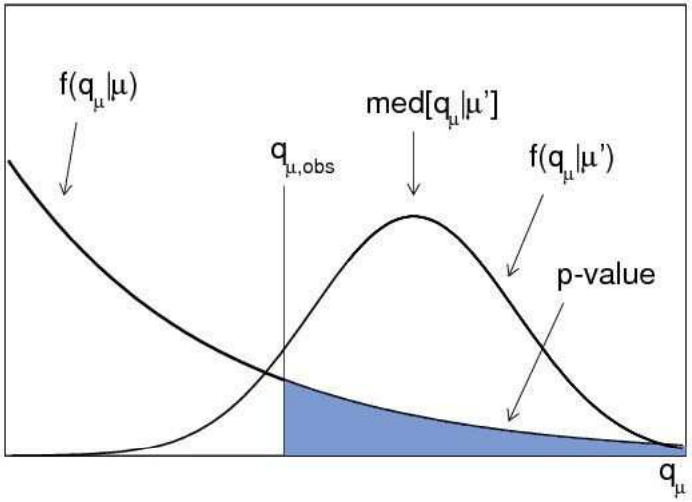
data disagree with hypothesized $\mu \rightarrow q_\mu$ large

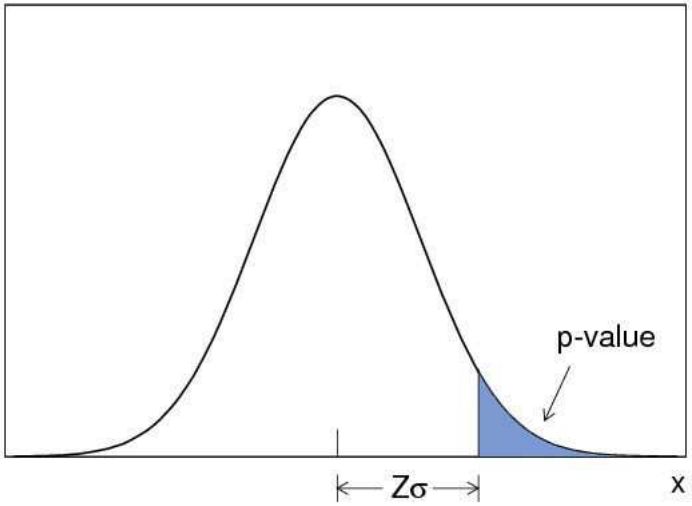
Distribution of q_μ under assumption of μ related to chi-square (Wilks' theorem, here approximation valid for roughly $L > 2$ fb^{-1}):

$$f(q_\mu|\mu) \approx \frac{1}{2} f_{\chi^2_1}(q_\mu) + \frac{1}{2}\delta(q_\mu)$$

p-value / significance of hypothesized μ

Test hypothesized μ by giving p-value, probability to see data with $\leq$ compatibility with μ compared to data observed:





Equivalently use significance, Z, defined as equivalent number of sigmas for a Gaussian fluctuation in one direction:

$$Z = \Phi^{-1}(1-p)$$

Sensitivity

Discovery:

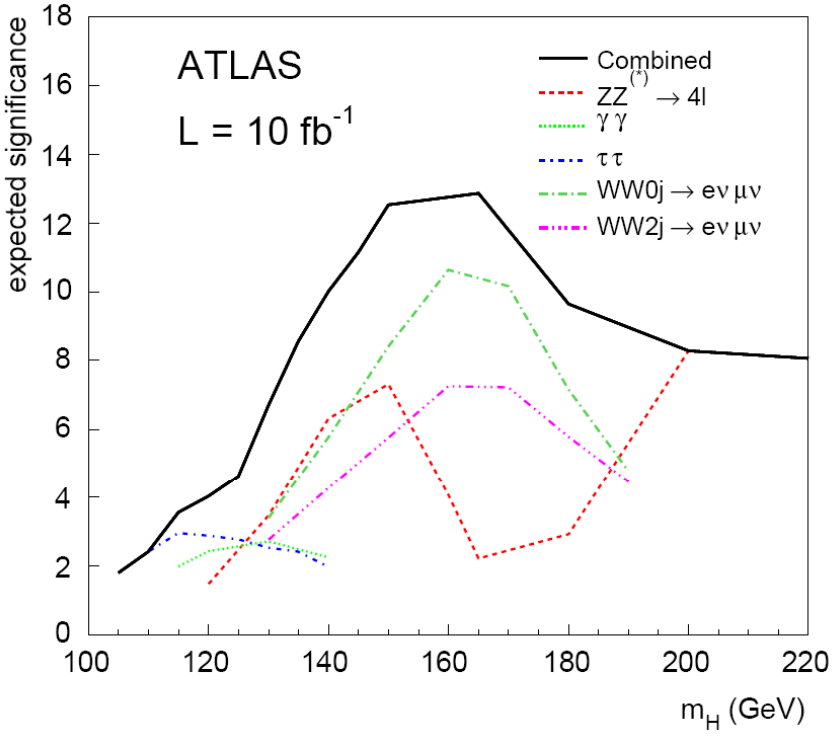
Generate data under s+b ($\mu = 1$) hypothesis;
Test hypothesis $\mu = 0 \rightarrow p$-value $\rightarrow Z$.

Exclusion:

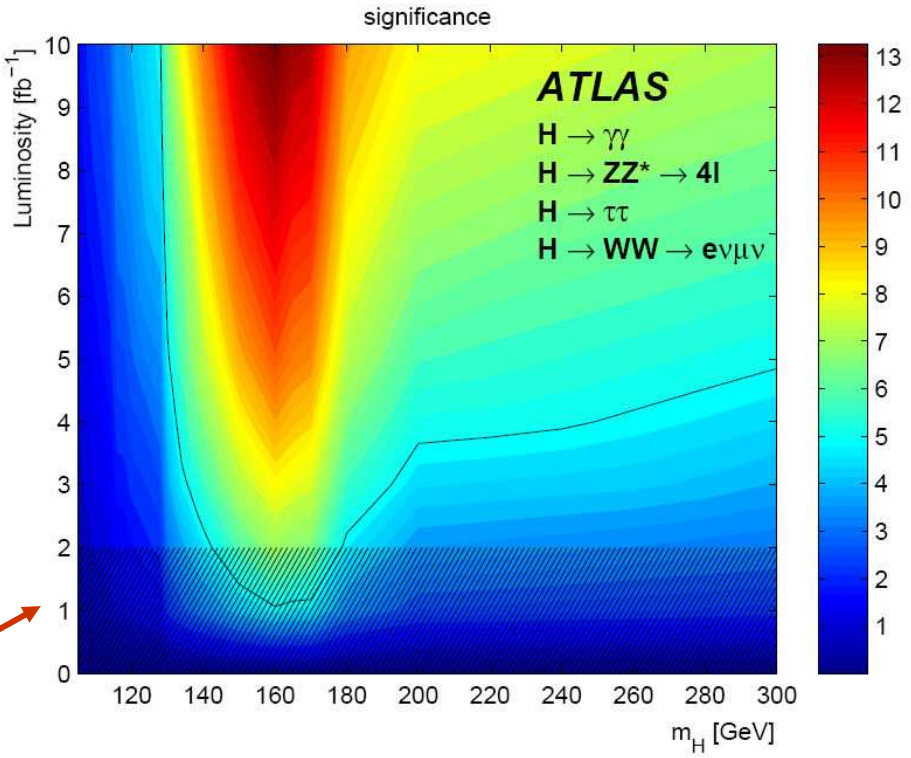
Generate data under background-only ($\mu = 0$) hypothesis;
Test hypothesis $\mu = 1$.
If $\mu = 1$ has p-value < 0.05 exclude m_{H} at 95% CL.

Presence of nuisance parameters leads to broadening of the profile likelihood, reflecting the loss of information, and gives appropriately reduced discovery significance, weaker limits.

Combined discovery significance



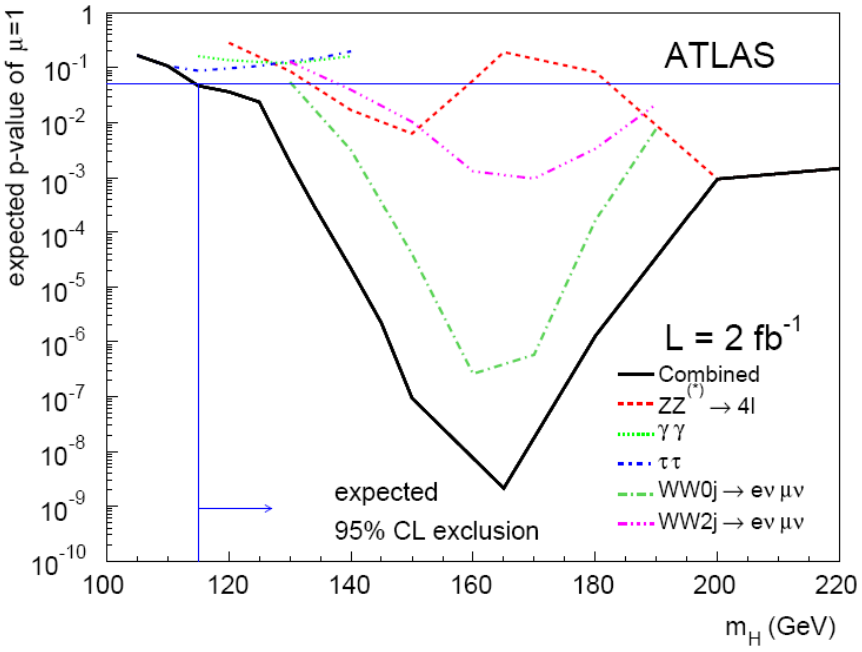
Discovery signficance (in colour) vs. L, m_H:

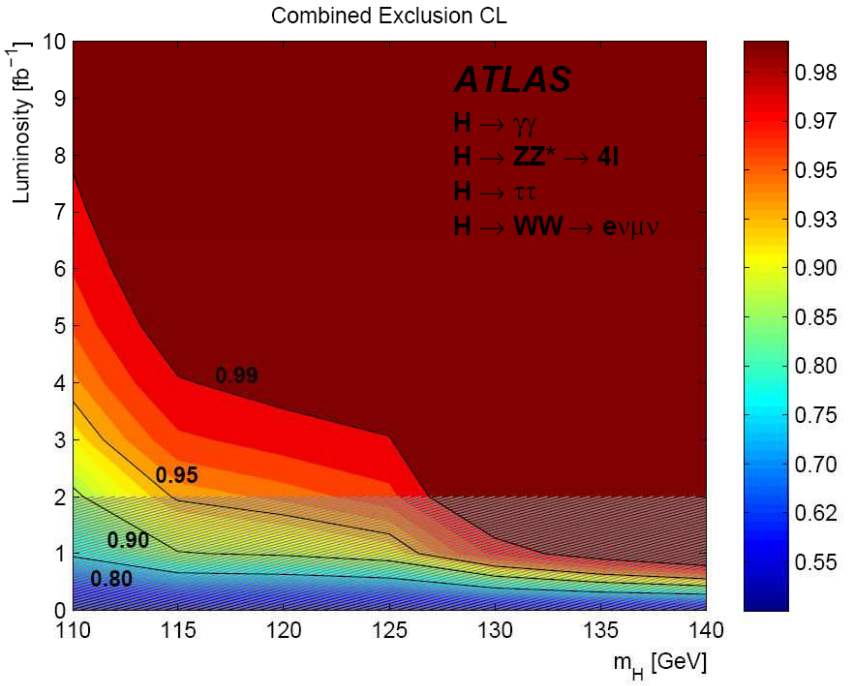


Approximations used here not always accurate for $L < 2$ fb^{-1} but in most cases conservative.

Combined 95% CL exclusion limits

1 – p-value of m_H (in colour) vs. L, m_H:





Summary on limits

Different sorts of limits answer different questions.

A frequentist confidence interval does not (necessarily) answer, "What do we believe the parameter's value is?"

Look at sensitivity, e.g., $E[s_{\text{up}} \mid s = 0]$; consider also:

need for consensus/conventions;
convenience and ability to combine results, ...

For any result, consumer will compute (mentally or otherwise):

$$p(\theta|\text{result}) \propto L(\text{result}|\theta)\pi(\theta)$$

Need likelihood (or summary thereof).

consumer's prior

Summary on discovery

Current convention: p-value of background-only $< 2.9 \times 10^{-7}$ (5σ)

This should really depend also on other factors:

Plausibility of signal
Confidence in modeling of background

Can also use Bayes factor

$$B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i)\pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j)\pi_j(\vec{\theta}_j)\, d\vec{\theta}_j}$$

Should hopefully point to same conclusion as p-value.

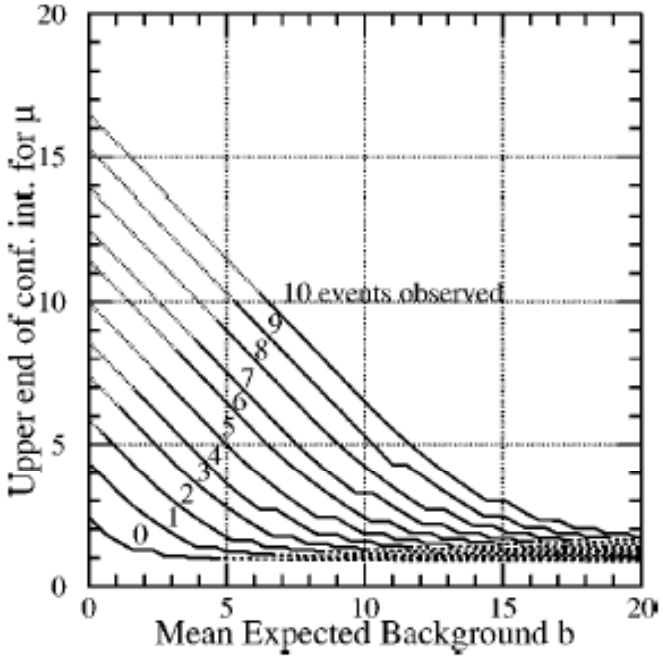
If not, need to understand why!

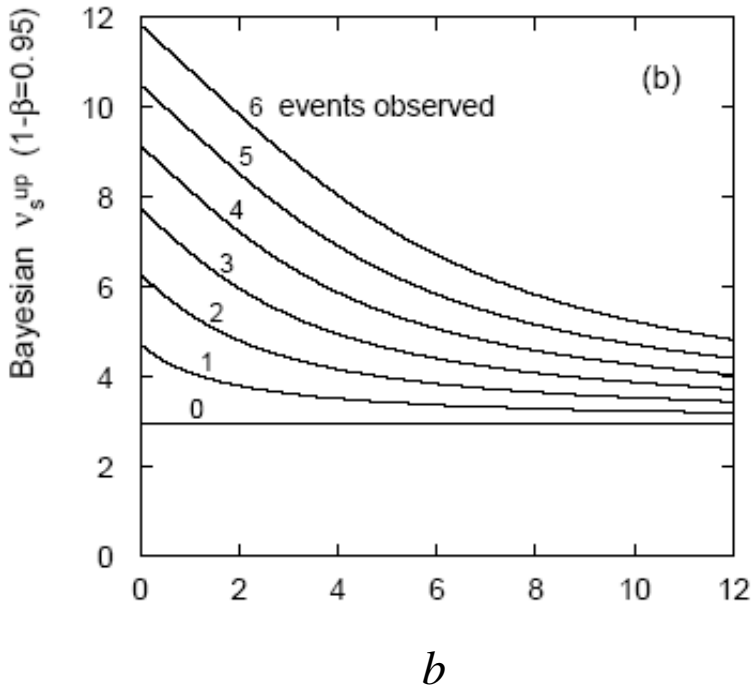
As yet not widely used in HEP, numerical issues not easy.

Extra slides

Upper limit versus *b*

Feldman & Cousins, PRD 57 (1998) 3873





If $n = 0$ observed, should upper limit depend on b?

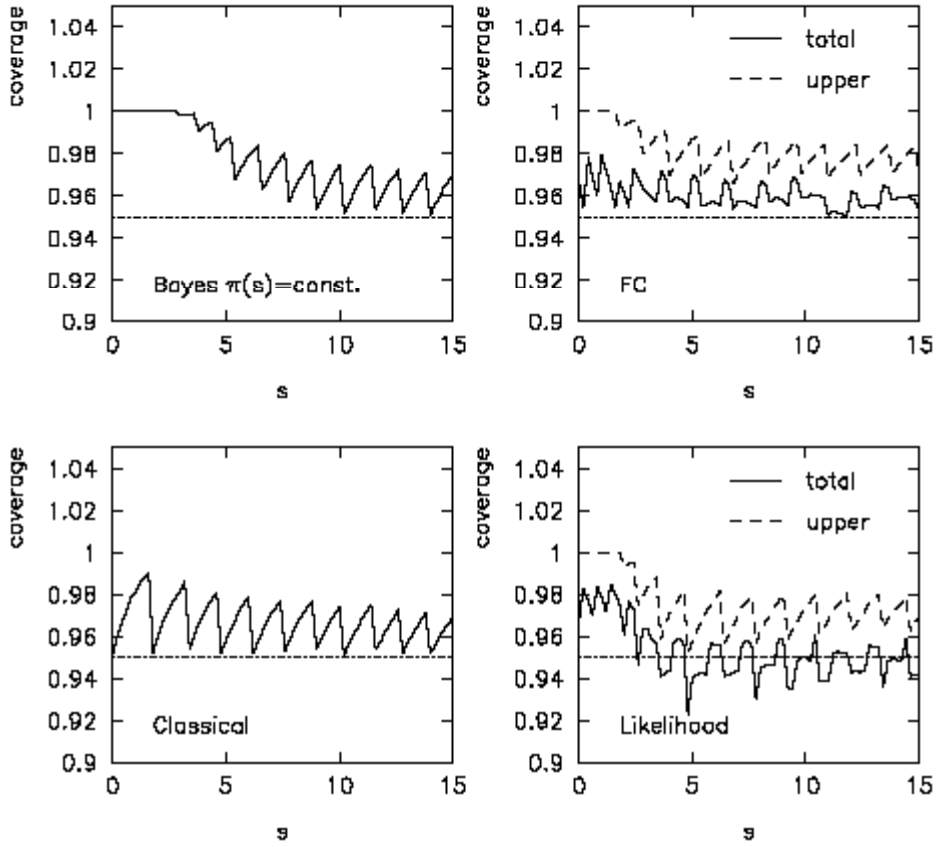
Classical: yes

Bayesian: no

FC: yes

Coverage probability of confidence intervals

Because of discreteness of Poisson data, probability for interval to include true value in general > confidence level ('over-coverage')



Cousins-Highland method

Regard b as 'random', characterized by pdf $\pi(b)$.

Makes sense in Bayesian approach, but in frequentist model b is constant (although unknown).

A measurement b_{meas} is random but this is not the mean number of background events, rather, b is.

Compute anyway $P(n;s) = \int P(n;s,b)\pi_b(b)\,db$

This would be the probability for n if Nature were to generate a new value of b upon repetition of the experiment with $\pi_b(b)$.

Now e.g. use this $P(n;s)$ in the classical recipe for upper limit at CL = $1 - \beta$: $\beta = P(n \leq n_{\text{obs}}; s_{\text{up}})$

Result has hybrid Bayesian/frequentist character.

'Integrated likelihoods'

Consider again signal s and background b, suppose we have uncertainty in b characterized by a prior pdf $\pi_b(b)$.

Define integrated likelihood as $L'(s) = \int L(s,b)\pi_b(b)\,db$,
also called modified profile likelihood, in any case not a real likelihood.

Now use this to construct likelihood ratio test and invert to obtain confidence intervals.

Feldman-Cousins & Cousins-Highland (FHC^2), see e.g. J. Conrad et al., Phys. Rev. D67 (2003) 012002 and Conrad/Tegenfeldt PHYSTAT05 talk.

Calculators available (Conrad, Tegenfeldt, Barlow).

Analytic formulae for limits

There are a number of papers describing Bayesian limits for a variety of standard scenarios

- Several conventional priors
- Systematics on efficiency, background
- Combination of channels

and (semi-)analytic formulae and software are provided.

Joel Heinrich, *Bayesian limit software: multi-channel with correlated backgrounds and efficiencies*, CDF/MEMO/STATISTICS/PUBLIC/7587 (2005).

Joel Heinrich et al., *Interval estimation in the presence of nuisance parameters. 1. Bayesian approach*, CDF/MEMO/STATISTICS/PUBLIC/7117, physics/0409129 (2004).

Luc Demortier, *A Fully Bayesian Computation of Upper Limits for Poisson Processes*, CDF/MEMO/STATISTICS/PUBLIC/5928 (2004).

But for more general cases we need to use numerical methods (e.g. L.D. uses importance sampling).

Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\boldsymbol{\theta})$ is normalized to unity so integrate both sides,

posterior expectation

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L]$$

Therefore sample $\boldsymbol{\theta}$ from the posterior via MCMC and estimate m with one over the average of $1/L$ (the harmonic mean of L).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). Gelfand & Dey propose variant:

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\boldsymbol{\theta})$:

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over $\boldsymbol{\theta}$:

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x}) = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right]$$

Improved convergence if tails of $f(\boldsymbol{\theta})$ fall off faster than $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Note harmonic mean estimator is special case $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

Importance sampling

Need pdf $f(\boldsymbol{\theta})$ which we can evaluate at arbitrary $\boldsymbol{\theta}$ and also sample with MC.

The marginal likelihood can be written

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_f\left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\right]$$

Best convergence when $f(\boldsymbol{\theta})$ approximates shape of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Use for $f(\boldsymbol{\theta})$ e.g. multivariate Gaussian with mean and covariance estimated from posterior (e.g. with MINUIT).

Statistical Methods in Particle Physics
Lecture 3: Multivariate Methods

SUSSP65

St Andrews

16–29 August 2009

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Lecture #1: An introduction to Bayesian statistical methods

- Role of probability in data analysis (Frequentist, Bayesian)
- A simple fitting problem : Frequentist vs. Bayesian solution
- Bayesian computation, Markov Chain Monte Carlo

Lecture #2: Setting limits, making a discovery

- Frequentist vs Bayesian approach,
- treatment of systematic uncertainties

Lecture #3: Multivariate methods for HEP

- Event selection as a statistical test
- Neyman-Pearson lemma and likelihood ratio test
- Some multivariate classifiers:
  - NN, BDT, SVM, ...

Resources on multivariate methods

Books:

C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006

T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2001

R. Duda, P. Hart, D. Stork, *Pattern Classification*, 2nd ed., Wiley, 2001

A. Webb, *Statistical Pattern Recognition*, 2nd ed., Wiley, 2002

Materials from some recent meetings:

PHYSTAT conference series (2002, 2003, 2005, 2007,...) see
`www.phystat.org`

Caltech workshop on multivariate analysis, 11 February, 2008
`indico.cern.ch/conferenceDisplay.py?confId=27385`

SLAC Lectures on Machine Learning by Ilya Narsky (2006)
`www-group.slac.stanford.edu/sluo/Lectures/Stat2006_Lectures.html`

Software for multivariate analysis

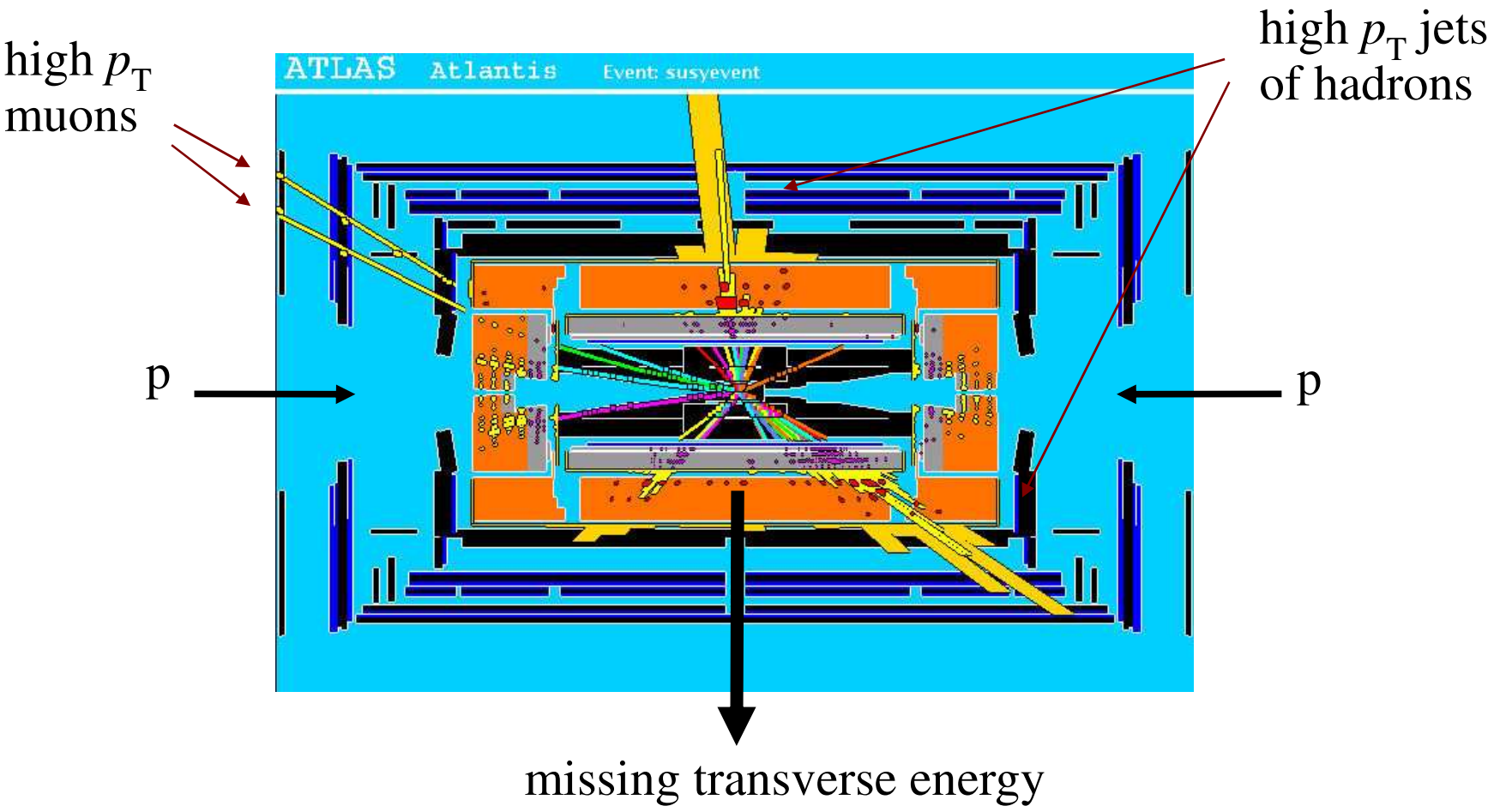
TMVA, Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

From **tmva.sourceforge.net**, also distributed with ROOT
Variety of classifiers
Good manual

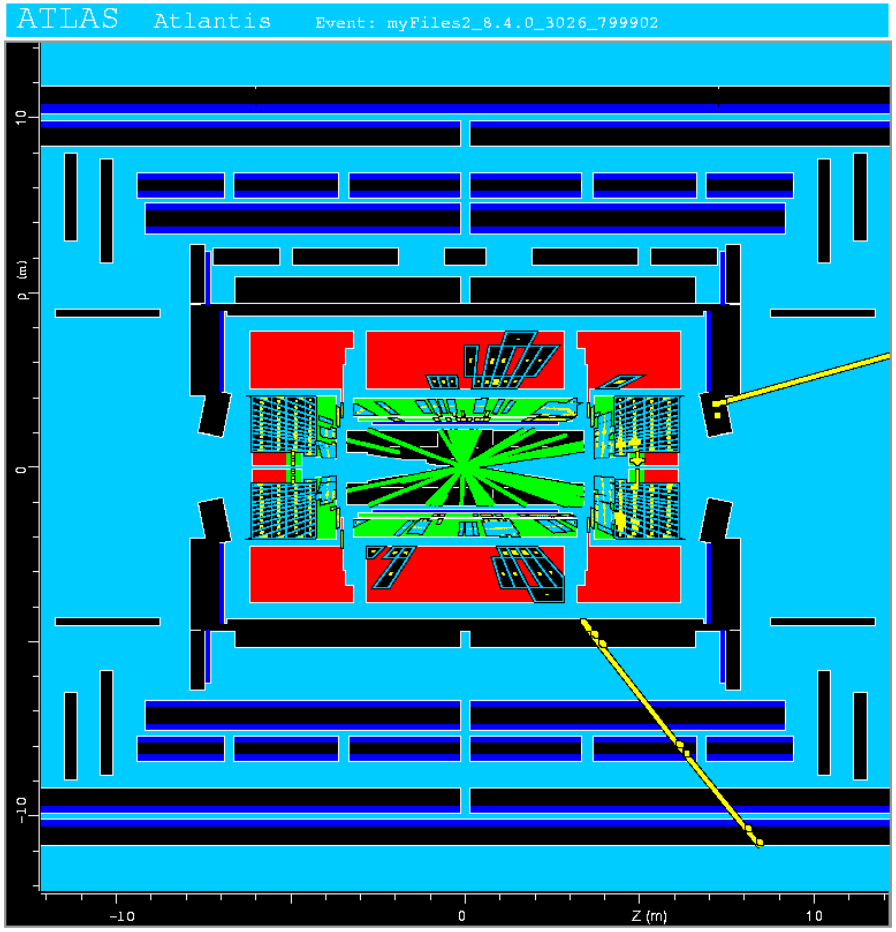
StatPatternRecognition, I. Narsky, physics/0507143

Further info from **www.hep.caltech.edu/~narsky/spr.html**
Also wide variety of methods, many complementary to **TMVA**
Currently appears project no longer to be supported

A simulated SUSY event in ATLAS



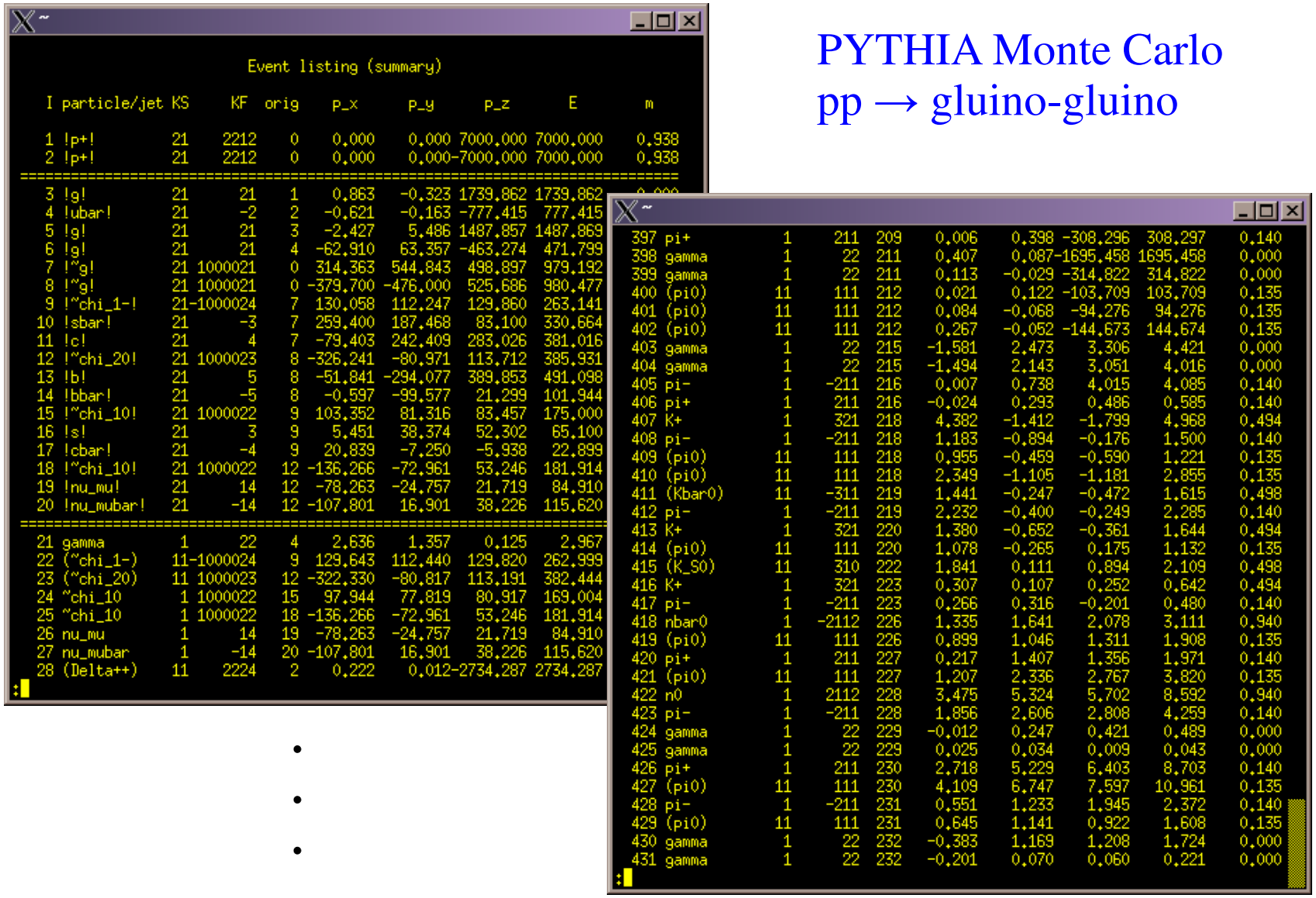
Background events



This event from Standard Model ttbar production also has high p_T jets and muons, and some missing transverse energy.

→ can easily mimic a SUSY event.

A simulated event



Event listing (summary)

I	particle/jet	KS	KF	orig	p_x	p_y	p_z	E	m
1	!p+!	21	2212	0	0.000	0.000	7000.000	7000.000	0.938
2	!p+!	21	2212	0	0.000	0.000	-7000.000	7000.000	0.938
3	!g!	21	21	1	0.863	-0.323	1739.862	1739.862	0.000
4	!ubar!	21	-2	2	-0.621	-0.163	-777.415	777.415	0.000
5	!g!	21	21	3	-2.427	5.486	1487.857	1487.869	0.000
6	!g!	21	21	4	-62.910	63.357	-463.274	471.799	0.000
7	!~g!	21	1000021	0	314.363	544.843	498.897	979.192	0.000
8	!~g!	21	1000021	0	-379.700	-476.000	525.686	980.477	0.000
9	!~chi_1-!	21	-1000024	7	130.058	112.247	129.860	263.141	0.000
10	!sbar!	21	-3	7	259.400	187.468	83.100	330.664	0.000
11	!c!	21	4	7	-79.403	242.409	283.026	381.016	0.000
12	!~chi_20!	21	1000023	8	-326.241	-80.971	113.712	385.931	0.000
13	!b!	21	5	8	-51.841	-294.077	389.853	491.098	0.000
14	!bbar!	21	-5	8	-0.597	-99.577	21.299	101.944	0.000
15	!~chi_10!	21	1000022	9	103.352	81.316	83.457	175.000	0.000
16	!s!	21	3	9	5.451	38.374	52.302	65.100	0.000
17	!cbar!	21	-4	9	20.839	-7.250	-5.938	22.899	0.000
18	!~chi_10!	21	1000022	12	-136.266	-72.961	53.246	181.914	0.000
19	!nu_mu!	21	14	12	-78.263	-24.757	21.719	84.910	0.000
20	!nu_mubar!	21	-14	12	-107.801	16.901	38.226	115.620	0.000
21	gamma	1	22	4	2.636	1.357	0.125	2.967	0.000
22	(~chi_1-)	11	-1000024	9	129.643	112.440	129.820	262.999	0.000
23	(~chi_20)	11	1000023	12	-322.330	-80.817	113.191	382.444	0.000
24	~chi_10	1	1000022	15	97.944	77.819	80.917	169.004	0.000
25	~chi_10	1	1000022	18	-136.266	-72.961	53.246	181.914	0.000
26	nu_mu	1	14	19	-78.263	-24.757	21.719	84.910	0.000
27	nu_mubar	1	-14	20	-107.801	16.901	38.226	115.620	0.000
28	(Delta++)	11	2224	2	0.222	0.012	-2734.287	2734.287	0.000

397	pi+	1	211	209	0.006	0.398	-308.296	308.297	0.140
398	gamma	1	22	211	0.407	0.087	-1695.458	1695.458	0.000
399	gamma	1	22	211	0.113	-0.029	-314.822	314.822	0.000
400	(pi0)	11	111	212	0.021	0.122	-103.709	103.709	0.135
401	(pi0)	11	111	212	0.084	-0.068	-94.276	94.276	0.135
402	(pi0)	11	111	212	0.267	-0.052	-144.673	144.674	0.135
403	gamma	1	22	215	-1.581	2.473	3.306	4.421	0.000
404	gamma	1	22	215	-1.494	2.143	3.051	4.016	0.000
405	pi-	1	-211	216	0.007	0.738	4.015	4.085	0.140
406	pi+	1	211	216	-0.024	0.293	0.486	0.585	0.140
407	K+	1	321	218	4.382	-1.412	-1.799	4.968	0.494
408	pi-	1	-211	218	1.183	-0.894	-0.176	1.500	0.140
409	(pi0)	11	111	218	0.955	-0.459	-0.590	1.221	0.135
410	(pi0)	11	111	218	2.349	-1.105	-1.181	2.855	0.135
411	(Kbar0)	11	-311	219	1.441	-0.247	-0.472	1.615	0.498
412	pi-	1	-211	219	2.232	-0.400	-0.249	2.285	0.140
413	K+	1	321	220	1.380	-0.652	-0.361	1.644	0.494
414	(pi0)	11	111	220	1.078	-0.265	0.175	1.132	0.135
415	(K_S0)	11	310	222	1.841	0.111	0.894	2.109	0.498
416	K+	1	321	223	0.307	0.107	0.252	0.642	0.494
417	pi-	1	-211	223	0.266	0.316	-0.201	0.480	0.140
418	nbar0	1	-2112	226	1.335	1.641	2.078	3.111	0.940
419	(pi0)	11	111	226	0.899	1.046	1.311	1.908	0.135
420	pi+	1	211	227	0.217	1.407	1.356	1.971	0.140
421	(pi0)	11	111	227	1.207	2.336	2.767	3.820	0.135
422	n0	1	2112	228	3.475	5.324	5.702	8.592	0.940
423	pi-	1	-211	228	1.856	2.606	2.808	4.259	0.140
424	gamma	1	22	229	-0.012	0.247	0.421	0.489	0.000
425	gamma	1	22	229	0.025	0.034	0.009	0.043	0.000
426	pi+	1	211	230	2.718	5.229	6.403	8.703	0.140
427	(pi0)	11	111	230	4.109	6.747	7.597	10.961	0.135
428	pi-	1	-211	231	0.551	1.233	1.945	2.372	0.140
429	(pi0)	11	111	231	0.645	1.141	0.922	1.608	0.135
430	gamma	1	22	232	-0.383	1.169	1.208	1.724	0.000
431	gamma	1	22	232	-0.201	0.070	0.060	0.221	0.000

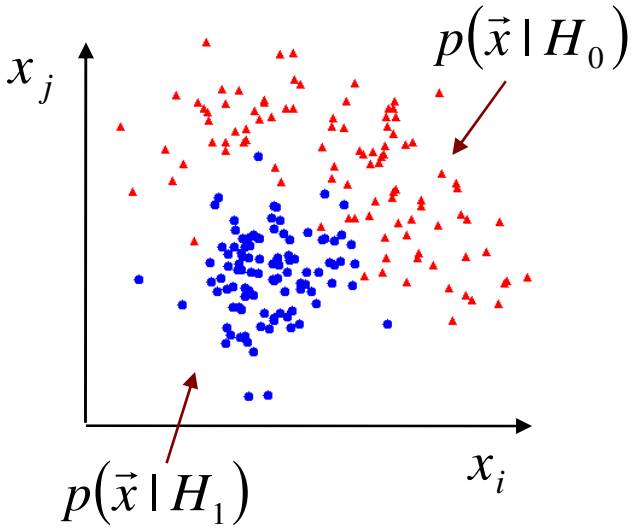
PYTHIA Monte Carlo
pp → gluino-gluino

Event selection as a statistical test

For each event we measure a set of numbers: $\vec{x} = (x_1, \ldots, x_n)$

x_1 = jet p_T
x_2 = missing energy
x_3 = particle i.d. measure, ...

$\vec{x}$ follows some n-dimensional joint probability density, which depends on the type of event produced, i.e., was it $pp \to t\bar{t}$, $pp \to \tilde{g}\tilde{g}$,…

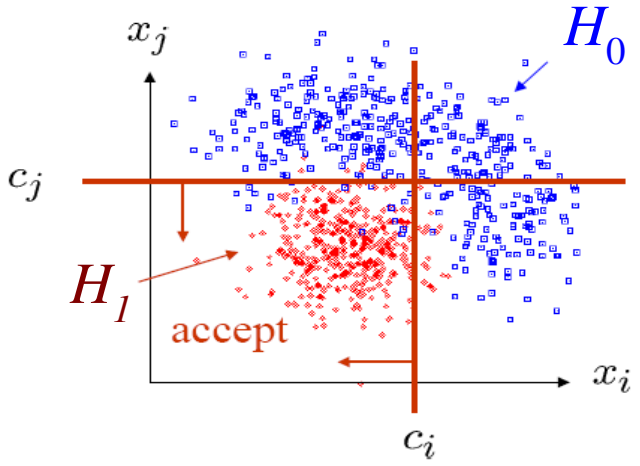


E.g. hypotheses H_0, H_1, ...
Often simply "signal", "background"

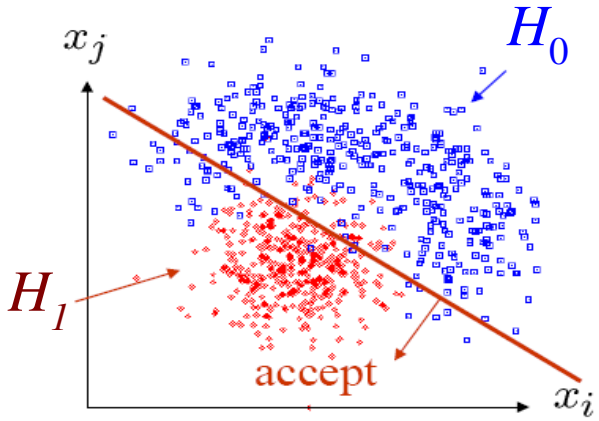
Finding an optimal decision boundary

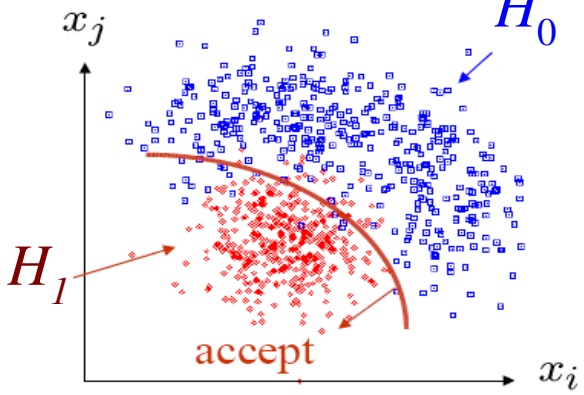
In particle physics usually start by making simple "cuts":

$$x_i < c_i$$
$$x_j < c_j$$



Maybe later try some other type of decision boundary:





Test statistics

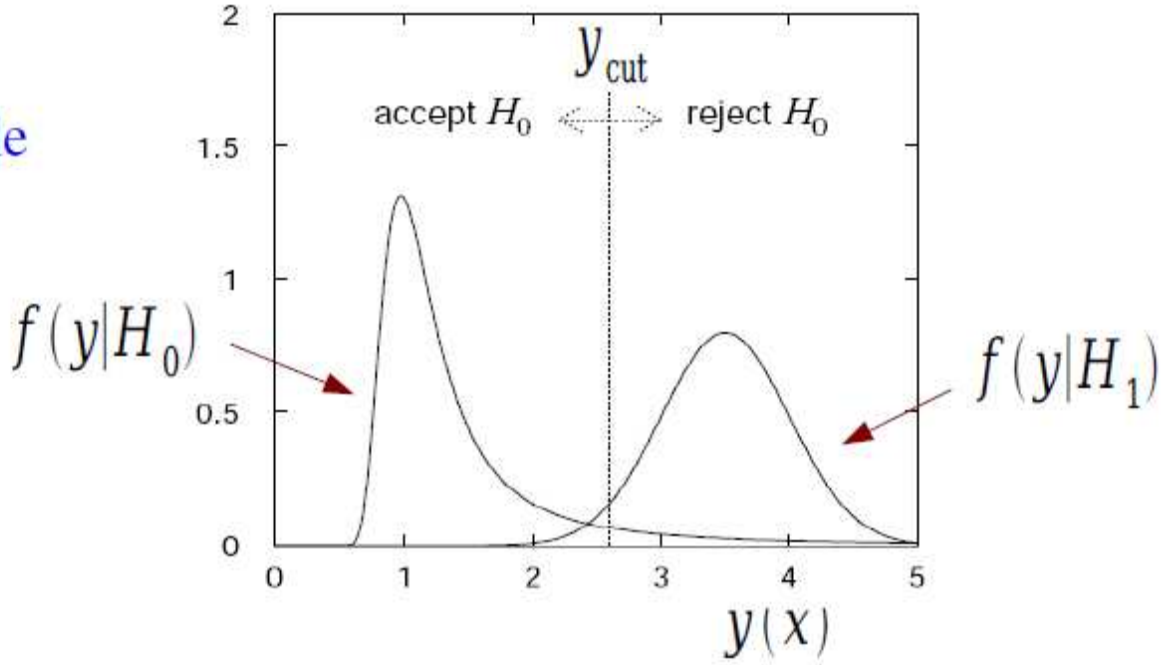
The decision boundary is a surface in the n-dimensional space of input variables, e.g., $y(\vec{x}) = \text{const}$.

We can treat the $y(x)$ as a scalar test statistic or discriminating function, and try to define this function so that its distribution has the maximum possible separation between the event types:

The decision boundary is now effectively a single cut on $y(\boldsymbol{x})$, dividing $\boldsymbol{x}$-space into two regions:

R_0 (accept H_0)

R_1 (reject H_0)



The optimal decision boundary

Try to best approximate optimal decision boundary based on likelihood ratio:

$$y(\mathbf{x}) = \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_1)} = \text{const.}$$

or equivalently think of the likelihood ratio as the optimal statistic for a test of H_0 vs H_1.

In general we don't have the pdfs $p(\boldsymbol{x}|H_0)$, $p(\boldsymbol{x}|H_1)$,...
Rather, we have Monte Carlo models for each process.

Usually training data from the MC models is cheap.

But the models contain many approximations:
predictions for observables obtained using perturbation theory (truncated at some order); phenomenological modeling of non-perturbative effects; imperfect detector description,...

Two distinct event selection problems

In some cases, the event types in question are both known to exist.

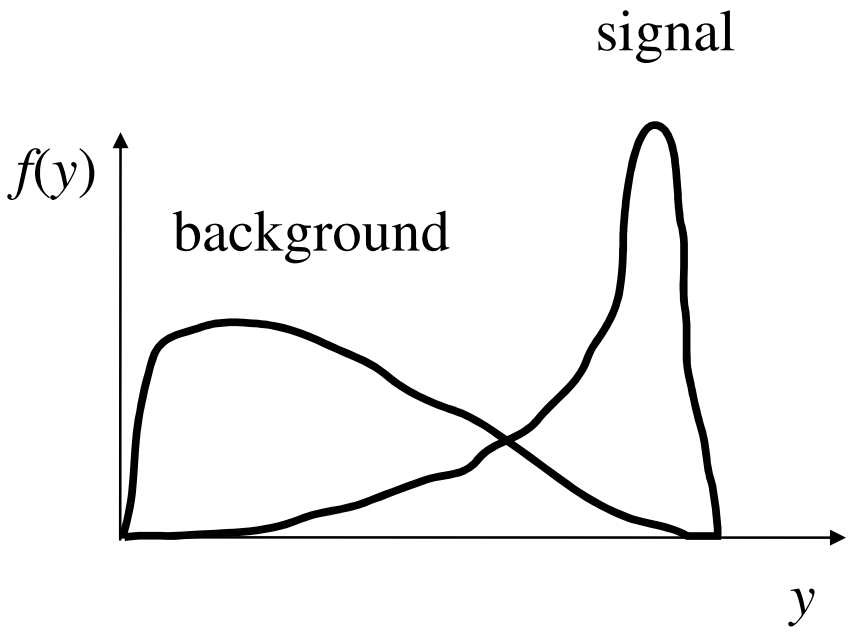
Example: separation of different particle types (electron vs muon)
Use the selected sample for further study.

In other cases, the null hypothesis H_0 means "Standard Model" events, and the alternative H_1 means "events of a type whose existence is not yet established" (to do so is the goal of the analysis).

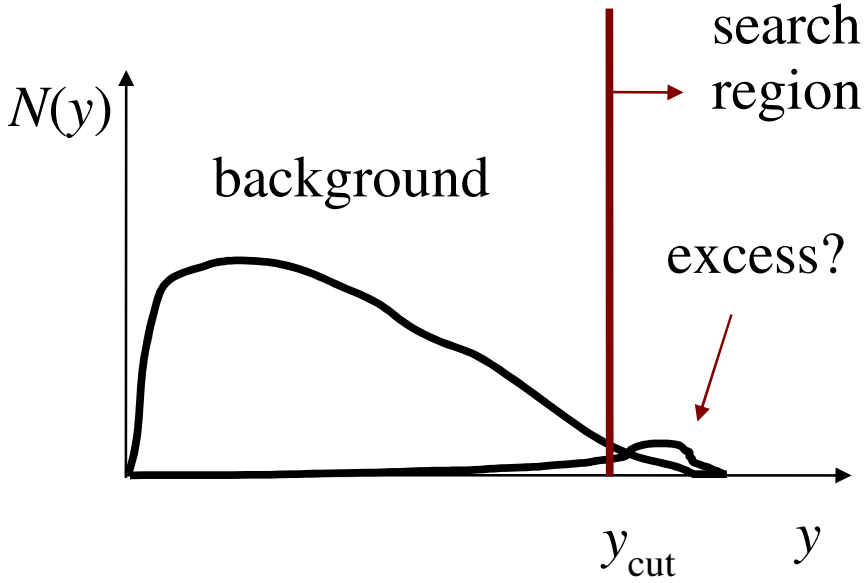
Many subtle issues here, mainly related to the heavy burden of proof required to establish presence of a new phenomenon.

Typically require p-value of background-only hypothesis below ~ 10^{-7} (a 5 sigma effect) to claim discovery of "New Physics".

Using classifier output for discovery



Normalized to unity



Normalized to expected number of events

Discovery = number of events found in search region incompatible with background-only hypothesis.

p-value of background-only hypothesis can depend crucially distribution $f(y|b)$ in the "search region".

Some "standard" multivariate methods

Place cuts on individual variables

Simple, intuitive, in general not optimal

Linear discriminant (e.g. Fisher)

Simple, optimal if the event types are Gaussian distributed with equal covariance, otherwise not optimal.

Probability Density Estimation based methods

Try to estimate $p(\boldsymbol{x}|\mathrm{s})$, $p(\boldsymbol{x}|\mathrm{b})$ then use $y(\vec{x}) = \hat{p}(\boldsymbol{x}|\mathrm{s}) / \hat{p}(\boldsymbol{x}|\mathrm{b})$.

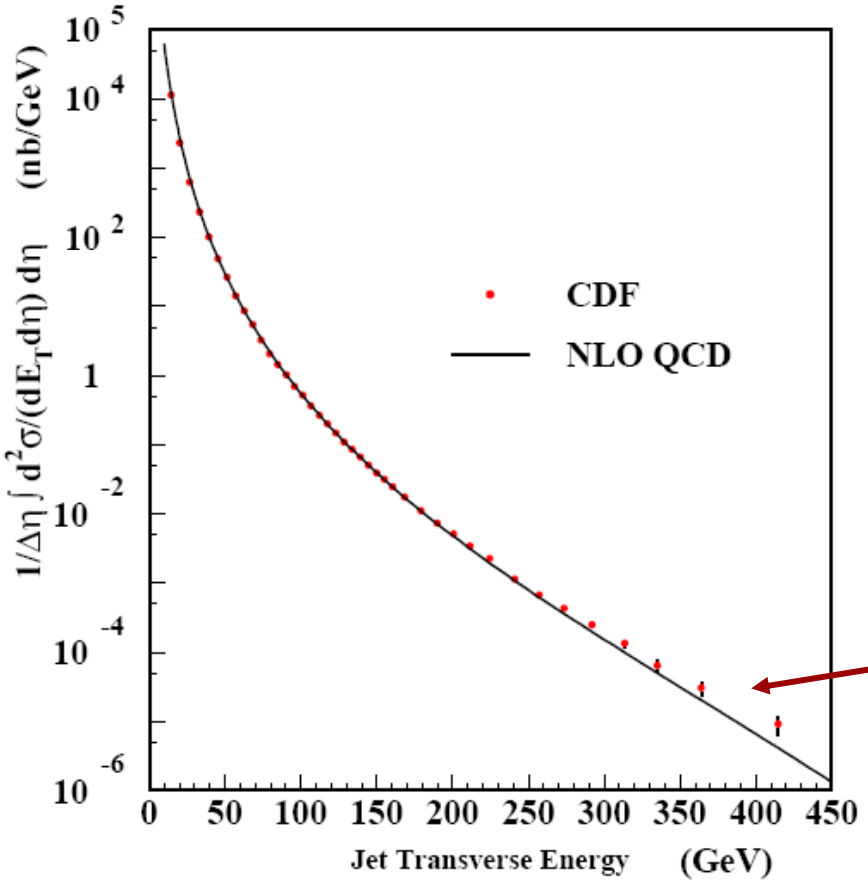
In principle best, difficult to estimate $p(\boldsymbol{x})$ for high dimension.

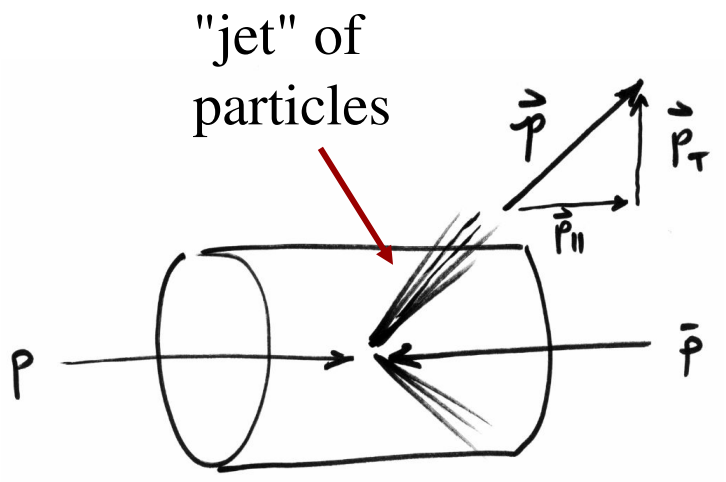
Neural networks

Can produce arbitrary decision boundary (in principle optimal), but can be difficult to train, result non-intuitive.

Example of a "cut-based" study

In the 1990s, the CDF experiment at Fermilab (Chicago) measured the number of hadron jets produced in proton-antiproton collisions as a function of their momentum perpendicular to the beam direction:

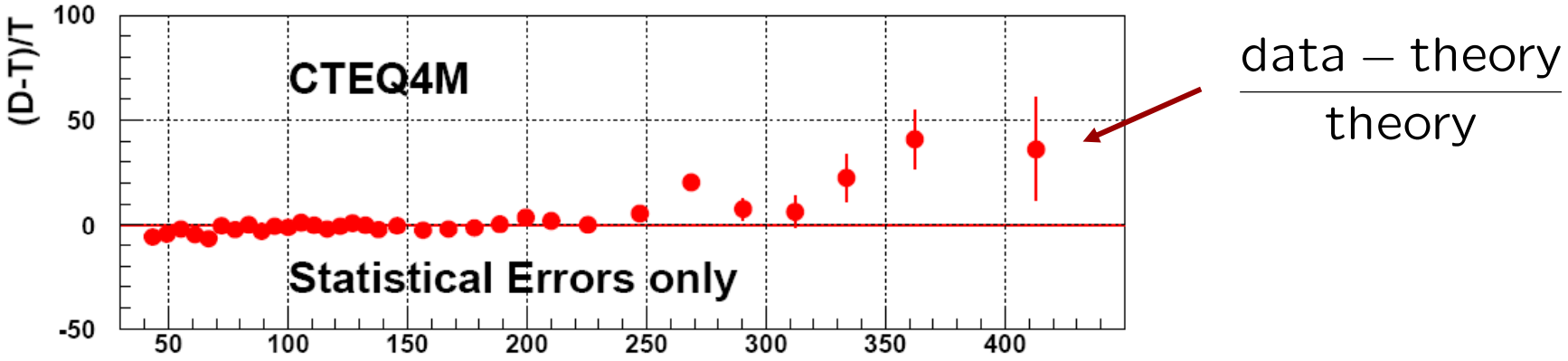




Prediction low relative to data for very high transverse momentum.

High p_T jets = quark substructure?

Although the data agree remarkably well with the Standard Model (QCD) prediction overall, the excess at high p_T appears significant:



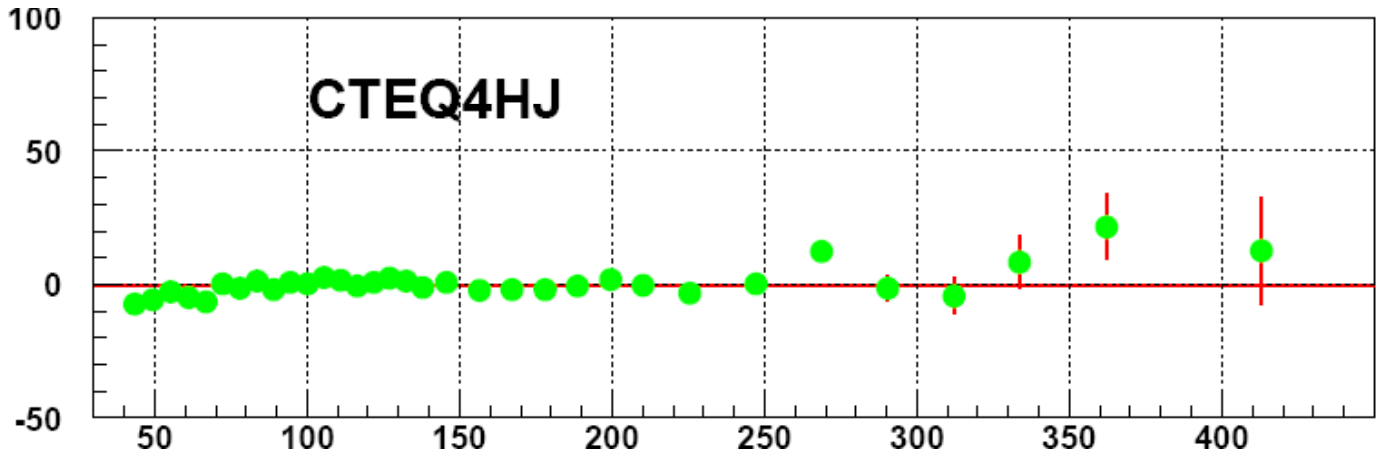
The fact that the variable is "understandable" leads directly to a plausible explanation for the discrepancy, namely, that quarks could possess an internal substructure.

Would not have been the case if the variable plotted was a complicated combination of many inputs.

High p_T jets from parton model uncertainty

Furthermore the physical understanding of the variable led one to a more plausible explanation, namely, an uncertain modeling of the quark (and gluon) momentum distributions inside the proton.

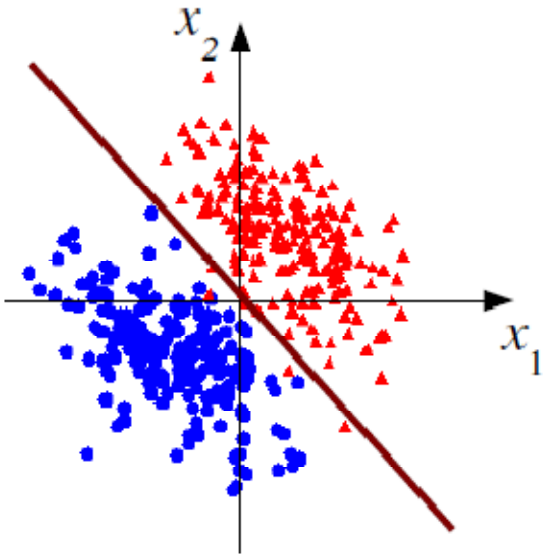
When model adjusted, discrepancy largely disappears:

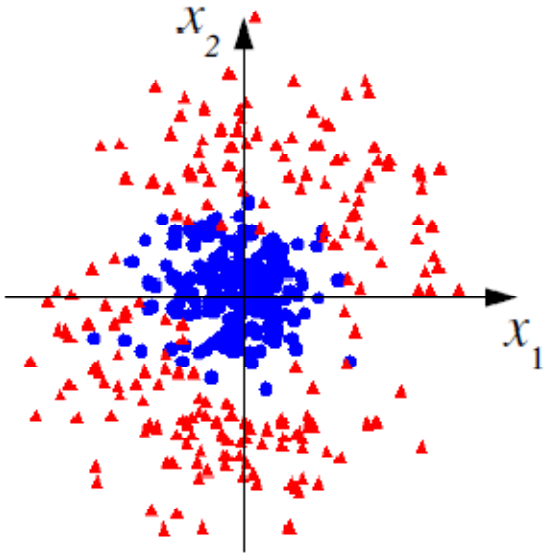


Can be regarded as a "success" of the cut-based approach. Physical understanding of output variable led to solution of apparent discrepancy.

Linear decision boundaries

A linear decision boundary is only optimal when both classes follow multivariate Gaussians with equal covariances and different means.





For some other cases a linear boundary is almost useless.

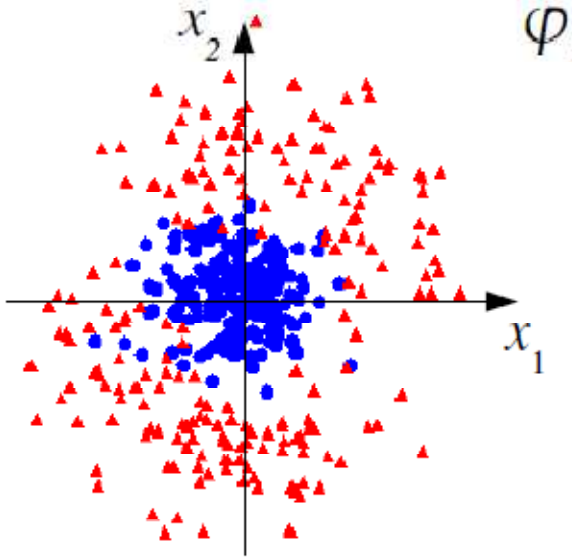
Nonlinear transformation of inputs

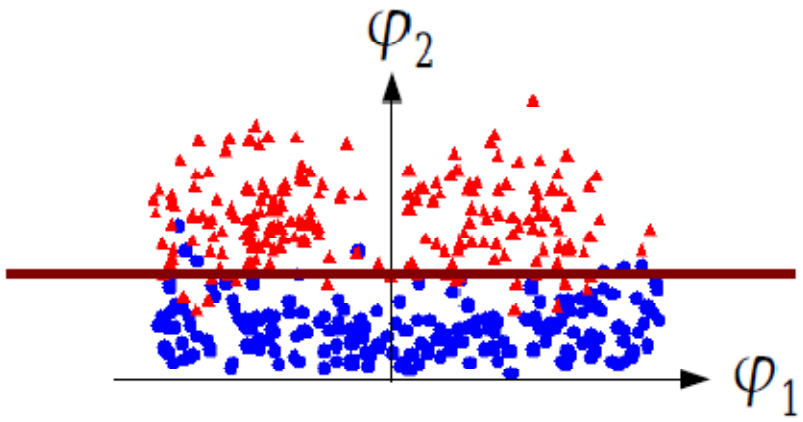
We can try to find a transformation, $x_1, \ldots, x_n \to \varphi_1(\vec{x}), \ldots, \varphi_m(\vec{x})$ so that the transformed "feature space" variables can be separated better by a linear boundary:

$$\varphi_1 = \tan^{-1}(x_2/x_1)$$

$$\varphi_2 = \sqrt{x_1^2 + x_2^2}$$

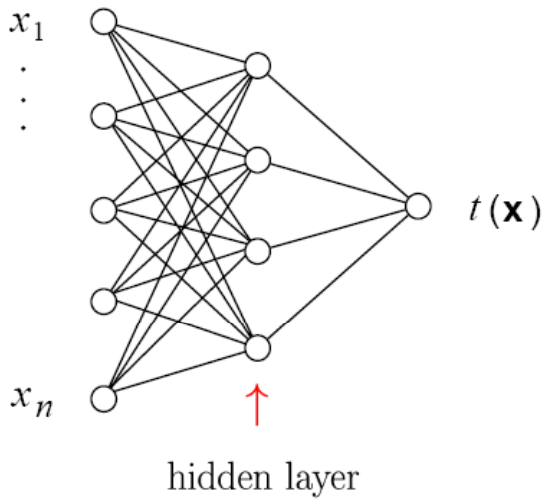
Here, guess fixed basis functions (no free parameters)





Neural networks in particle physics

For many years, the only "advanced" classifier used in particle physics.

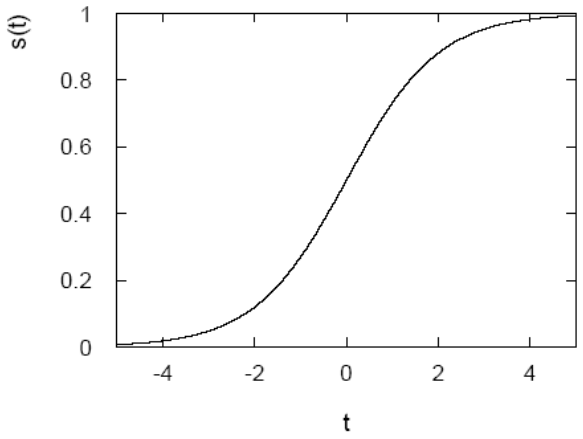


$$h_i(\vec{x}) = s\left(w_{i0} + \sum_{j=1}^{n} w_{ij}x_j\right) ,$$

$$t(\vec{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i h_i(\vec{x})\right) .$$

Usually use single hidden layer,
logistic sigmoid activation function:

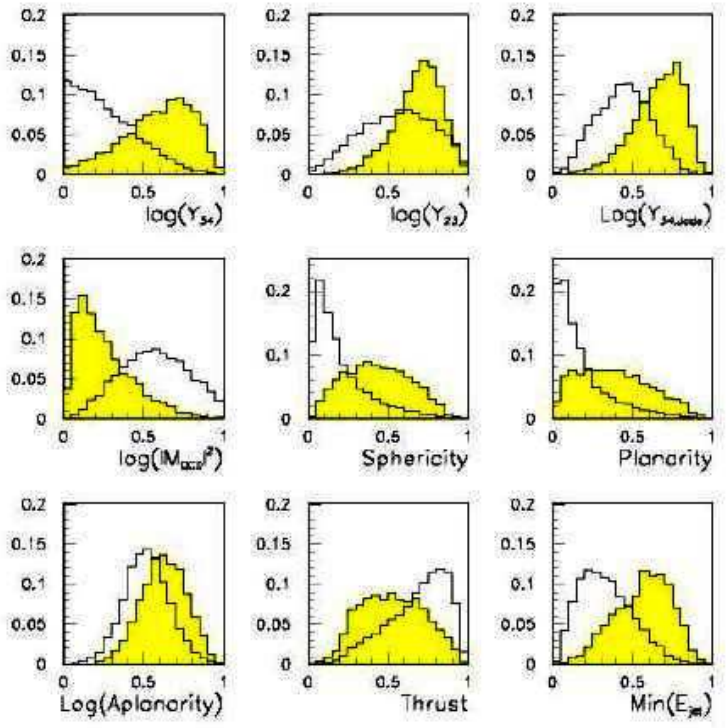
$$s(u) \equiv (1 - e^{-u})^{-1} .$$



Neural network example from LEP II

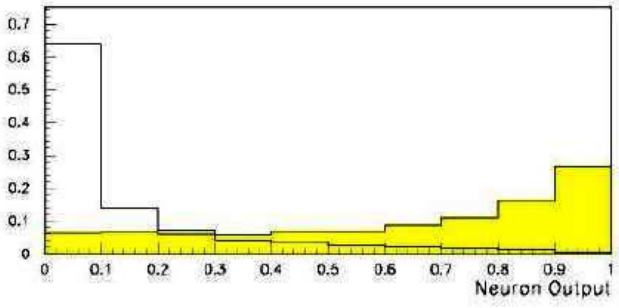
Signal: $e^+e^- \to W^+W^-$ (often 4 well separated hadron jets)

Background: $e^+e^- \to qqgg$ (4 less well separated hadron jets)



← input variables based on jet structure, event shape, ...
none by itself gives much separation.

Neural network output:



(Garrido, Juste and Martinez, ALEPH 96-144)

Some issues with neural networks

In the example with WW events, goal was to select these events so as to study properties of the W boson.

> Needed to avoid using input variables correlated to the properties we eventually wanted to study (not trivial).

In principle a single hidden layer with an sufficiently large number of nodes can approximate arbitrarily well the optimal test variable (likelihood ratio).

> Usually start with relatively small number of nodes and increase until misclassification rate on validation data sample ceases to decrease.

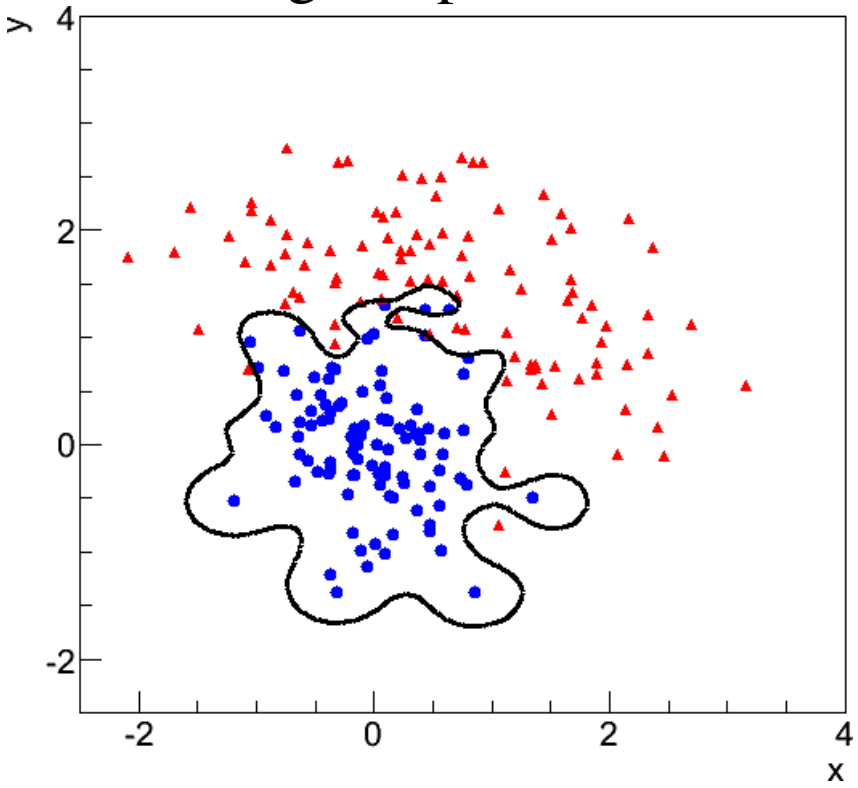
Often MC training data is cheap -- problems with getting stuck in local minima, overtraining, etc., less important than concerns of systematic differences between the training data and Nature, and concerns about the ease of interpretation of the output.

Overtraining

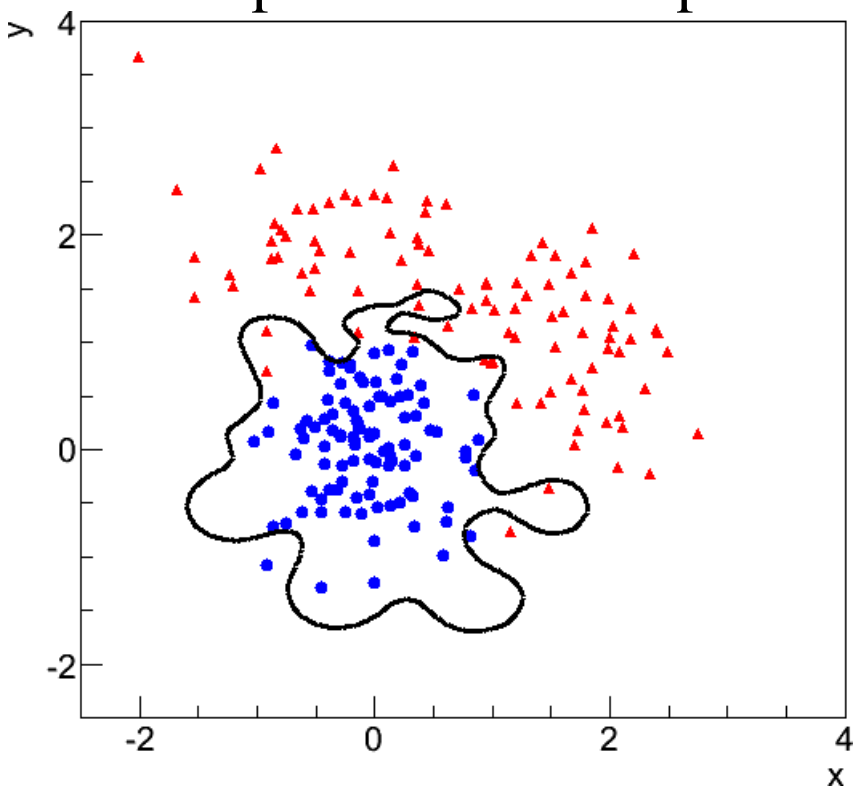
If decision boundary is too flexible it will conform too closely to the training points → overtraining.

Monitor by applying classifier to independent test sample.

training sample



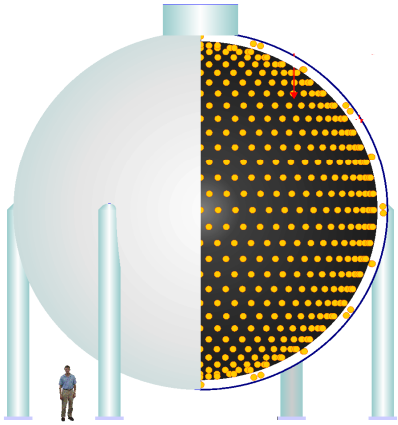
independent test sample



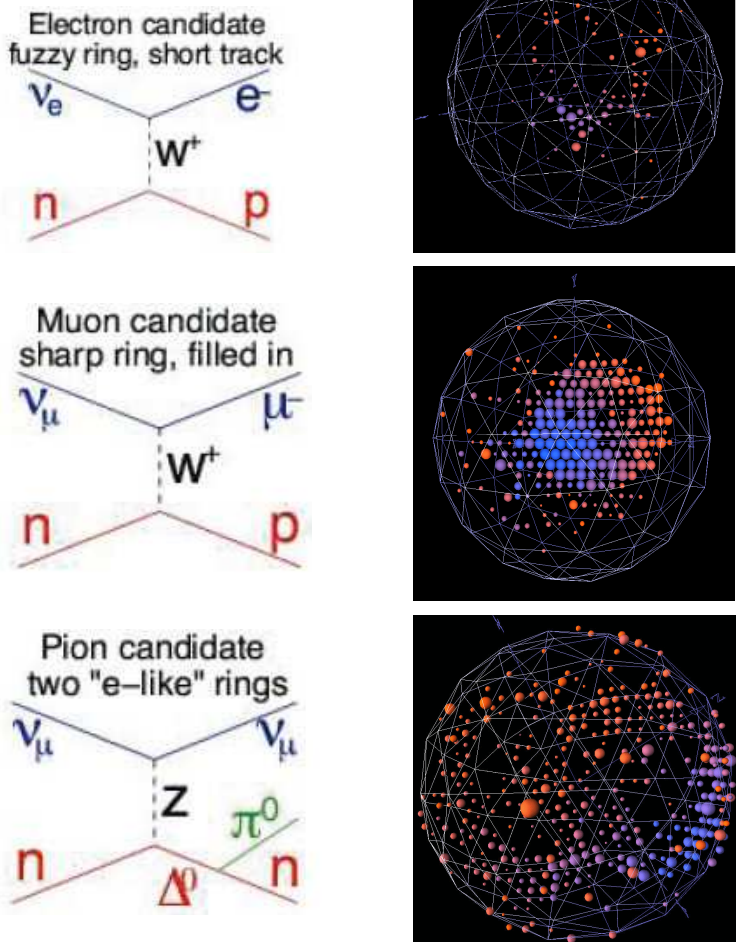
Particle i.d. in MiniBooNE

Detector is a 12-m diameter tank of mineral oil exposed to a beam of neutrinos and viewed by 1520 photomultiplier tubes:

MiniBooNE Detector



Search for ν_μ to ν_e oscillations required particle i.d. using information from the PMTs.



H.J. Yang, MiniBooNE PID, DNP06

Decision trees

Out of all the input variables, find the one for which with a single cut gives best improvement in signal purity:

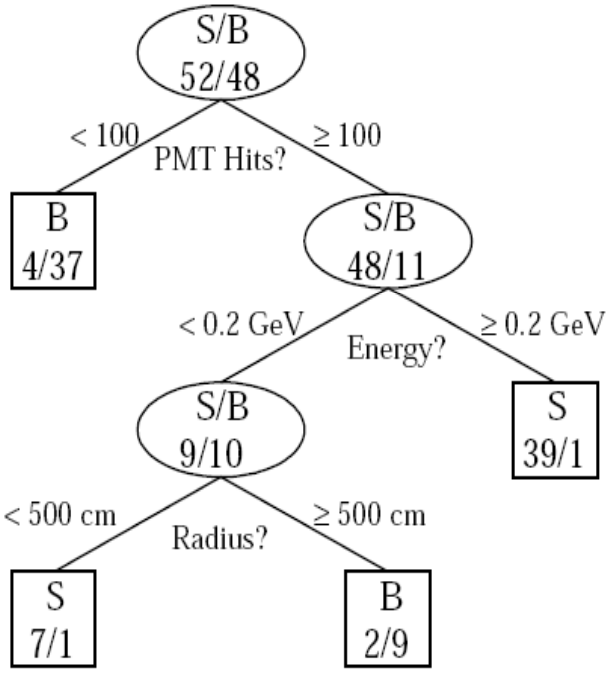
$$P = \frac{\sum_{\text{signal}} w_i}{\sum_{\text{signal}} w_i + \sum_{\text{background}} w_i}$$

where w_i. is the weight of the ith event.

Resulting nodes classified as either signal/background.

Iterate until stop criterion reached based on e.g. purity or minimum number of events in a node.

The set of cuts defines the decision boundary.



Example by MiniBooNE experiment, B. Roe et al., NIM 543 (2005) 577

Decision trees (2)

The terminal nodes (leaves) are classified as signal or background depending on majority vote (or e.g. signal fraction greater than a specified threshold).

This classifies every point in input-variable space as either signal or background, a decision tree classifier, with the discriminant function

$$f(\boldsymbol{x}) = 1 \text{ if } \boldsymbol{x} \in \text{signal region}, -1 \text{ otherwise}$$

Decision trees tend to be very sensitive to statistical fluctuations in the training sample.

Methods such as boosting can be used to stabilize the tree.

Boosting

Boosting is a general method of creating a set of classifiers which can be combined to achieve a new classifier that is more stable and has a smaller error than any individual one.

Often applied to decision trees but, can be applied to any classifier.

Suppose we have a training sample T consisting of N events with

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ event data vectors (each $\boldsymbol{x}$ multivariate)

$y_1, \ldots, y_N$ true class labels, +1 for signal, −1 for background

$w_1, \ldots, w_N$ event weights

Now define a rule to create from this an ensemble of training samples $T_1, T_2, \ldots$, derive a classifier from each and average them.

AdaBoost

A successful boosting algorithm is AdaBoost (Freund & Schapire, 1997).

First initialize the training sample T_1 using the original

$\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N$ event data vectors

$y_1,\ldots,y_N$ true class labels (+1 or -1)

$w_1^{(1)},\ldots,w_N^{(1)}$ event weights

with the weights equal and normalized such that $\sum_{i=1}^{N} w_i^{(1)} = 1$.

Train the classifier $f_1(\boldsymbol{x})$ (e.g. a decision tree) using the weights $\boldsymbol{w}^{(1)}$ so as to minimize the classification error rate,

$$\varepsilon_1 = \sum_{i=1}^{N} w_i^{(1)} I(y_i f_1(\boldsymbol{x}_i) \leqslant 0),$$

where $I(X) = 1$ if X is true and is zero otherwise.

Updating the event weights (AdaBoost)

Assign a score to the kth classifier based on its error rate:

$$\alpha_k = \ln \frac{1-\varepsilon_k}{\varepsilon_k}$$

Define the training sample for step k+1 from that of k by updating the event weights according to

$$w_i^{(k+1)} = w_i^{(k)} \frac{e^{-\alpha_k f_k(\mathbf{x}_i) y_i / 2}}{Z_k}$$

i = event index

k = training sample index

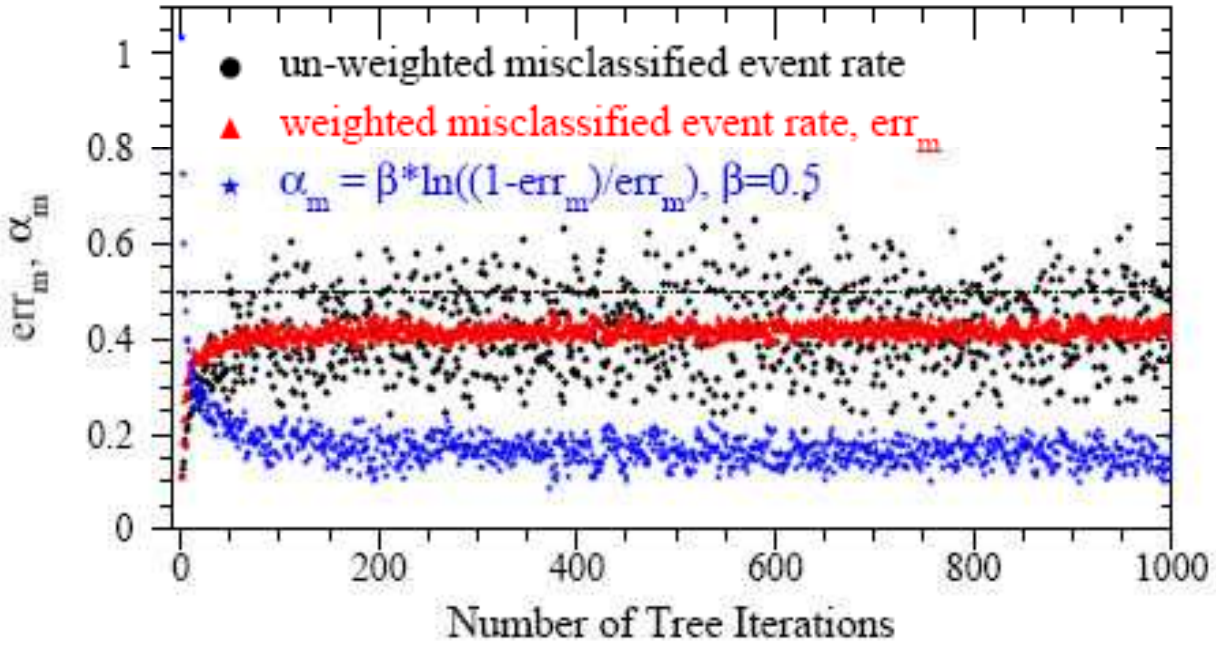
Normalize so that $\sum_i w_i^{(k+1)} = 1$

Iterate K times, final classifier is $y(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k f_k(\mathbf{x}, T_k)$

BDT example from MiniBooNE

~200 input variables for each event (ν interaction producing e, μ or π).

Each individual tree is relatively weak, with a misclassification error rate ~ 0.4 – 0.45

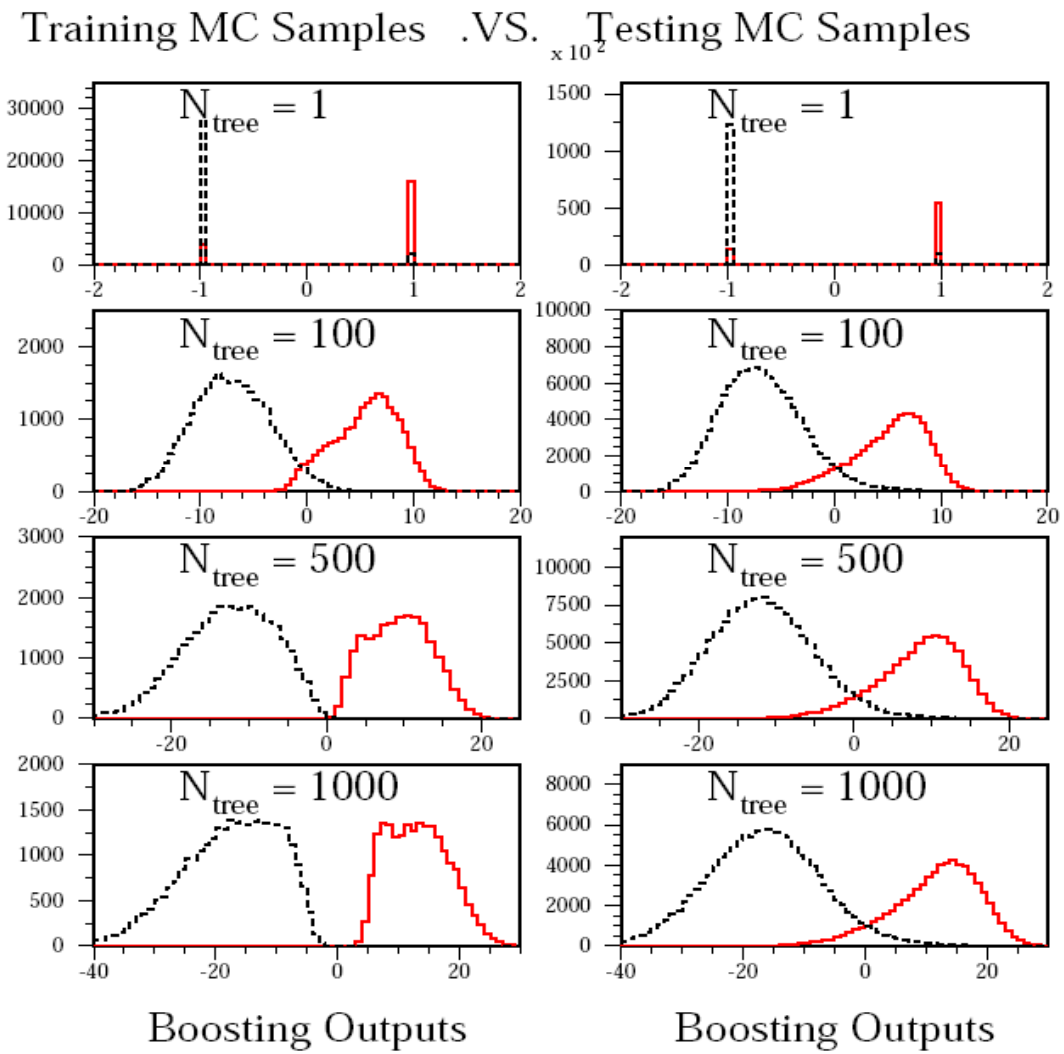


B. Roe et al., NIM 543 (2005) 577

Monitoring overtraining

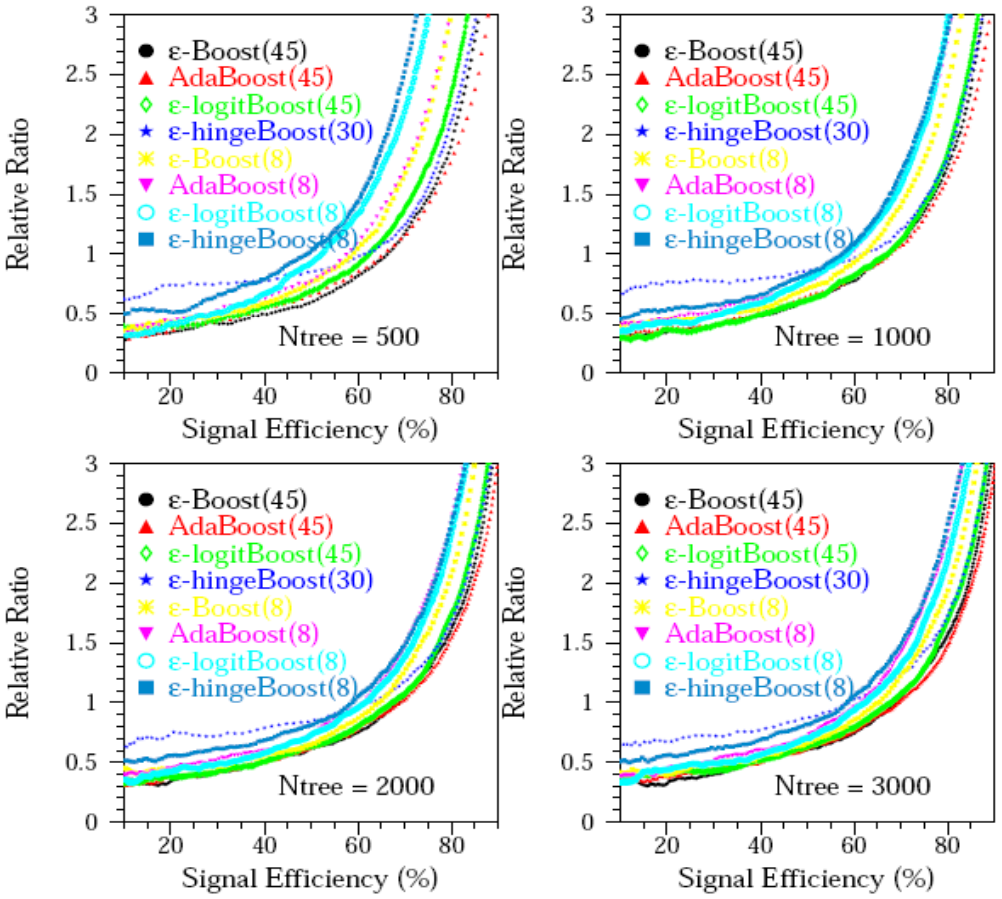
From MiniBooNE example:

Performance stable after a few hundred trees.



Comparison of boosting algorithms

A number of boosting algorithms on the market; differ in the update rule for the weights.



Boosted decision tree summary

Advantage of boosted decision tree is it can handle a large number of inputs. Those that provide little/no separation are rarely used as tree splitters are effectively ignored.

Easy to deal with inputs of mixed types (real, integer, categorical...).

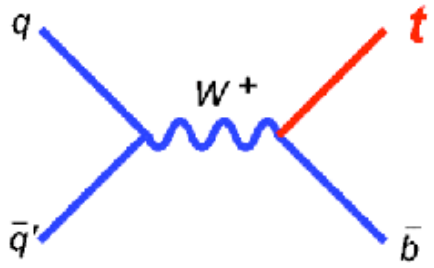
If a tree has only a few leaves it is easy to visualize (but rarely use only a single tree).

There are a number of boosting algorithms, which differ primarily in the rule for updating the weights (ε-Boost, LogitBoost,...)

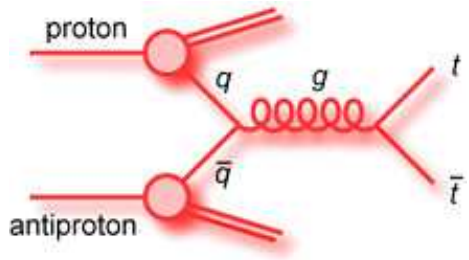
Other ways of combining weaker classifiers: Bagging (Boostrap-Aggregating), generates the ensemble of classifiers by random sampling with replacement from the full training sample.

Single top quark production (CDF/D0)

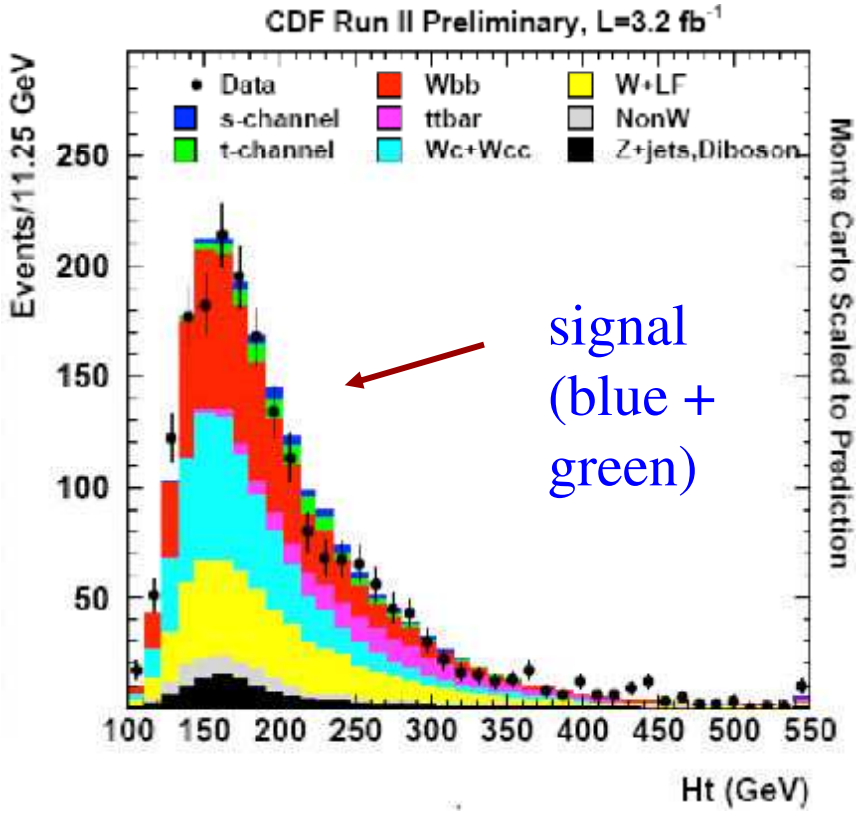
Top quark discovered in pairs, but
SM predicts single top production.



Pair-produced tops are now
a background process.

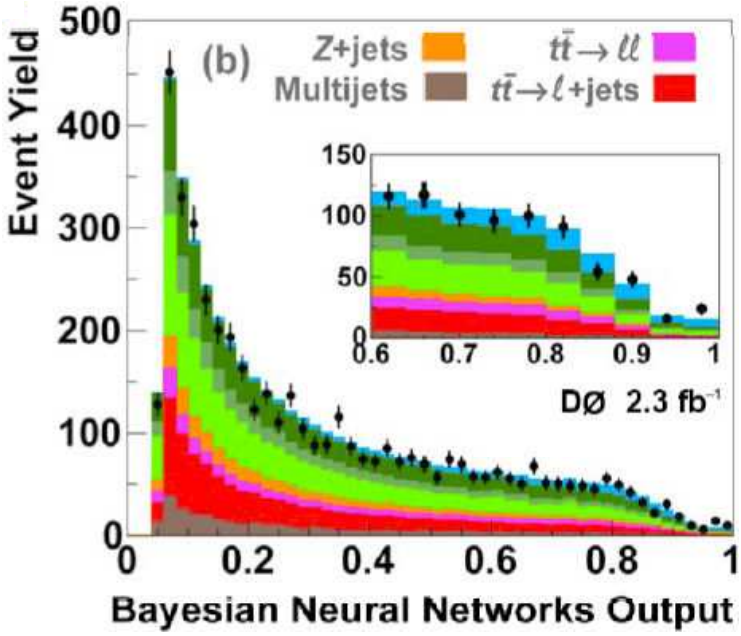


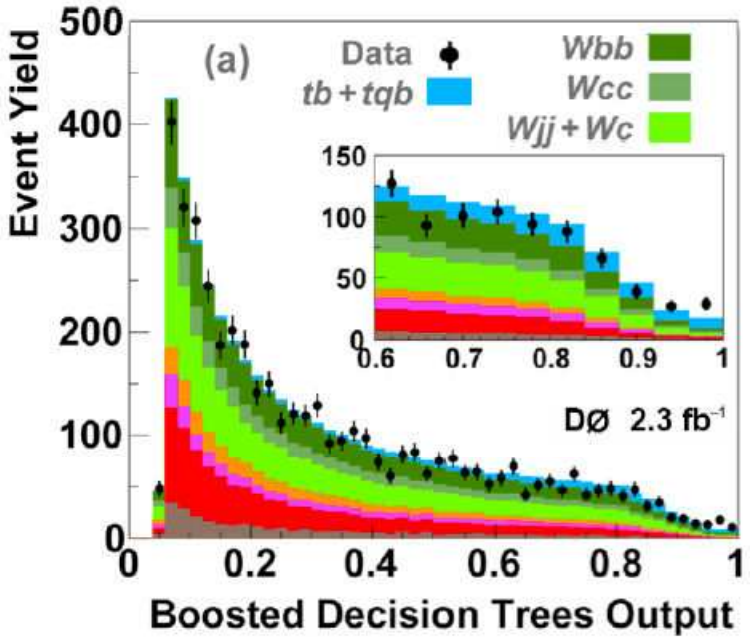
Use many inputs based on
jet properties, particle i.d., ...



signal
(blue +
green)

Different classifiers for single top





Also Naive Bayes and various approximations to likelihood ratio,....

Final combined result is statistically significant (>5σ level) but not easy to understand classifier outputs.

Support Vector Machines

Map input variables into high dimensional feature space: $\boldsymbol{x} \rightarrow \boldsymbol{\phi}$

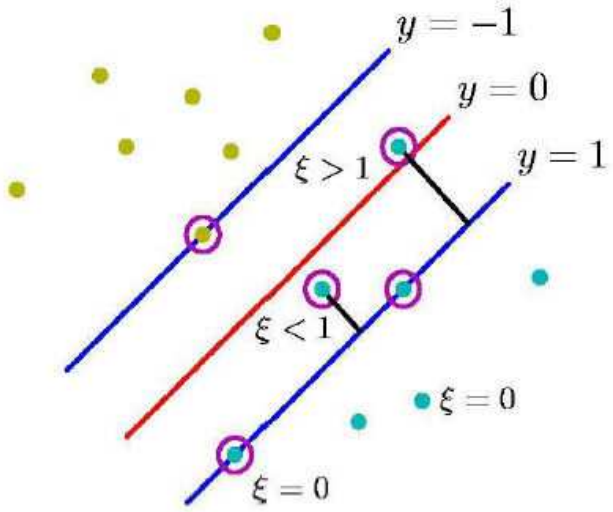
Maximize distance between separating hyperplanes (margin) subject to constraints allowing for some misclassification.

Final classifier only depends on scalar products of $\boldsymbol{\phi}(\boldsymbol{x})$:

$$y(\boldsymbol{x}) = \text{sign}\left(\sum_i \alpha_i y_i \vec{\varphi}(\boldsymbol{x}) \cdot \vec{\varphi}(\boldsymbol{x}_i) + b \right)$$

So only need kernel

$$K(\boldsymbol{x}, \boldsymbol{x}') = \vec{\varphi}(\boldsymbol{x}) \cdot \vec{\varphi}(\boldsymbol{x}')$$



Support Vector Machines

Support Vector Machines (SVMs) are an example of a kernel-based classifier, which exploits a nonlinear mapping of the input variables onto a higher dimensional feature space.

The SVM finds a linear decision boundary in the higher dimensional space.

But thanks to the "kernel trick" one does not every have to write down explicitly the feature space transformation.

Some references for kernel methods and SVMs:

The books mentioned in `www.pp.rhul.ac.uk/~cowan/mainz_lectures.html`

C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, `research.microsoft.com/~cburges/papers/SVMTutorial.pdf`

N. Cristianini and J.Shawe-Taylor. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.

The TMVA manual (!)

Linear SVMs

Consider a training data set consisting of

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ event data vectors

$y_1, \ldots, y_N$ true class labels (+1 or −1)

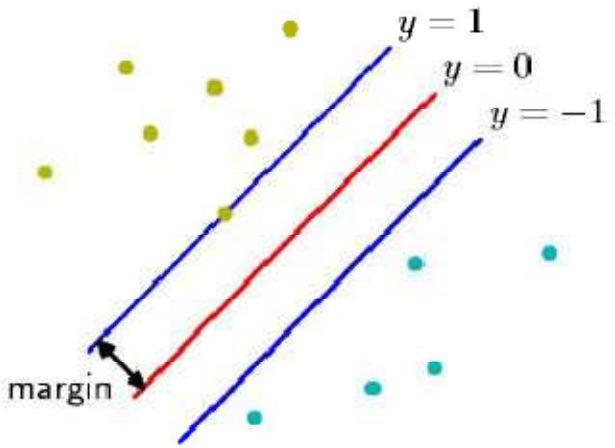
Suppose the classes can be separated by a hyperplane defined by a normal vector $\boldsymbol{w}$ and scalar offset b (the "bias"). We have

$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \geqslant +1$ for all $y_i = +1$

$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \leqslant -1$ for all $y_i = -1$

or equivalently

$y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) - 1 \geqslant 0$ for all i

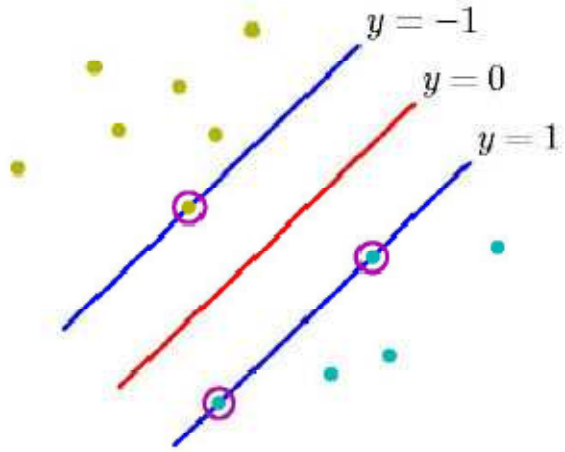


Bishop Ch. 7

Margin and support vectors

The distance between the hyperplanes defined by $y(\boldsymbol{x}) = \boldsymbol{x}\cdot\boldsymbol{w} + b = +1$ and $y(\boldsymbol{x}) = -1$ is called the margin, which is:

$$\text{margin} = \frac{2}{\|\boldsymbol{w}\|}$$



If the training data are perfectly separated then this means there are no points inside the margin.

Suppose there are points on the margin (this is equivalent to defining the scale of $\boldsymbol{w}$). These points are called support vectors.

Linear SVM classifier

We can define the classifier using

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{x} \cdot \boldsymbol{w} + b)$$

which is +1 for points on one side of the hyperplane and −1 on the other.

The best classifier should have a large margin, so to maximize

$$\text{margin} = \frac{2}{\|\boldsymbol{w}\|}$$

we can minimize $\|\boldsymbol{w}\|^2$ subject to the constraints

$$y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) - 1 \geqslant 0 \quad \text{for all } i$$

Lagrangian formulation

This constrained minimization problem can be reformulated using a Lagrangian

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(\boldsymbol{x_i}\cdot\boldsymbol{w} + b) - 1)$$

positive Lagrange multipliers α_i

We need to minimize L with respect to $\boldsymbol{w}$ and b and maximize with respect to α_i.

There is an α_i for every training point. Those that lie on the margin (the support vectors) have $\alpha_i > 0$, all others have $\alpha_i = 0$. The solution can be written

$$\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x_i}$$

(sum only contains support vectors)

Dual formulation

The classifier function is thus

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{x}\cdot\boldsymbol{w} + b) = \text{sign}\left(\sum_i \alpha_i y_i \boldsymbol{x}\cdot\boldsymbol{x}_i + b\right)$$

It can be shown that one finds the same solution a by minimizing the dual Lagrangian

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$$

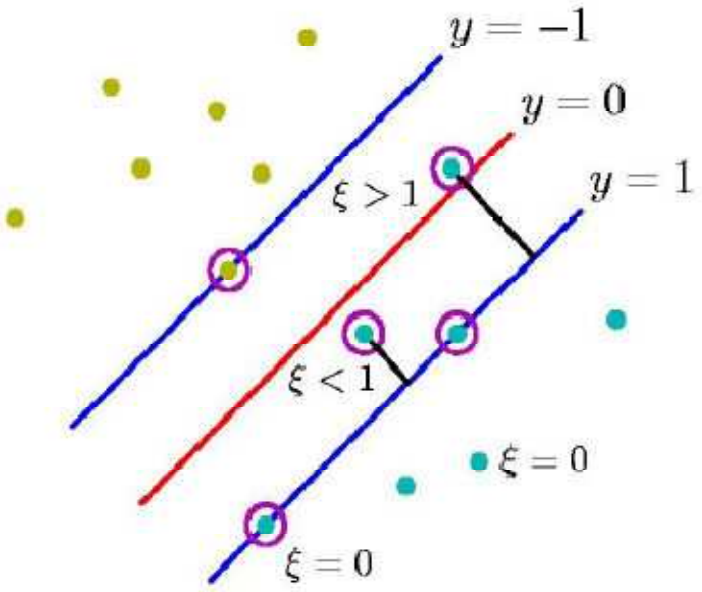
So this means that both the classifier function and the Lagrangian only involve dot products of vectors in the input variable space.

Nonseparable data

If the training data points cannot be separated by a hyperplane, one can redefine the constraints by adding slack variables ξ_i:

$$y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) + \xi_i - 1 \geqslant 0 \text{ with } \xi_i \geqslant 0 \text{ for all } i$$

Thus the training point $\boldsymbol{x}_i$ is allowed to be up to a distance ξ_i on the wrong side of the margin, and $\xi_i = 0$ at or on the right side.



For an error to occur we have $\xi_i > 1$, so

$$\sum_i \xi_i$$

is an upper bound on the number of training errors.

Cost function for nonseparable case

To limit the magnitudes of the ξ_i we can define the error function that we minimize to determine $\boldsymbol{w}$ to be

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_i \xi_i\right)^k$$

where C is a cost parameter we must choose that limits the amount of misclassification. It turns out that for k=1 or 2 this is a quadratic programming problem and furthermore for k=1 it corresponds to minimizing the same dual Lagrangian

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$$

where the constraints on the α_i become $0 \leqslant \alpha_i \leqslant C$.

Nonlinear SVM

So far we have only reformulated a way to determine a linear classifier, which we know is useful only in limited circumstances.

But the important extension to nonlinear classifiers comes from first transforming the input variables to feature space:

$$\vec{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \ldots, \varphi_m(\boldsymbol{x}))$$

These will behave just as our new "input variables". Everything about the mathematical formulation of the SVM will look the same as before except with $\boldsymbol{\phi}(\boldsymbol{x})$ appearing in the place of $\boldsymbol{x}$.

Only dot products

Recall the SVM problem was formulated entirely in terms of dot products of the input variables, e.g., the classifier is

$$f(\boldsymbol{x}) = \mathrm{sign}\left(\sum_i \alpha_i y_i \boldsymbol{x} \cdot \boldsymbol{x}_i + b \right)$$

so in the feature space this becomes

$$f(\boldsymbol{x}) = \mathrm{sign}\left(\sum_i \alpha_i y_i \vec{\varphi}(\boldsymbol{x}) \cdot \vec{\varphi}(\boldsymbol{x}_i) + b \right)$$

The Kernel trick

How do the dot products help? It turns on that a broad class of kernel functions can be written in the form:

$$K(\boldsymbol{x},\boldsymbol{x}')=\vec{\varphi}(\boldsymbol{x})\cdot\vec{\varphi}(\boldsymbol{x}')$$

Functions having this property must satisfy Mercer's condition

$$\int K(\boldsymbol{x},\boldsymbol{x}')g(\boldsymbol{x})g(\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}'\geqslant 0$$

for any function g where $\int g^2(\boldsymbol{x})d\boldsymbol{x}$ is finite.

So we don't even need to find explicitly the feature space transformation $\boldsymbol{\phi}(\boldsymbol{x})$, we only need a kernel.

Finding kernels

There are a number of techniques for finding kernels, e.g., constructing new ones from known ones according to certain rules (cf. Bishop Ch 6).

Frequently used kernels to construct classifiers are e.g.

$$K(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{x}\cdot\boldsymbol{x}' + \theta)^p \qquad \text{polynomial}$$

$$K(\boldsymbol{x},\boldsymbol{x}') = \exp\left(\frac{-\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right) \qquad \text{Gaussian}$$

$$K(\boldsymbol{x},\boldsymbol{x}') = \tanh\left(\kappa(\boldsymbol{x}\cdot\boldsymbol{x}') + \theta\right) \qquad \text{sigmoidal}$$

Using an SVM

To use an SVM the user must as a minimum choose

a kernel function (e.g. Gaussian)
any free parameters in the kernel (e.g. the σ of the Gaussian)
a cost parameter C (plays role of regularization parameter)

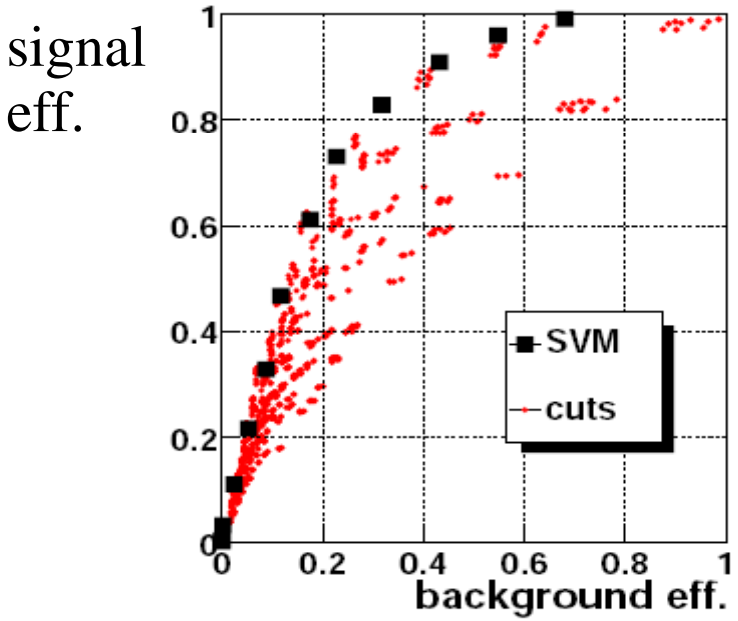
The training is relatively straightforward because, in contrast to neural networks, the function to be minimized has a single global minimum.

Furthermore evaluating the classifier only requires that one retain and sum over the support vectors, a relatively small number of points.

The advantages/disadvantages and rationale behind the choices above is not always clear to the particle physicist -- help needed here.

SVM in particle physics

SVMs are very popular in the Machine Learning community but have yet to find wide application in HEP. Here is an early example from a CDF top quark anlaysis (A. Vaiciulis, contribution to PHYSTAT02).



Summary on multivariate methods

Particle physics has used several multivariate methods for many years:

- linear (Fisher) discriminant
- neural networks
- naive Bayes

and has in the last several years started to use a few more

- *k*-nearest neighbour
- boosted decision trees
- support vector machines

The emphasis is often on controlling systematic uncertainties between the modeled training data and Nature to avoid false discovery.

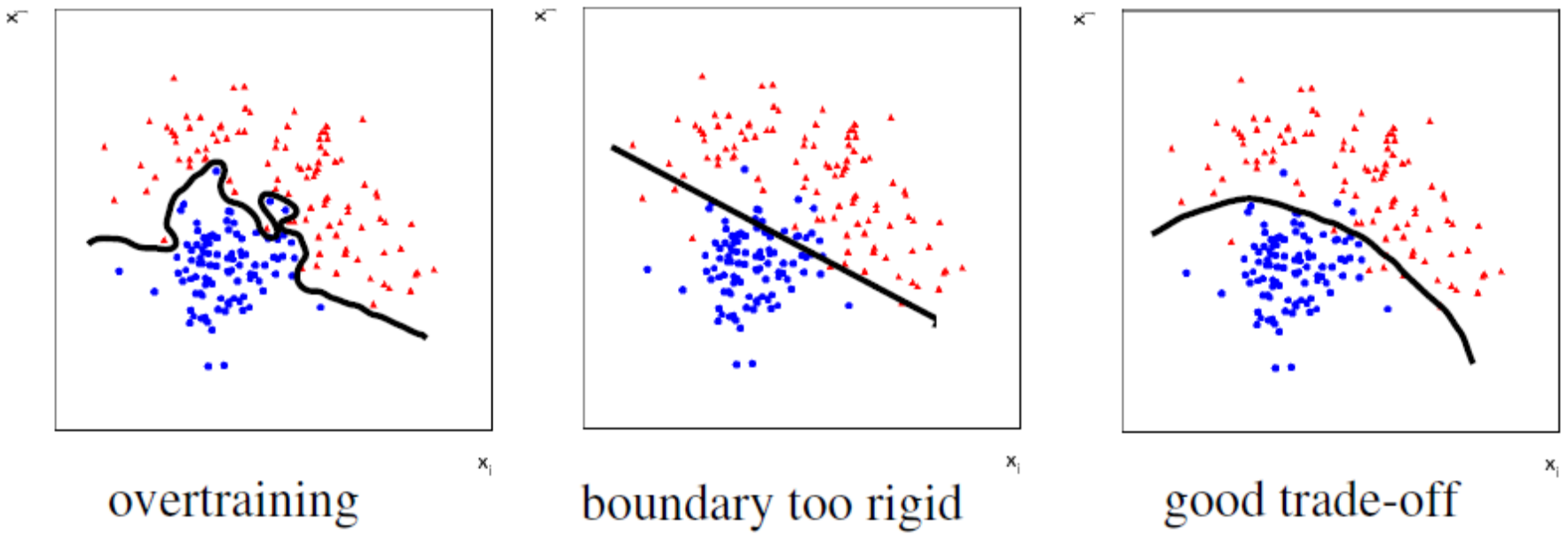
Although many classifier outputs are "black boxes", a discovery at 5σ significance with a sophisticated (opaque) method will win the competition if backed up by, say, 4σ evidence from a cut-based method.

Extra slides

Decision boundary flexibility

The decision boundary will be defined by some free parameters that we adjust using training data (of known type) to achieve the best separation between the event types.

Goal is to determine the boundary using a finite amount of training data so as to best separate between the event types for an unseen data sample.



overtraining boundary too rigid good trade-off