Multidimensional Data

- Every event or Pattern or object of interest can be represented as a point in a multi-dimensional "feature space" spanned by the "feature variables" that characterize the event, pattern or object.

$$\vec{X} \equiv (x_1, x_2, x_3, \ldots)$$

- Usually correlations exist between variables. These correlations are useful information about the data, and need to be exploited.

- So, an optimal analysis is multivariate

Multivariate Discriminant Analysis

Bayes Decision Rule:

Assign a vector $\vec{x}$ (an event or data point or pattern) to class C_k if $P(C_k|\vec{x})$ is the highest of all classes.

$P(C_k|\vec{x}) \equiv$ Class posterior probability, i.e., the probability that $\vec{x}$ belongs to class C_k.

If there are 2 classes then x belongs to C_1 if $P(C_1|x) > P(C_2|x)$

It can be shown that

$$R(x) = \frac{P(C_1|x)}{P(C_2|x)} = \text{Constant defines an optimal boundary}$$

Bayes Discriminant

$$= \frac{P(x|C_1)\,P(C_1)}{P(x|C_2)\,P(C_2)}$$

$P(C_k|x)$: Posterior Probability

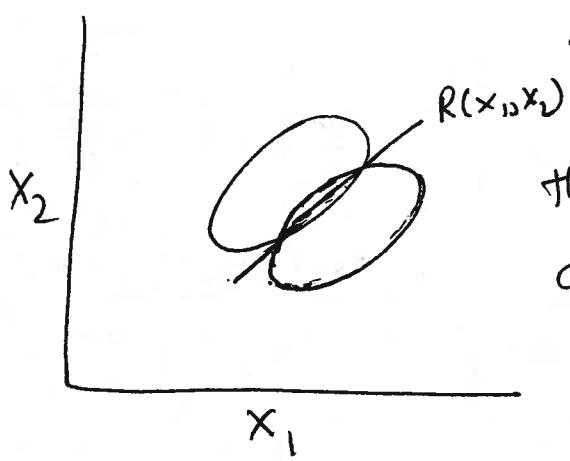
$P(x|C_k)$: Likelihood functions or class conditional Probabilities

$P(C_k)$: Prior Probability

Classification can be accomplished by applying a cut on $R(x)$ or any equivalent function thereof, such as $P(C_k | x)$.

$\therefore$ Calculate $R(x)$ or Calculate class conditional Probabilities.

We considered a simple 2D example where the two classes have Gaussian distributions.



If $\Sigma_1 = \Sigma_2$, i.e., If the covariances of the 2 Gaussians are equal, then the discriminant is a linear discriminant

$$R(x) = R(x_1, x_2)$$

$$= a x_1 + b x_2 + c$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} \Sigma^{-1}$$

Neural Networks

- Mapping of n-dimensional to m-dimensional space using

$$\mathbb{R}^n \rightarrow \mathbb{R}^m \qquad \text{"Supervised Learning"}$$

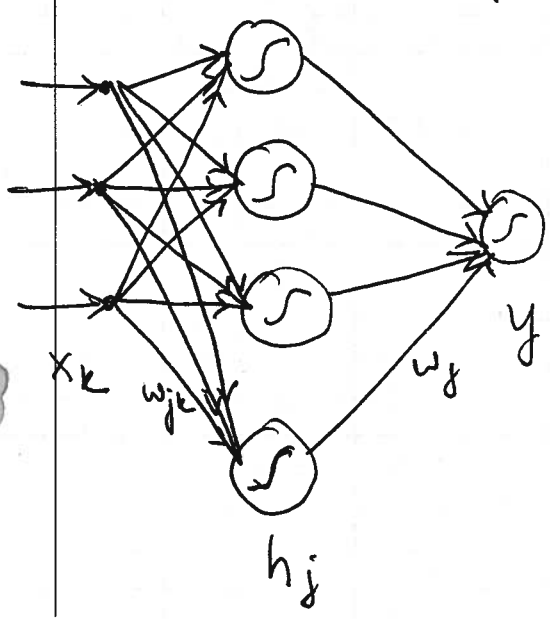
- Write output (desired output) as a non-linear function of all the inputs

$$\underset{\text{inputs}}{x_k} \longrightarrow \underset{\text{hidden functions}}{h_j} \longrightarrow \underset{\text{output}}{y}$$

$$y \equiv f(x_k)$$
$$= g\left(\sum w_j \, g\left(\sum w_{jk} x_k\right)\right)$$

where $g(a) = \frac{1}{1+e^{-a}}$ or $g(a) = \tanh a$ etc.

(I have ignored thresholds for simplicity.)



All that neural network algorithm does is to calculate the parameters w_{jk} and w_j by minimizing the error function

$$E = \frac{1}{2} \sum_p \left(y^p - t^p\right)^2 \qquad p \equiv \text{pattern.}$$

desired output

You can minimize the error function and determine the parameters using MINUIT if you like!

But, it is much more efficient to use the stochastic minimization methods used in Neural Network algorithms.

⟹ Standard Back Propagation

→ Start with w_{jk}, w_j randomly chosen in $\{-1, +1\}$

then update weights after each pattern/example

$$w_{new} = w_{old} + \Delta w$$

$$\Delta w = -\eta \frac{\partial E}{\partial w} + k \cdot w_{old}$$

↑ learning strength

↳ partial derivative of E w.r.t. w

↳ momentum term

→ There are many other algorithms for estimating the network parameters

Principal Component Analysis

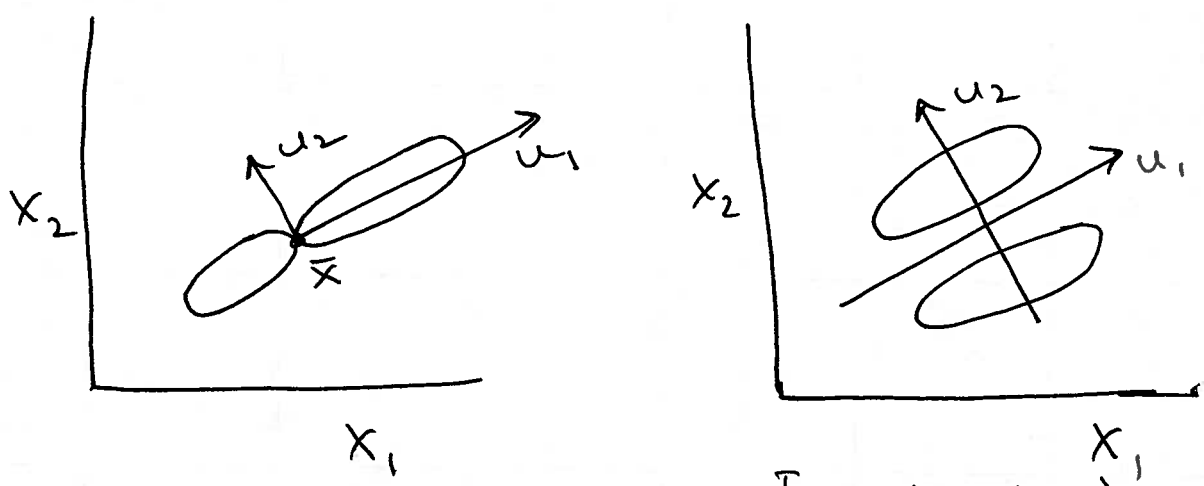
Unsupervised technique for dimensionality reduction

Goal: Find a new set of orthonormal vectors (or variables) for the given data vector $\vec{x}$ so that some dimensions (that are not important can be ignored.

$$\vec{x} = \sum_{i=1}^{d} z_i u_i \implies \tilde{x} = \sum_{i=1}^{M} z_i u_i + \sum_{i=M+1}^{d} b_i u_i$$

$M < d$

less important



Ignoring $u_2 \Rightarrow$ no discrimination

PCA $\equiv$ Rotation of coordinate system to new orthonormal system

Independent Component Analysis (ICA)

The new axes need not be orthogonal.
Find axes so that the projections are statistically independent.

Self-Organizing Maps (SOM)

- Mapping of n-dimensional data onto 2D (Unsupervised learning)

$$\mathbb{R}^n \rightarrow \mathbb{R}^2$$ (A regular array of nodes in 2D)

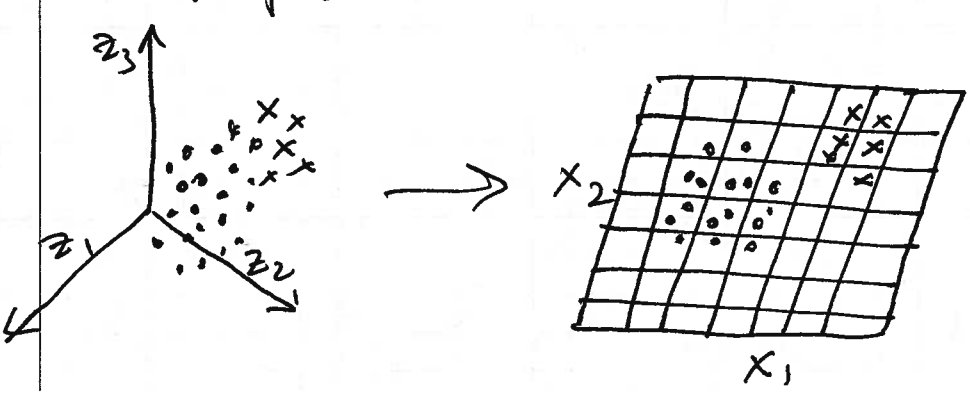
- with every node i a reference vector m_i in $\mathbb{R}^n$ is associated

- Input vector $\vec{x} \in \mathbb{R}^n$ is compared with m_i; the input is mapped to the best match.

$$\min\{\|\vec{x} - m_i\|\}$$ is the best match

- SOM performs "non-linear" projection of the Prob. density function $p(\vec{x})$ onto 2D.
- The mapping should preserve local structure of $p(\vec{x})$

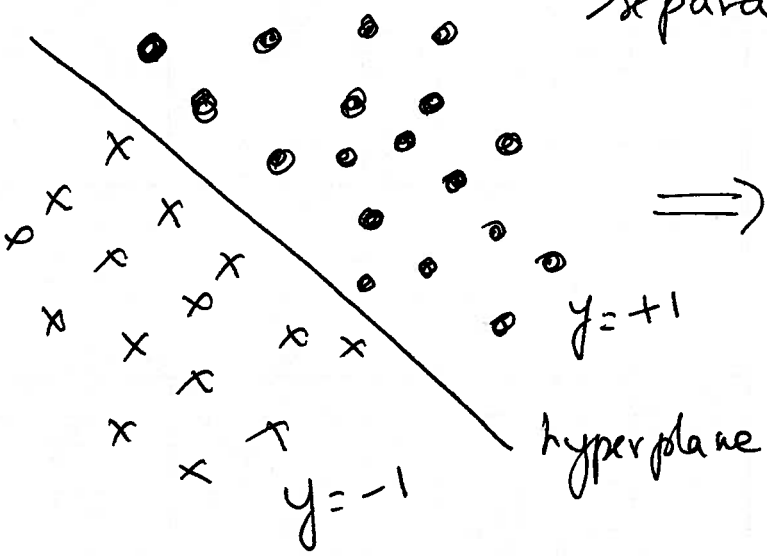
$$\mathbb{R}^3 \rightarrow \mathbb{R}^2$$



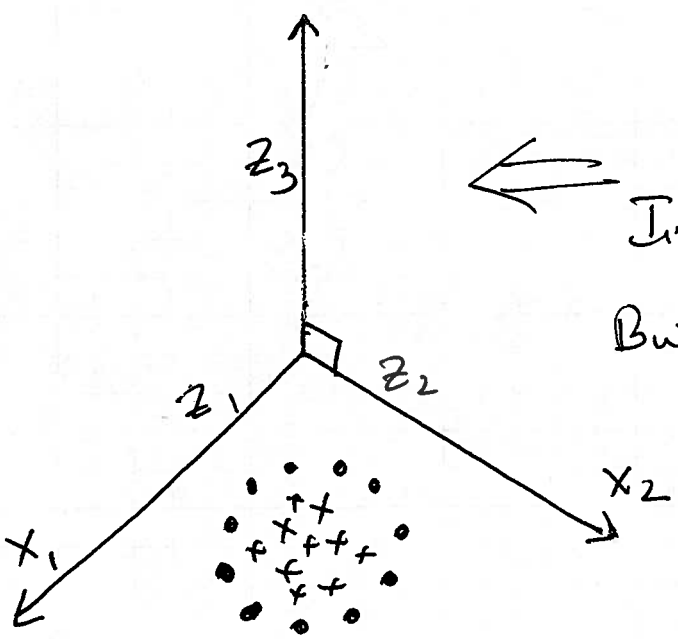
$$\{z_1, z_2, z_3\} \rightarrow \{x_1, x_2\}$$

Support Vector Machines

Problem: Find a hyperplane that optimally separates ● from X



$\Longrightarrow$ easy!



$\Longleftarrow$ Hard in 2-dimensions! Impossible!

But, easy if you map to 3 dimensions

$\therefore$ Map

$$\phi : (x_1, x_2) \rightarrow (z_1, z_2, z_3)$$

$$\mathbb{R}^2 \longrightarrow \mathbb{R}^3$$

so that in the space $\mathbb{R}^3$ the points can be separated with a hyper-plane

$$w_1 z_1 + w_2 z_2 + w_3 z_3 + b = \phi$$

SVM, contd

The discriminant surface is a hyper-plane in a higher dimensions than the original, i.e., the problem can be linearized and

$$R(x) = \sum_i w_i x_i + w_0$$

Cover's Theorem:

Any set of points in a multi-dimensional space can be separated with high probability by mapping the points to a higher dimensional space provided that

- the map is non-linear
- the dimensionality of the image space is sufficiently high

Analysis Examples

Cross section Measurement ($\sigma_{t\bar{t}}$)

$$\sigma_{t\bar{t}}(m_t) = \frac{N - B}{A(m_t) \cdot \mathcal{L}}$$

N = No. of data events observed in the analysis

B = Estimated Background

$A(m_t)$ = Total acceptance of the experiment including the analysis (I would like to call this ϵ_{tot} → total efficiency)

$$A = \epsilon_{trig} \cdot \epsilon_{ID1} \cdot \epsilon_{ID2} \cdots \epsilon_{cut} \cdot \epsilon_{geom} \cdot BR$$

↑ Branching Ratio

$\mathcal{L}$ = Integrated Luminosity of the data sample

If you combine many channels to calculate the cross section,

$$\sigma_{tt}(m_t) = \frac{\sum_i N_i - B_i}{\sum_i A_i(m_t) \cdot \mathcal{L}_i}$$

i denotes a channel

If you do not know the mass of the top quark, you are obliged to calculate the cross section at all possible m_t.

TABLE I. Triggers used to collect top signal sample ...

Name	Run period	Exposure (pb^{-1})	Level 1	Level 2	Used by
ELE-HIGH	1a	11.0	1 EM tower, $E_T > 10$ GeV GB	1 isolated e, $E_T > 20$ GeV	$e+$jets
ELE-JET	1a	14.4	1 EM tower, $E_T > 10$ GeV , $\|\eta\| < 2.6$ 2 jet towers, $E_T > 5$ GeV MRBS	1 e, $E_T > 15$ GeV , $\|\eta\| < 2.5$ 2 jets ($\Delta R = 0.3$), $E_T > 10$ GeV , $\|\eta\| < 2.5$ $\not{E}_T^{cal} > 10$ GeV	ee,$e\mu$,$e\nu$ $e+$jets $e+$jets/μ
ELE-JET-HIGH	1b	98.0	1 EM tower, $E_T > 12$ GeV , $\|\eta\| < 2.6$ 2 jet towers, $E_T > 5$ GeV , $\|\eta\| < 2.0$ ML	1 e, $E_T > 15$ GeV , $\|\eta\| < 2.5$ 2 jets ($\Delta R = 0.3$), $E_T > 10$ GeV , $\|\eta\| < 2.5$ $\not{E}_T^{cal} > 14$ GeV	ee,$e\mu$,$e\nu$ $e+$jets $e+$jets/μ
ELE-JET-HIGH	1c	1.9	1 EM tower, $E_T > 12$ GeV , $\|\eta\| < 2.6$ 2 jet towers, $E_T > 5$ GeV , $\|\eta\| < 2.0$ ML	1 e, $E_T > 15$ GeV , $\|\eta\| < 2.5$ 2 jets ($\Delta R = 0.3$), $E_T > 10$ GeV , $\|\eta\| < 2.5$ $\not{E}_T^{cal} > 14$ GeV	ee,$e\mu$,$e\nu$ $e+$jets/μ
ELE-JET-HIGHA	1c	11.0	1 EM tower, $E_T > 12$ GeV , $\|\eta\| < 2.6$ 2 jet towers, $E_T > 5$ GeV , $\|\eta\| < 2.0$ 1 EX tower, $E_T > 15$ GeV ML	1 e, $E_T > 17$ GeV , $\|\eta\| < 2.5$ 2 jets ($\Delta R = 0.3$), $E_T > 10$ GeV , $\|\eta\| < 2.5$ $\not{E}_T^{cal} > 14$ GeV	ee,$e\mu$,$e\nu$ $e+$jets/μ
EM1-EISTRKCC-MS	1b	93.4	1 EM tower, $E_T > 10$ GeV 1 EX tower, $E_T > 15$ GeV GC, NoL0	1 isolated e w/track, $E_T > 20$ GeV $\not{E}_T^{cal} > 15$ GeV	$e\nu$ $e+$jets
MU-ELE	1a	13.7	1 EM tower, $E_T > 7$ GeV 1 μ, $\|\eta\| < 2.4$ MRBS	1 e, $E_T > 7$ GeV 1 μ, $E_T > 5$ GeV , $\|\eta\| < 2.4$	$e\mu$
	1b	93.9	1 EM tower, $E_T > 7$ GeV 1 μ, $\|\eta\| < 2.4$ GC	1 e, $E_T > 7$ GeV , $\|\eta\| < 2.5$ 1 μ, $p_T > 8$ GeV/c , $\|\eta\| < 2.4$	$e\mu$
MU-ELE-HIGH	1c	10.6	1 EM tower, $E_T > 10$ GeV , $\|\eta\| < 2.5$ 1 μ, $\|\eta\| < 2.4$ GC	1 e, $E_T > 10$ GeV , $\|\eta\| < 2.5$ 1 μ, $p_T > 8$ GeV/c , $\|\eta\| < 1.7$	$e\mu$
MU-JET-HIGH	1a	10.2	1 μ, $\|\eta\| < 2.4$ 1 jet tower, $E_T > 5$ GeV GB	1 μ, $p_T > 8$ GeV/c , $\|\eta\| < 1.7$ 1 jet ($\Delta R = 0.7$), $E_T > 15$ GeV	$e\mu$,$\mu\mu$ $\mu+$jets $\mu+$jets/μ
	1b	66.4	1 μ, $p_T > 7$ GeV/c , $\|\eta\| < 1.7$ 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.0$ GC	1 μ, $p_T > 10$ GeV/c , $\|\eta\| < 1.7$, scint 1 jet ($\Delta R = 0.7$), $E_T > 15$ GeV , $\|\eta\| < 2.5$	$e\mu$,$\mu\mu$ $\mu+$jets $\mu+$jets/μ
MU-JET-CAL	1b	88.0	1 μ, $p_T > 7$ GeV/c , $\|\eta\| < 1.7$ 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.0$ GC	1 μ, $p_T > 10$ GeV/c , $\|\eta\| < 1.7$ cal confirm, scint 1 jet ($\Delta R = 0.7$), $E_T > 15$ GeV , $\|\eta\| < 2.5$	$\mu\mu$,$\mu+$jets $\mu+$jets/μ
MU-JET-CENT	1b	48.5	1 μ, $\|\eta\| < 1.0$ 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.0$ GC	1 μ, $p_T > 10$ GeV/c , $\|\eta\| < 1.0$, scint 1 jet ($\Delta R = 0.7$), $E_T > 15$ GeV , $\|\eta\| < 2.5$	$e\mu$,$\mu\mu$ $\mu+$jets $\mu+$jets/μ
	1c	8.9	1 μ, $\|\eta\| < 1.0$ 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.0$ 2 jet towers, $E_T > 3$ GeV GC	1 μ, $p_T > 12$ GeV/c , $\|\eta\| < 1.0$, scint 1 jet ($\Delta R = 0.7$), $E_T > 15$ GeV , $\|\eta\| < 2.5$	$e\mu$,$\mu\mu$
MU-JET-CENCAL	1b	51.2	1 μ, $\|\eta\| < 1.0$ 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.0$ GC	1 μ, $p_T > 10$ GeV/c , $\|\eta\| < 1.0$ cal confirm, scint 1 jet ($\Delta R = 0.7$), $E_T > 15$ GeV , $\|\eta\| < 2.5$	$\mu\mu$, $\mu+$jets $\mu+$jets/μ
	1c	11.4	1 μ, $\|\eta\| < 1.0$ 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.0$ 2 jet towers, $E_T > 3$ GeV GC	1 μ, $p_T > 12$ GeV/c , $\|\eta\| < 1.0$ cal confirm, scint 1 jet ($\Delta R = 0.7$), $E_T > 15$ GeV , $\|\eta\| < 2.5$	$e\mu$, $\mu\mu$
JET-3-MU	1b	11.9	3 jet towers, $E_T > 5$ GeV $\not{E}_T^{cal} > 20$ GeV ML	3 jets ($\Delta R = 0.7$), $E_T > 15$ GeV , $\|\eta\| < 2.5$ $\not{E}_T^{cal} > 17$ GeV	$\mu+$jets $\mu+$jets/μ
JET-3-MISS-LOW	1b	57.8	3 large tiles, $E_T > 15$, $\|\eta\| < 2.4$ 3 jet towers, $E_T > 7$ GeV , $\|\eta\| < 2.6$ MB	3 jets ($\Delta R = 0.5$), $E_T > 15$ GeV , $\|\eta\| < 2.5$ $\not{E}_T^{cal} > 17$ GeV	$\mu+$jets ? $\mu+$jets/μ
JET-3-L2MU	1b	25.8	3 large tiles, $E_T > 15$, $\|\eta\| < 2.4$ 3 jet towers, $E_T > 7$ GeV , $\|\eta\| < 2.6$ MB	1 μ, $p_T > 6$ GeV/c , $\|\eta\| < 1.7$ cal confirm, scint 3 jets ($\Delta R = 0.5$), $E_T > 15$ GeV , $\|\eta\| < 2.5$ $\not{E}_T^{cal} > 17$ GeV	$\mu+$jets $\mu+$jets/μ
MISSING-ET	1a	13.7	$\not{E}_T^{cal} > 30$ GeV 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.6$ MRBS	$\not{E}_T^{cal} > 35$ GeV	$e\nu$
	1b	83.6	$\not{E}_T^{cal} > 40$ GeV 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.6$ GB	$\not{E}_T^{cal} > 40$ GeV	$e\nu$
MISSING-ET-HIGH	1c	0.7	$\not{E}_T^{cal} > 50$ GeV 1 jet tower, $E_T > 5$ GeV , $\|\eta\| < 2.6$ GB	$\not{E}_T^{cal} > 50$ GeV	$e\nu$

Estimating Backgrounds:

$$B = \sum_{i=\text{runs}} \left(\epsilon_{trig} \cdot \epsilon_{ID} \cdot \epsilon_{cut} \cdot \epsilon_{geom}\right)_i \cdot \mathcal{L}_i$$

using MC samples (data in some cases) that has trigger simulation, detector simulation applied and reconstructed

Expected number of signal events:

$$N_s = \left(\sum_{i=\text{runs}} A_i^{(m_t)} \mathcal{L}_i\right) \cdot \sigma_{tt}(m_t)$$

Then one expects,

$$N_{tot}^{predicted} = N_s + \underbrace{B_1 + B_2 + \ldots}_{\text{backgrounds}}$$

No. of signal events expected assuming some m_t and $\sigma_{t\bar{t}}(m_t)$

Limit or Discovery?

Calculate the Probability that $B \pm \Delta B$ (σ_B) events fluctuate to N^{obs} or more events

$$\mathbb{P} = \int_{N^{obs}}^{\infty} P(N^{obs} \mid B, \sigma_B) = \text{very small} \quad (\sim 10^{-6}.\,!)$$

$2 \times 10^{-6} \Rightarrow 4.6\,\sigma$ (Gaussian approx.)

→ we have a signal ☺

→ Calculate σ, Significance

→ claim discovery, if new.

$\mathbb{P}$ = large

⇒ data is consistent with background

You can still calculate cross section, but not very useful (large errors on σ)

⇒ Set a limit on the cross section

The two types of errors:

Statistical: Coming from data statistics,

Monte Carlo statistics for modeling of Backgrounds for background estimates

And for signal modeling to calculate signal efficiency or acceptance

Systematic:

Sources: Jet energy corrections
Validity of assumptions such as scaling
Errors in fit procedures
Parameters in the modeling of events
errors on the efficiencies

Normal practice is to sum all the individual systematic errors in quadrature

$$\Delta\sigma_{syst} = \sqrt{\Delta\sigma_{syst1}^2 + \Delta\sigma_{syst2}^2 + \cdots}$$

Result: $\sigma_{t\bar{t}}(m_t) \pm \Delta\sigma_{stat} \pm \Delta\sigma_{syst}$

$\sigma_{t\bar{t}}^{stat}$ $\sigma_{t\bar{t}}^{syst}$

Setting Limits

We did not find a signal ☹

But, not many care! You can still write a PRL!

Calculate limits on cross section and hence mass of the sought particle.

We have measured:

Signal acceptance and error $A \pm \sigma_A$

Background 1 : $B_1 \pm \sigma_{B_1}$ (Number of events)

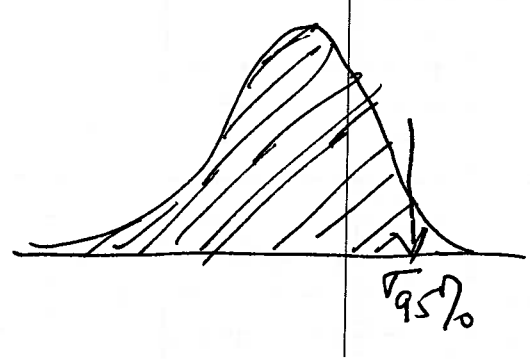
Background 2 : $B_2 \pm \sigma_{B_2}$

observed : N^{obs} events

Calculate $P(\sigma \mid N^{obs}) \equiv$ The probability of the cross section given that we observed N events

Then, set $\int_0^{\sigma_{up}} P(\sigma \mid N^{obs})\, d\sigma = 0.95$

$\Rightarrow \sigma_{95\%}$



At 95% C.L., the cross section is less than σ_{95}

Setting Limits : Some details

$$P(\sigma, a, b_1, b_2 \mid N^{obs}) = \overbrace{P(N^{obs} \mid \sigma, a, b_1, b_2)}^{\text{likelihood function}} * \underbrace{P(\sigma, a, b_1, b_2 \mid A, B_1, B_2)}_{\text{Prior probability}}$$

a, b_1, b_2 are the "true" but unknown values of acceptance and backgrounds.

$$P(N^{obs} \mid \sigma, a, b_1, b_2) = \frac{e^{-\bar{N}} \cdot \bar{N}^{N_{obs}}}{N_{obs}!}$$

$\Uparrow$ Corresponds to true mean

$$\text{Prior}(\sigma, a, b_1, b_2 \mid A, B_1, B_2) = \text{Prior}(\sigma) * \text{Prior}(a, b_1, b_2)$$
$$= d\sigma * \exp\left(-\frac{1}{2} \Delta C^T \Sigma_c^{-1} \Delta C\right) dc$$

$$C = \begin{pmatrix} a \\ b_1 \\ b_2 \end{pmatrix} ; \quad \bar{C} = \begin{pmatrix} A \\ B_1 \\ B_2 \end{pmatrix} ; \quad \Delta C = C - \bar{C}$$

$\Sigma_c =$ Error matrix

$$\boxed{P(\sigma \mid N^{obs}) = \iiint_{a\, b_1\, b_2} P(\sigma, a, b_1, b_2 \mid N^{obs})}$$

Multi-dimensional integration is done using Monte Carlo integration

Analysis Examples: Cont'd

Measurement of the Top Quark Mass

$$t\bar{t} \to W^+ b W^- \bar{b} \to l\nu b\, jj\, \bar{b}$$

General Strategy:

- kinematic fit
  - $\longrightarrow$ To the extent that you can identify and match final state objects with original partons, you can perform a kinematic fit to the $t\bar{t}$ decay hypothesis
  - $\Longrightarrow$ m_t^{fit} fitted Top mass
    (very good sensitivity to top mass)
- Mass-dependent variables

  Use kinematic quantities that are mass-dependent

  e.g., $H_T = \sum E_T^{jets}$

  Disadvantage: Broader distributions, hence lesser sensitivity to true top mass.

- Both give a mass dependent variable x,

$\Rightarrow$ $f_s(x \mid m_t)$; $f_b(x)$ $\Longleftarrow$ Serve as Templates

f_s ↑ signal ; f_b ↑ background

Step 1:

Apply corrections required + event selection

Step 2:

Kinematic fit

hypothesis:

$$p\bar{p} \to t\bar{t} + X \to (W^+ b)(W^- \bar{b}) + X \to (\ell \nu b)(q\bar{q}\bar{b}) + X$$

Use all the measured variables and following constraints:

$$\vec{p}_T^{\,t\bar{t}} = \vec{\not{E}}_T + \vec{E}_T^{\,lep} + \sum \vec{E}_T^{\,jets} \qquad 2 \text{ Constraints}$$

$$m(t \to \ell \nu b) = m(\bar{t} \to q\bar{q}\bar{b}) \qquad 1 \text{ Constraint}$$

$$m(\ell \nu) = M_W \qquad 1 \quad ''$$

$$m(q\bar{q}) = M_W \qquad 1 \quad ''$$

5 Constraints

Unknowns: Momentum of neutrino

$\Rightarrow$ 3 unknowns

$\Longrightarrow$ 2C fit.

Minimize $\chi^2 = (x - x^m)^T G (x - x^m)$

x^m: measured variables

$\Longrightarrow$ m_{fit} and a χ^2

Twelve assignments of the 4 jets to partons.

Step 3:

Build a Top Discriminant
or top probability function

Use 4 variables: $\not{E}_T$, $\mathcal{A}$, H_{T2}/H_2, $\frac{\Delta R_{jj}^{min} \cdot E_T^{min}}{E_T^{LJ}}$.

(see DØ TOP Nam PRD)

$P(top \mid D)$ using NN and a log likelihood method

↑ Data

Step 4:

Bin $P(top \mid D)$ vs M_{fit}

Perform a Bayesian fit

$$\mathcal{L}(D \mid A, a, p) = \prod_{j=1}^{M} q(N_j, p_s a_j^s + p_b a_j^b)\, q(A_j^s, a_j^s) * q(A_j^b, a_j^b)$$

$q(N, a) = \frac{e^{-a} a^N}{N!}$; p_s, p_b: signal & background strength

a_j^s, a_j^b : "true" counts for signal and background

A_j^s, A_j^b : observed counts for signal & background

j = bin index $1, \ldots M$.

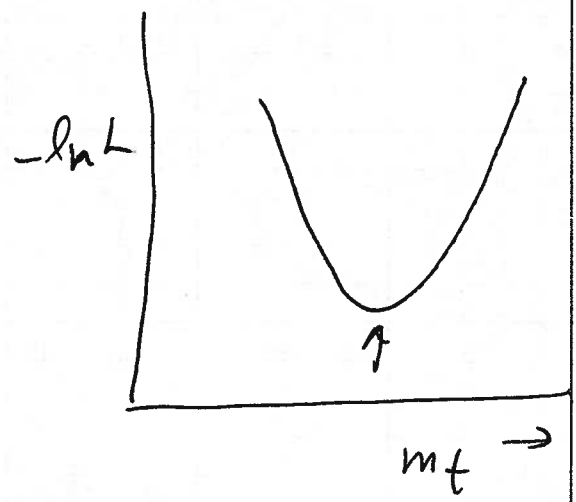
Eliminate unknown a_j^s and a_j^b by integrating over them

Analytical integration possible (Bhat et al PLB, 407, 73 (1997))

Then maximize $\mathcal{L}(D/A, p)$ or minimize $-\ln \mathcal{L}$.

$$\hat{m}_t = m_t(-\ln \mathcal{L}_{min})$$

$$\sigma_{m_t} = \pm\left[m_t(-\ln \mathcal{L}_{min} + 1) - \hat{m}_t\right]$$



Cross checks of methods by Ensemble Tests

- fake Experiments simulated → 10,000 or so
- In each case, extract m_t, σ_{m_t}, n_s, n_b etc.
- Study the distributions, pulls $\dfrac{m_t^{meas} - m_t^{true}}{\sigma(m_t)}$
- Extract confidence intervals and/or errors.
- Work out correlations with results of methods if needed.

...