Statistics Topics for Particle Physics

1) Combining results

2) Understanding Neural Networks

Louis Lyons
Imperial College & Oxford

RAL
April 2021

Combining uncorrelated exptl results

Different uncorrelated measurements $x_i \pm \sigma_i$

$x_{best} = \{\Sigma x_i/\sigma_i^2\} / \{\Sigma 1/\sigma_i^2\}$ [1]

$1/\sigma^2 = \Sigma(1/\sigma_i^2)$ [2]

{This comes from minimising (wrt x_{best}) $S = \Sigma\{(x_i - x_{best})^2/\sigma_i^2\}$}

Commonly know as χ^2

Define $w_i = 1/\sigma_i^2$ = weight ~ 'information content'

Eqns [1] and [2] become:

$x_{best} = \Sigma w_i\, x_i / \Sigma w_i$ [1']

= weighted average of x_i

$w = \Sigma w_i$ [2']

Example: All σ_i equal

x_{best} = simple average of x_i

$\sigma = \sigma_i/\sqrt{n}$

BLUE is equivalent to χ^2, but also outputs weights. Useful for assessing statistical and systematic uncertainties on x_{best}.

N. B. Better to combine data

Difference between weighted and standard averaging

Isolated island with conservative inhabitants

How many married people ?

Number of married men = 100 ± 5 K

Number of married women = 80 ± 30 K

Total = 180 ± 30 K

Wtd average = 99 ± 5 K **CONTRAST**

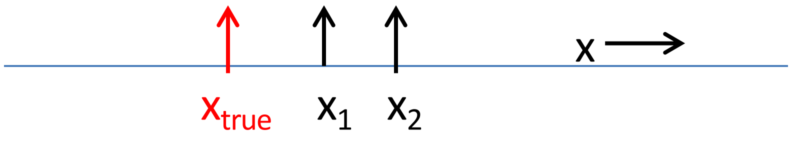
Total = 198 ± 10 K

GENERAL POINT: Adding (uncontroversial) theoretical input can improve precision of answer

Compare "kinematic fitting"

Combining: oddities

- 1 variable :
  Best combination of 2 correlated measurements can be outside range of measurements
  Peelle's Pertinent Puzzle



- 2 parameters, α β
  Uncertainties on α_{best} and β_{best} much smaller than individual uncertainties.

- 2 parameters, α β
  $\alpha_{best} > \alpha_1$ and α_2 $\beta_{best} > \beta_1$ and β_2
  Yule-Simpson Paradox

COMBINING RESULTS

- Better to combine data than combine results
  (Problems with non-Gaussian estimates
  dealing with correlations
  uncertainty estimates)

- **BEWARE** of uncertainty estimates that depend on parameter estimate

  e.g. $n \pm \sqrt{n}$ 100 ± 10 and 80 ± 9

  or $\tau \pm \tau/\sqrt{N}$ 1.00 ± 0.10 and 1.20 ± 0.12 (N=100)

  Likelihood works better

BEWARE:

Counting experiment, records in 2 separate days: 100 ± 10 and 80 ± 9 counts

Standard formulae → 88.8 ± 6.7 [1] Biassed

Sensible (and correct) approach → $(180 \pm 13.4)/2 = 90.0 \pm 6.7$ [2]

(Part of reason why PDG average b-lifetime used to be ~1ps, rather than current 1.5ps)

Solution 1:

Needs $w = 1/\sigma^2$ to be real accuracies, not estimated accuracy.

If counting for 2 equal periods with equal efficiency, etc, then expected accuracies are equal → equal weights → solution [2]

See LL, A. J. Martin and D. H. Saxon, Phys. Rev. D **41** (1990) 982

Deals with B lifetime example, and recalculates (essentially iteratively) what each experiment's uncertainties would have been as a function of lifetime

i.e. What part of the uncertainty scales with τ, and what is independent of τ.

Solution 2:

Use likelihood approach.

Combining correlated exptl results

Different uncorrelated measurements $x_i \pm \sigma_i$

$x_{best} = \{\Sigma x_i/\sigma_i^2\} / \{\Sigma 1/\sigma_i^2\}$ [1]

$1/\sigma^2 = \Sigma(1/\sigma_i^2)$ [2]

{This comes from minimising (wrt x_{best}) $S = \Sigma\{(x_i - x_{best})^2/\sigma_i^2\}$ }

For correlated variables, minimise

$$S' = \Sigma_i\Sigma_j\,(x_i - x_{best})\,M_{ij}\,(x_j - x_{best})$$

where M is the inverse of the covariance matrix $C = \begin{pmatrix} \sigma_i^2 & \mathrm{Cov} \\ \mathrm{Cov} & \sigma_j^2 \end{pmatrix}$

x_{best} outside range of x_1 and x_2 when Cov> smaller σ^2

or $\rho > \sigma_{small}/\sigma_{large}$

So if 2 similar analyses on same data, don't combine but instead use 'better' result, and use other as confirmatory. Highly correlated combination → extrapolation. Sensitive to exact values of σs and ρ.

Nice example of $\rho = \sigma_1/\sigma_2 \rightarrow w_2 = 0$

Sample 2 is subsample of Sample 1

Sensible that sample 2 is ignored in 'combination'.

Peelle's Pertinent Puzzle

Combination outside range of individual measurements

- Oak Ridge Nat Lab Memorandum, 1987
- Combining neutron + nuclei cross-sections
- Sometimes reasonable
- Sometimes unreasonable e.g. luminosity systematic for cross-sections
- Numerous solutions to Puzzle
- Again using estimated uncertainties

Combined uncertainty very small: Danger of combining profile $\mathcal{L}$s

Experiments quote $\mathcal{L}$ikelihood, profiled over nuisance parameters, so that combinations can be performed.

Very simple 'tracking' example:

* No magnetic field
* 2-D fit of straight line y = a + bx

  a = parameter of interest, b = nuisance param
* Track hits in 2 subdetectors, each of 3 planes

Straight line fit to red points has large uncertainties on intercept and on gradient

Straight line fit to blue points has large uncertainties on intercept and on gradient

Combined straight line fit to red and blue points has much smaller uncertainties on intercept and on gradient

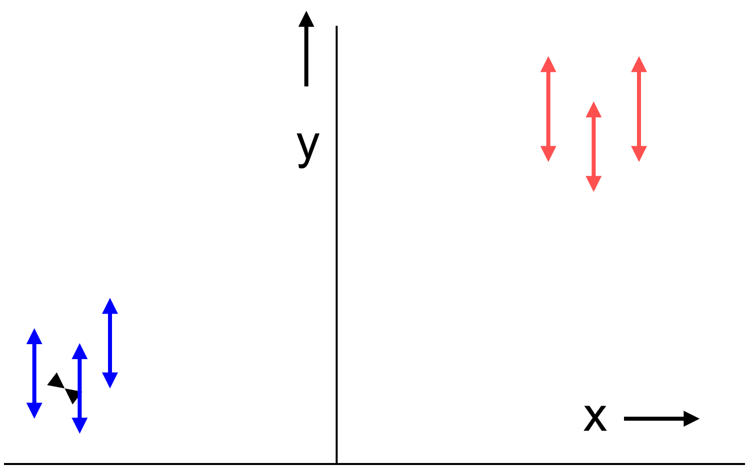
2 sub-detectors each of 3 planes.

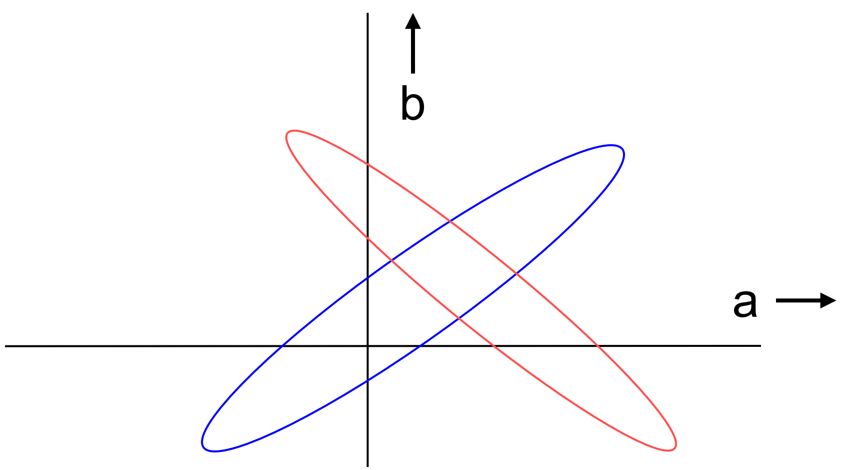
(a) Straight line fits for L1, L2 and combination.

(b) Covariance ellipses, large for L1 and L2, small for combination

Covariance of gradient and intercept proportional to minus weighted mean x

Uncertainties from different subdetectors are uncorrelated



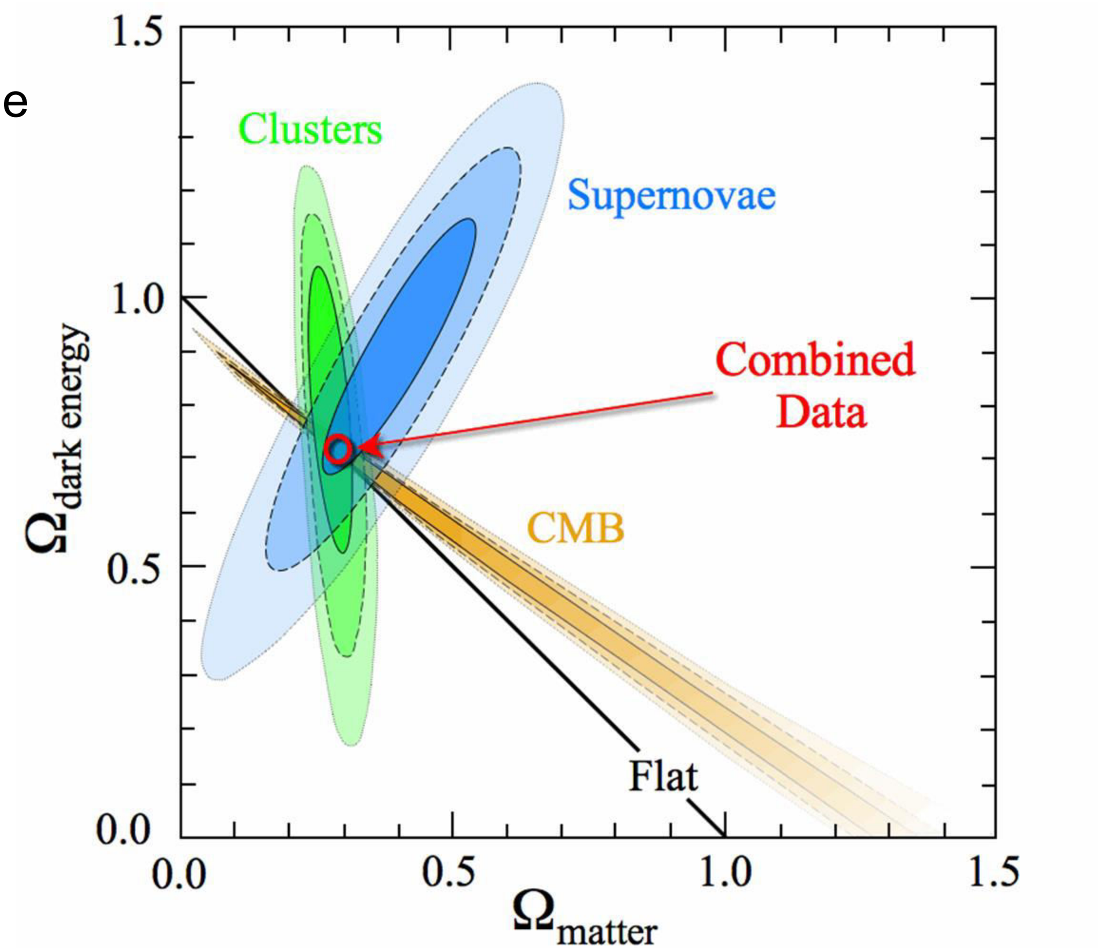


Uncertainty on $\Omega_{\text{dark energy}}$

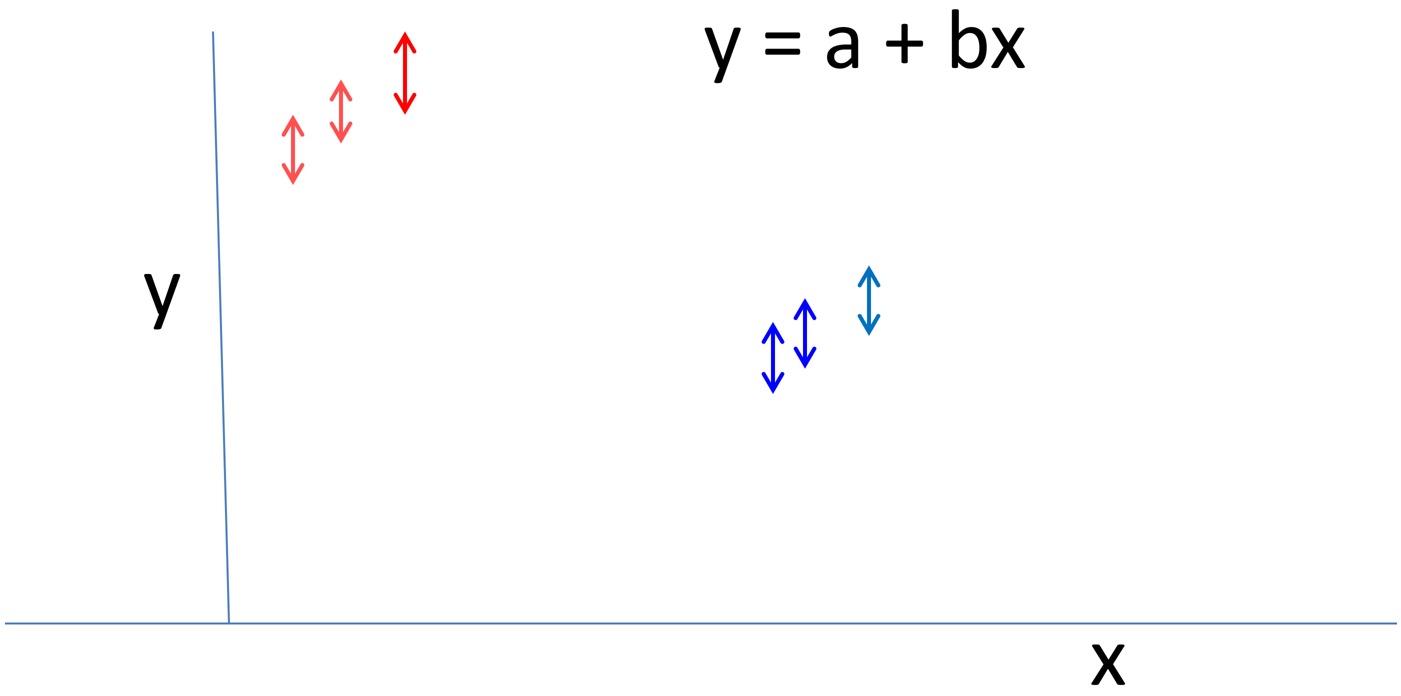
When combining pairs of variables, the uncertainties on the combined parameters can be **much** smaller than any of the individual uncertainties
e.g. $\Omega_{\text{dark energy}}$

Plot of dark energy fraction versus dark matter fraction by various methods. Each determines dark energy fraction poorly, but combination is fine, because of different correlations

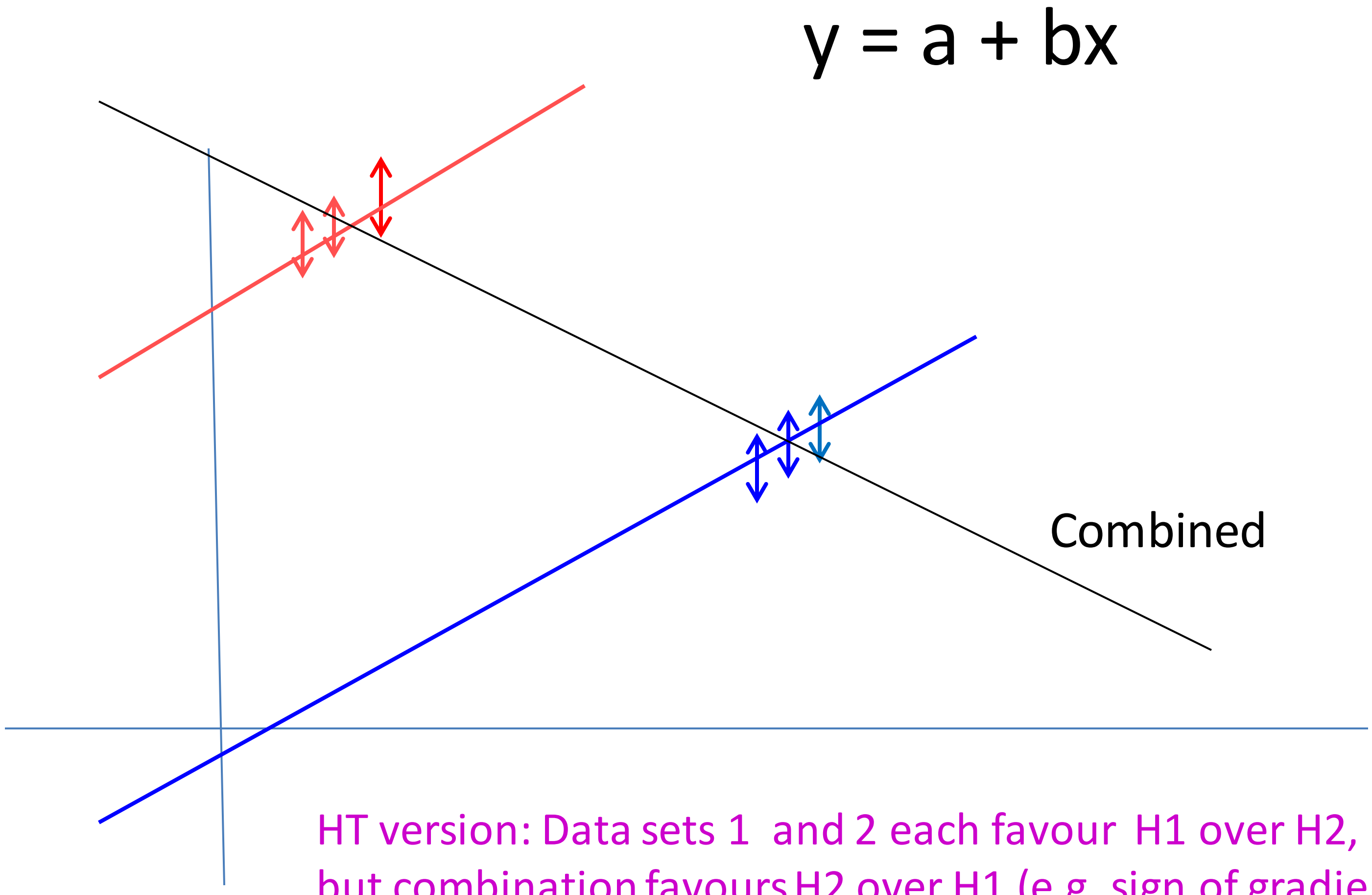
Combining Profile Likelihoods would give very large uncertainty on dark energy fraction



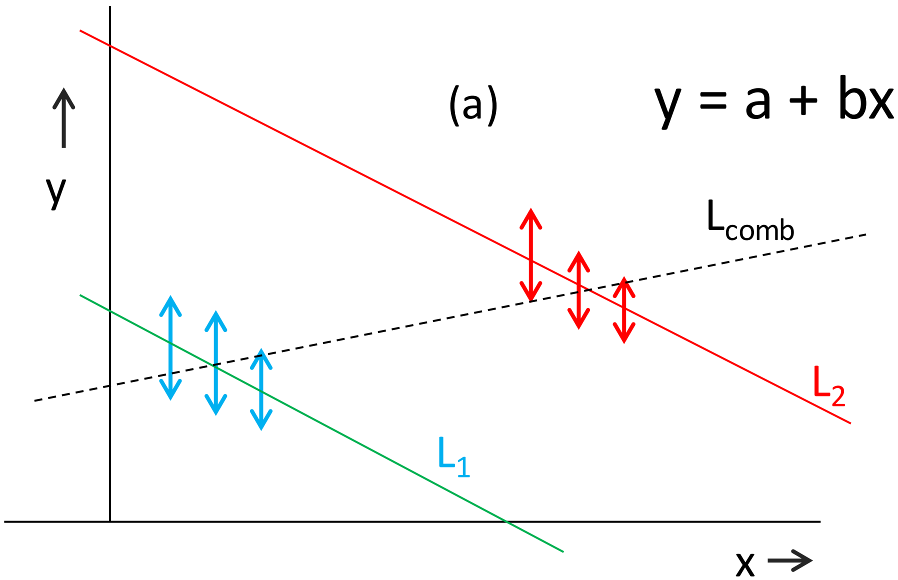
Best values of params a and b outside range of individual values (Remember PPP)



Best values of params a and b outside range of individual values



HT version: Data sets 1 and 2 each favour H1 over H2,
but combination favours H2 over H1 (e.g. sign of gradient).
Relevant for Nova and T2K on neutrino mass hierarchy?



Example where best values of a and b are outside ranges of individual values.

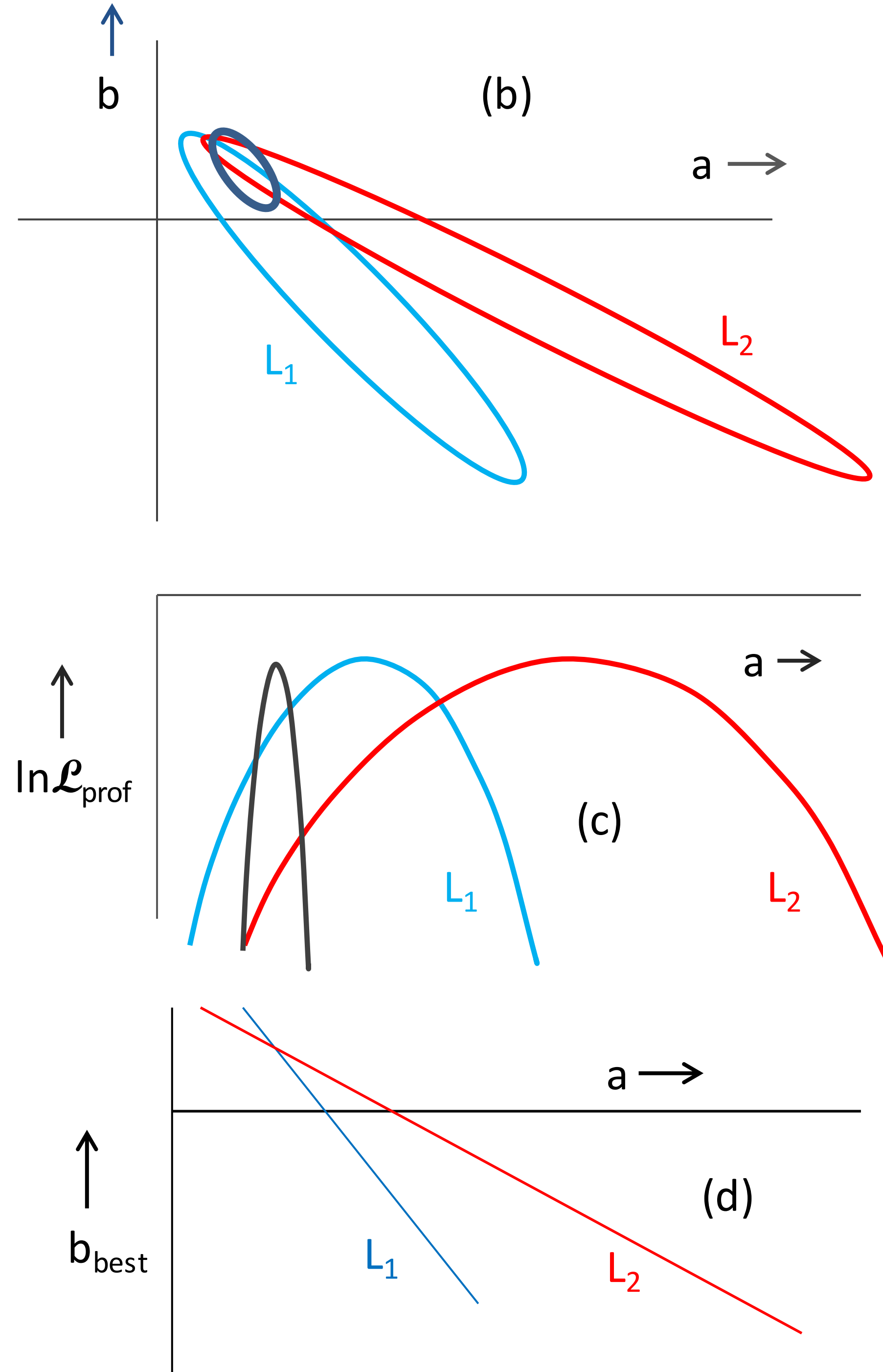
(a) Hits in sub-detectors

(b) Covariance ellipses

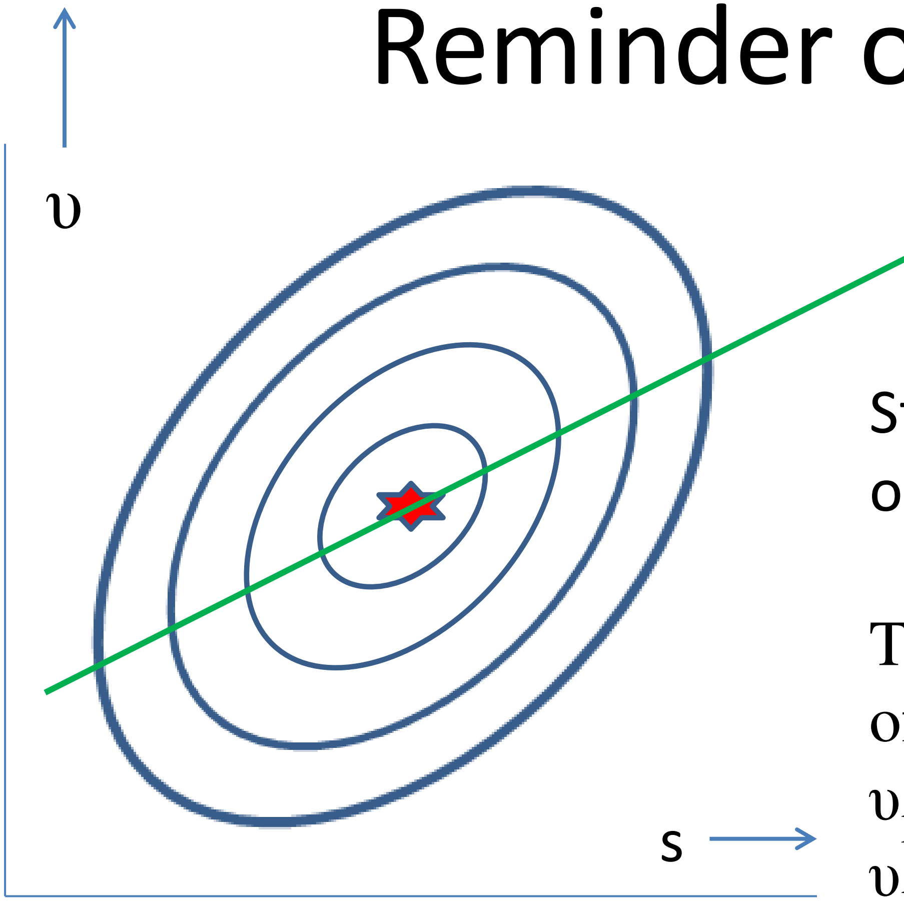
(c) $\ln\mathcal{L}_{prof}$ as function of a

(d) b_{best} as a function of a

BEWARE: Combining profile $\mathcal{L}$s will give poor result



Reminder of Profile $\mathcal{L}$



Stat uncertainty on s from width of $\mathcal{L}$ fixed at υ_{best}

Total uncertainty on s from width of $\mathcal{L}(s,\upsilon_{prof(s)}) = \mathcal{L}_{prof}$

$\upsilon_{prof(s)}$ is best value of υ at that s

$\upsilon_{prof(s)}$ as fn of s lies on green line

Contours of $\ln\mathcal{L}(s,\upsilon)$

s = physics param

υ = nuisance param

Total uncert $\geq$ stat uncertainty

Simpler example of PPP, without correlations (Yule-Simpson paradox)

Results of studies on effectiveness of drug, depending on whether patient had asthma in childhood.

The outcome for each patient is assigned a 'mark'. Higher mark means that the drug is more effective.

Numbers in 'table' below are: total 'marks for drug' divided by number of patients = average.

	No Asthma	With Asthma	Combined
Drug A	**80/2 = 40**	**640/8 = 80**	**720/10 = 72**
Drug B	**400/8 = 50**	**180/2 = 90**	**580/10 = 58**

(In both cases, the combined result lies between the separate results for the different asthma histories , as required for uncorrelated measurements.

It's just that the weighting of the two histories is different for the two drugs)

For people who have asthma in their childhood, Drug B is better than Drug A in treating this disease

For people who did not have asthma in their childhood, Drug B is better than Drug A in treating this disease

But overall, Drug A is better than Drug B in treating this disease.

Then the doctor's dilemma is:

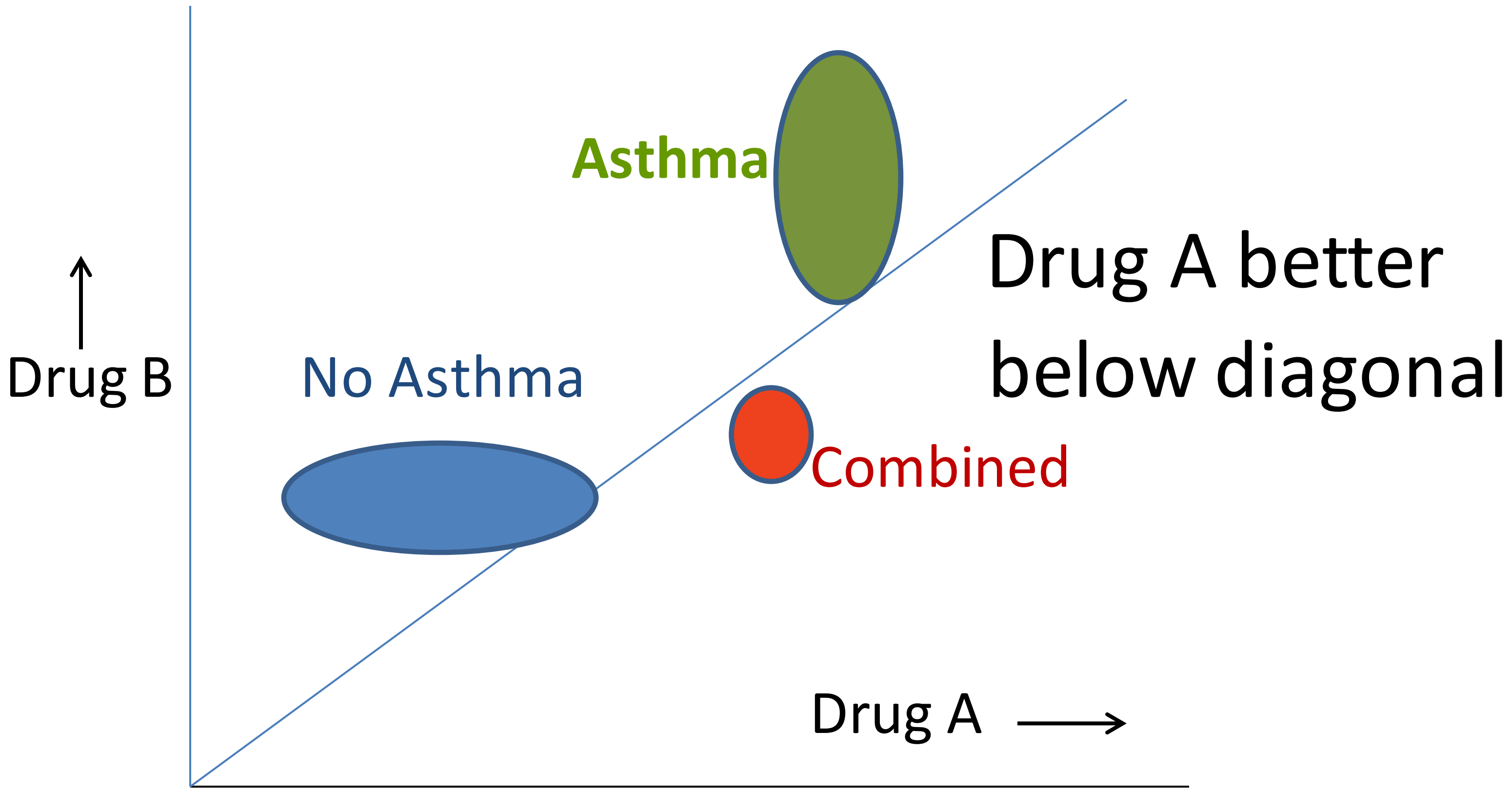
For people who have asthma in their childhood, prescribe Drug B

For people who did not have asthma in their childhood, prescribe Drug B

For people who did not know whether they had asthma in their childhood, prescribe Drug A

(even though they either had asthma or they didn't. In either case, the doctor would have prescribed Drug A)

Medical tests



For each class of patients, drug B is better

For combined set of patients, drug A is better

Doctor's Dilema?

Comments on Drug Test example

The dilemma arises even though here there are no correlations.

Also combined values are within ranges on individual values i.e. no PPP

Common feature with tracking: In both cases, major axes of covariance ellipses not parallel

Rotation of axes is even sensible in medical case.

Unknown ρ

- What to do if correlations are unknown?
  e.g. Old neutrino cross-section data

New archive note by Lukas Koch (Oxford)

"Robust test statistics for data with missing correlation information"

https://arxiv.org/abs/2102.06172 (Feb 2021)

Summary of Combination Oddities

- Including theory can help
- Estimated uncertainties: 100 ± 10 and 80 ± 9
- PPP: Combination outside range of individual σ
- Extrapolation can be correct
- Combined σ can be $<<$ individual σ
- Profile Likelihood loses information
- Extrapolation can occur without correlations (e.g. doctor's dilemma)

Combining p-values

For comparing hypothesis H with data, p = probability of obtaining result = data, or more extreme.

p is **NOT** probability that H = true, given the data

Much better to combine data e.g.

1) Small p-values from different analyses could result from very different discrepancies.
2) Correlated systematics
3) Bob Cousins: Combination method is ambiguous:

p_i are supposedly uniformly distributed and independent.

How to construct $p_{comb}(p_i)$ such that it is uniformly distributed over hyper-cube?

Optimal method depends on other information, e.g.

Data set 1. Histogram of 100 bins. H = constant

Weighted sum of squares S = 90, p_1=0.4

Data set 2. One measurement. H predicts 49 events. Observe 84 events. $p_2 = 3\ 10^{-6}$

p_{comb} likely to be small. But S_{comb} =115 → p_{comb}= 0.16

Combination method for p-values

1) Don’t combine p-values

2) Select smallest p_i (and calculate prob)

3) Use Π = product of p_i, and calculate
   p_{comb} = prob that $\Pi < \Pi_{obs}$

e.g. For 2 p-values, $p_{comb} = p_1 p_2(1 - \ln(p_1 p_2)) >= p_1 p_2$

4) Stouffer: $z_{comb} = \Sigma z_i/\sqrt{N}$,
where z_i is z-score corresponding to p_i
(e.g. $z_i = 5$ for $p_i = 3\ 10^{-7}$)

For longer list, see Heard & Rubin-Delanchy (2017)
“Choosing Between Methods of Combining p-values”
https://core.ac.uk/download/pdf/146459765.pdf

MULTIVARIATE ANALYSIS

Example: Aim to separate signal from background

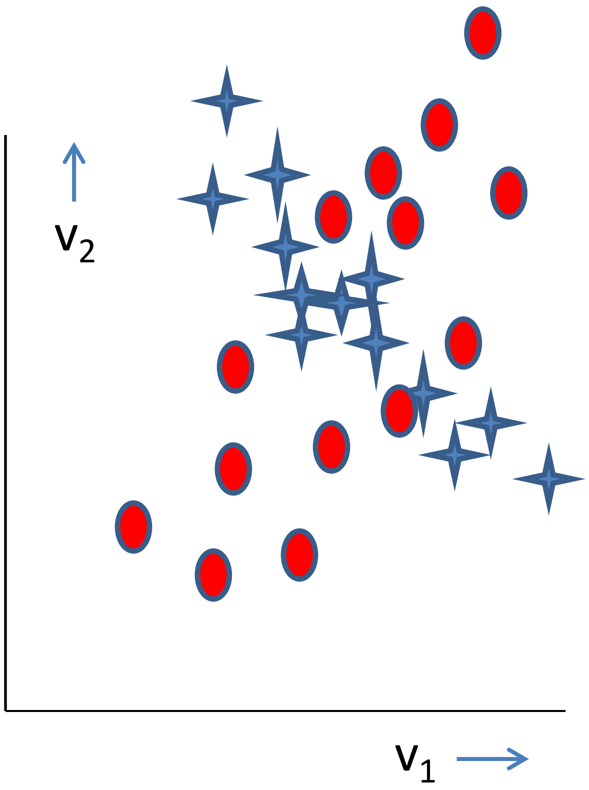
Neyman-Pearson Lemma:
Imagine all possible contours that select signal with efficiency ε (Loss = Error of 1st Kind)
Best is one containing minimal amount of background (Contamination = Error of 2nd Kind)

Equivalent to ordering data by

$\mathcal{L}$-ratio $= \mathcal{L}_s(v_1, v_2, \ldots\ldots) / \mathcal{L}_b(v_1, v_2, \ldots)$

IF variables are independent

$\mathcal{L}$-ratio $= \{\mathcal{L}_s(v_1)/\mathcal{L}_b\{v_1)\}$ x $\{\mathcal{L}_s(v_2)/\mathcal{L}_b(v_2)\}$ x .....



PROBLEM:

Don't know $\mathscr{L}$-ratio exactly because:

1) Signal & bdg generated by M.C. with finite statistics
2) Nuisance params (systematics) and signal params
3) Neglected sources of bgd
4) Hard to implement in many dimensions

METHODS TO DEAL WITH THIS

Cuts

Kernel Density Estimation

Fisher Discriminant

Principal Component Analysis

Boosted Decision Trees

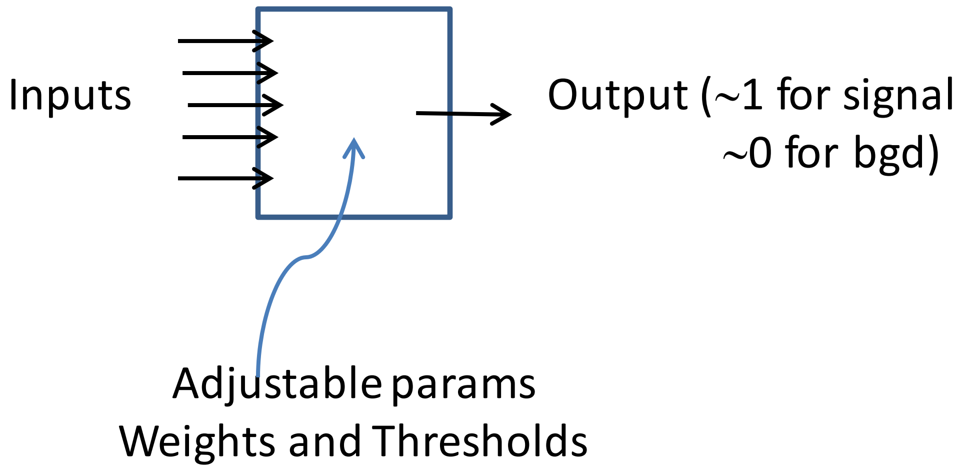
Support Vector Machines

Neural Nets

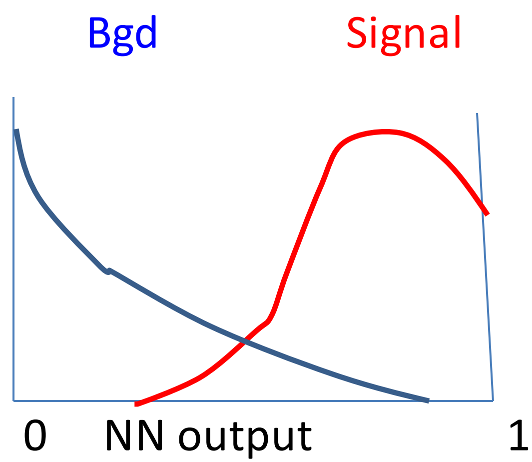
Deep Nets

NEURAL NETWORKS

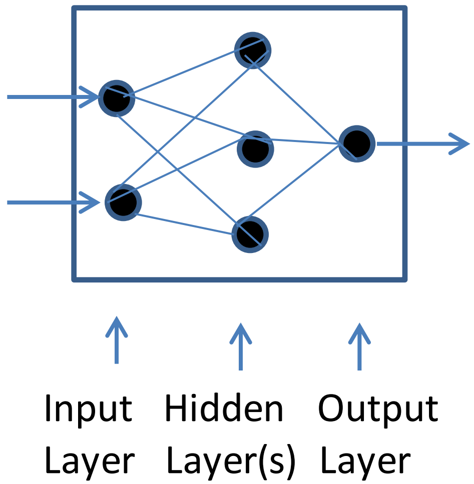
Typical application: Classify events as signal or bgd



- Learning process:
  Input = Known signal & bgd (e.g M.C.)
  Adjust params → ‘Best’ output
- Testing process
  Make sure not ‘overtraining’
- Use trained network on actual data
  Classify events as signal if output > cut



HOW DOES IT WORK?



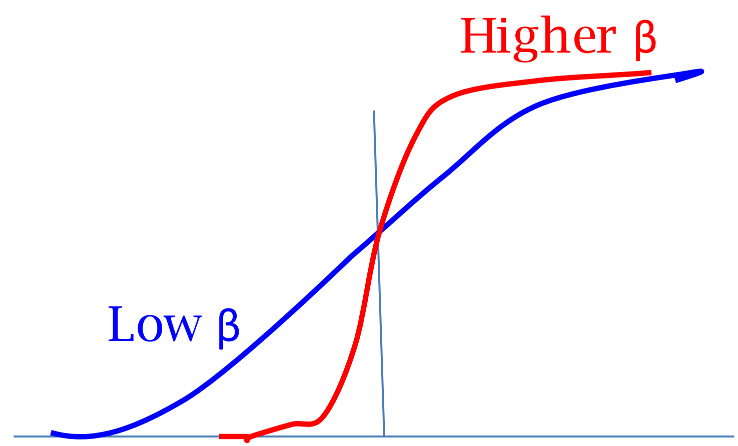
For each hidden or output node j

$\text{Output}_j = F\,[\Sigma\ \text{Input}_i * W_{ij} \ \ + T_j]$

(W and T = network params)

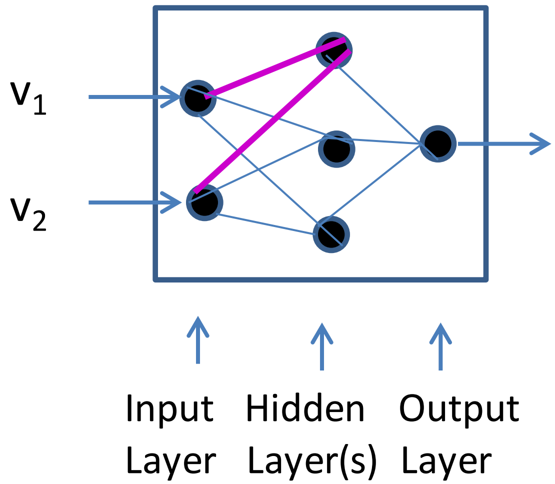
Typical $F(x) = 1/(1+ e^{-\beta x})$ Sigmoid

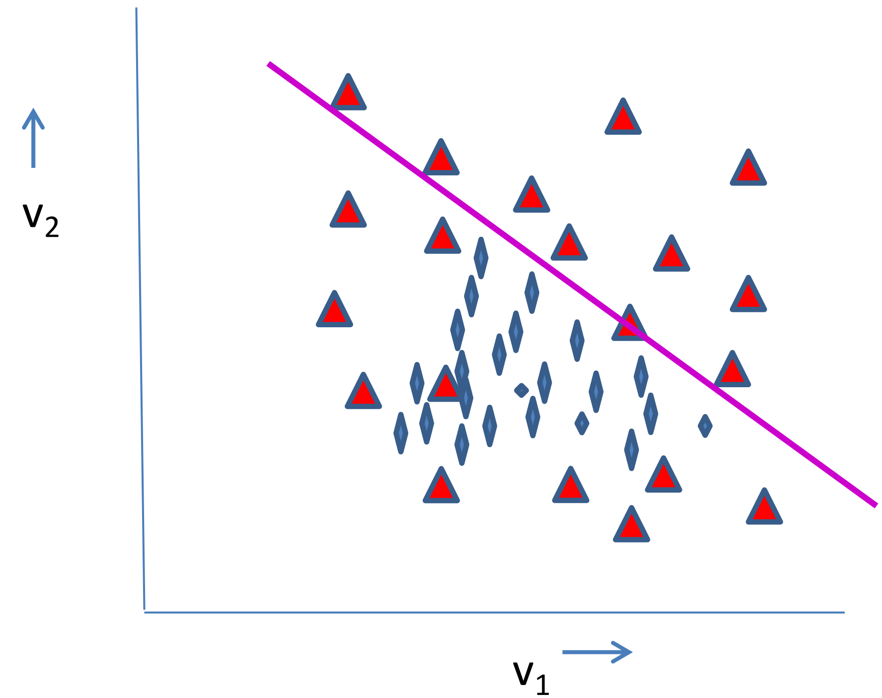
For large β, output of node j is 'ON' if $\Sigma\ I_i\, w_{ij} + T_j > 0$, and 'OFF' otherwise



Dividing contour is 'hyper-plane' in I space

HOW DOES IT WORK?





For First hidden node

Straight line is

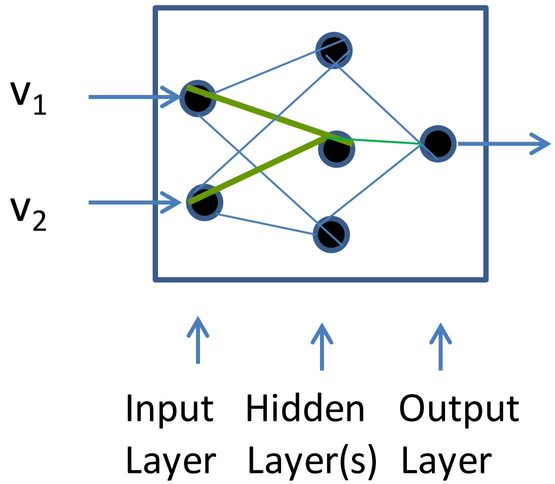
$$w_{\mathbf{11}}*v_1 + w_{21}*v_2 + T_{10} = 0$$

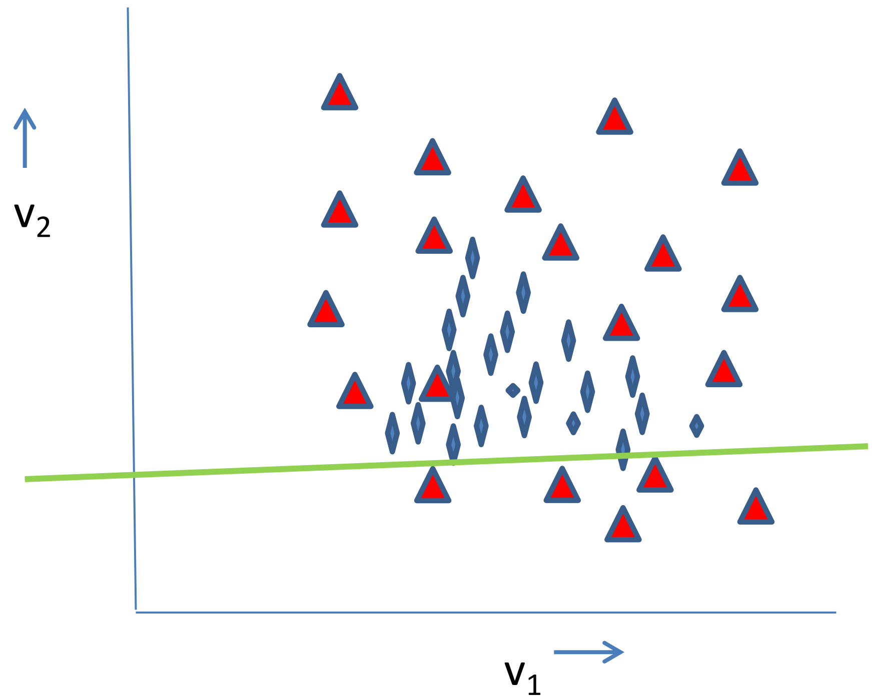
where

w_{ij} is weight from i^{th} input node to j^{th} hidden node

T_{k0} is threshold for k^{th} hidden node

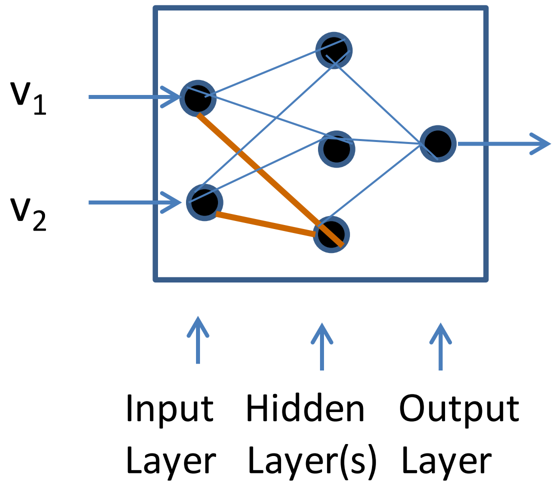
HOW DOES IT WORK?



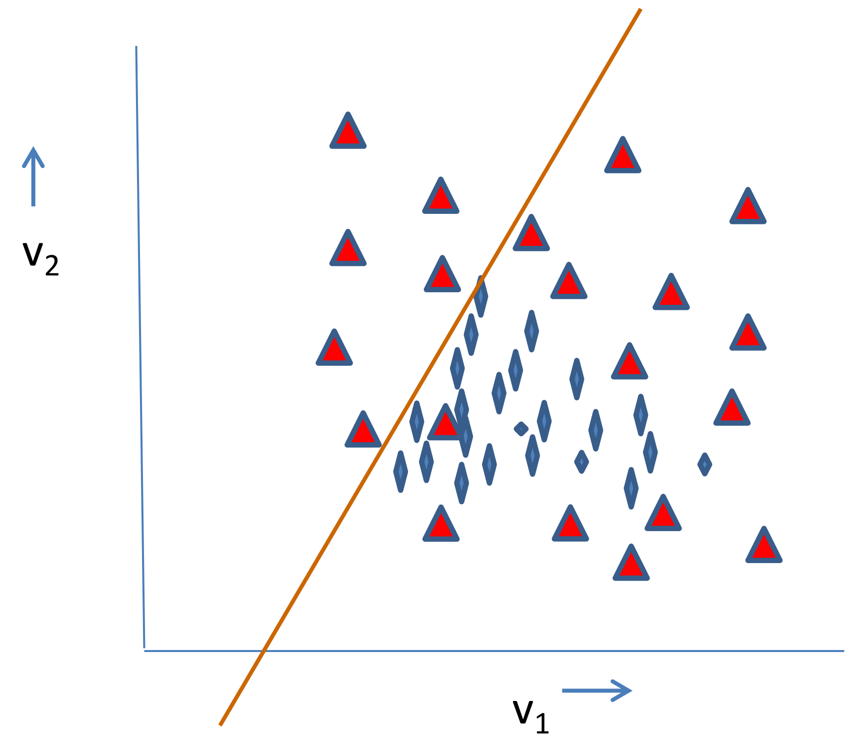


For second hidden node

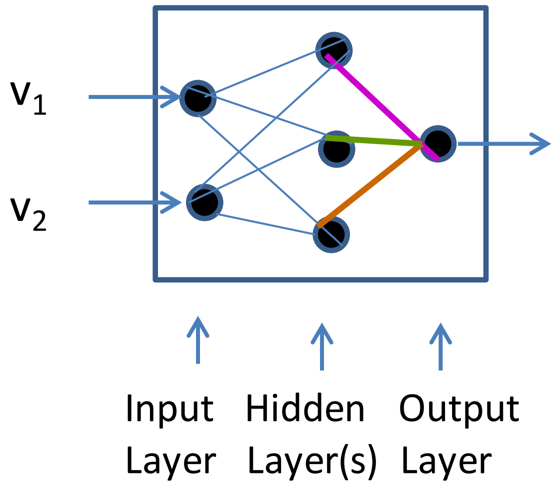
HOW DOES IT WORK?

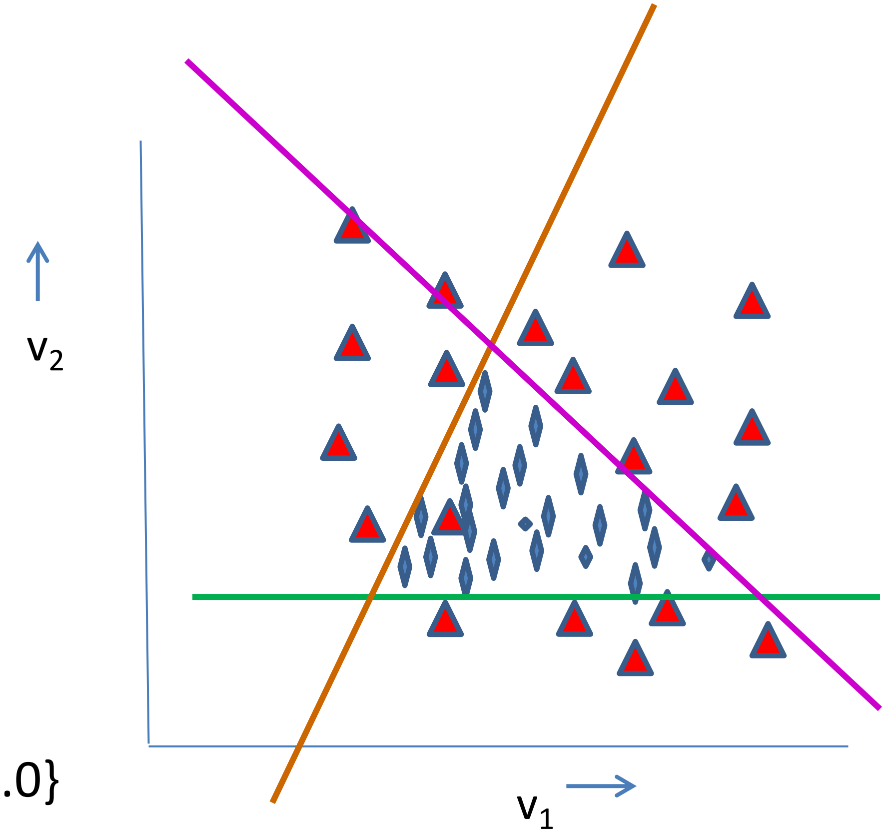


For third hidden node



HOW DOES IT WORK?

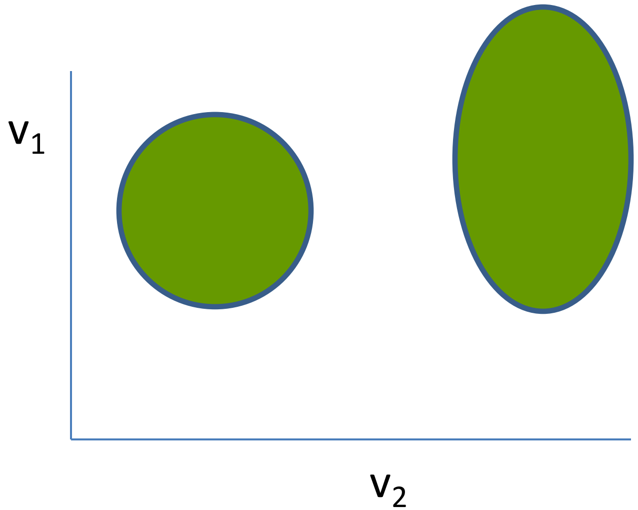


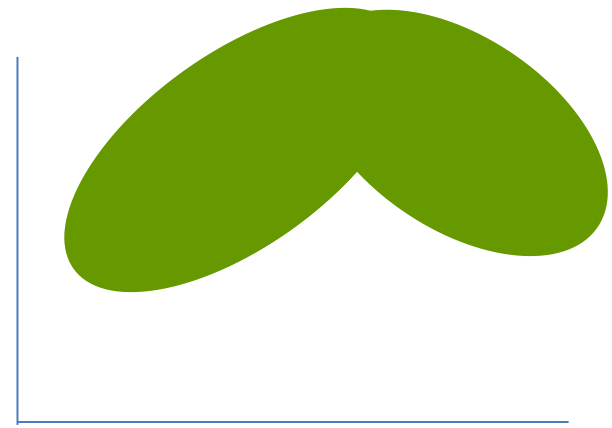


Output = Sigmoid{$0.4H_1 + 0.4H_2 + 0.4H_3 - 1.0$}
Output is 'On' only if H_1 H_2 H_3 all are 'On'

N.B.
* Complexity of final region depends on number of hidden nodes.
* Finite β → rounded edges for selected region; and contours of constant output in (v_1, v_2) plane.

When do we need more than one Hidden Layer?

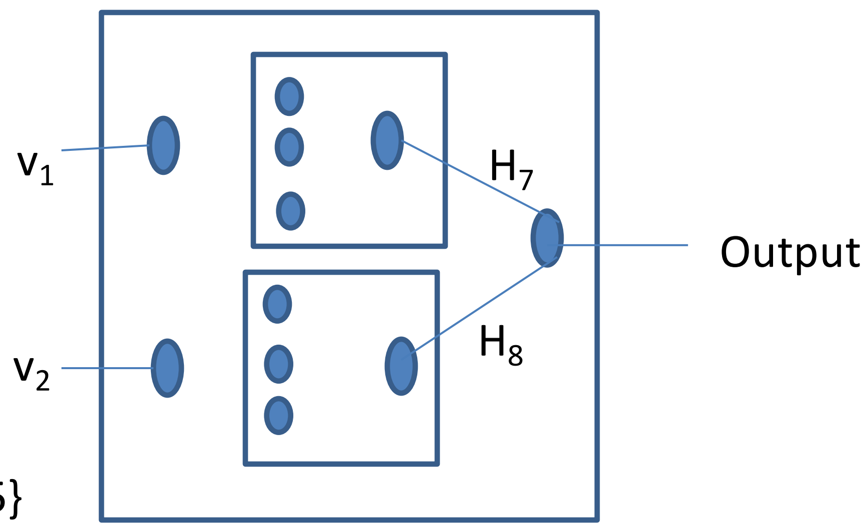




Input nodes connected to all 1st hidden layer nodes

1st hidden layer nodes connected to 2nd hidden layer nodes in same rectangle

Output = Sigmoid$\{H_7 + H_8 - 0.5\}$
i.e. Output is **ON** if either or both of H4 and H5 are **ON** (logical OR)



BEWARE

- Training sets are reliable
- Don’t train with variable you want to measure
- Data does not extend outside range of training samples (in multi-dimensions)
- Don’t overtrain
- Approx equal numbers of signal and bgd

Is NN better* than simple cuts?

In principle, NO

Can cut on complicated variable e.g. NN output

In practice: YES

But:

Better NN performance → more work by ‘Cuts’ analysis to improve performance

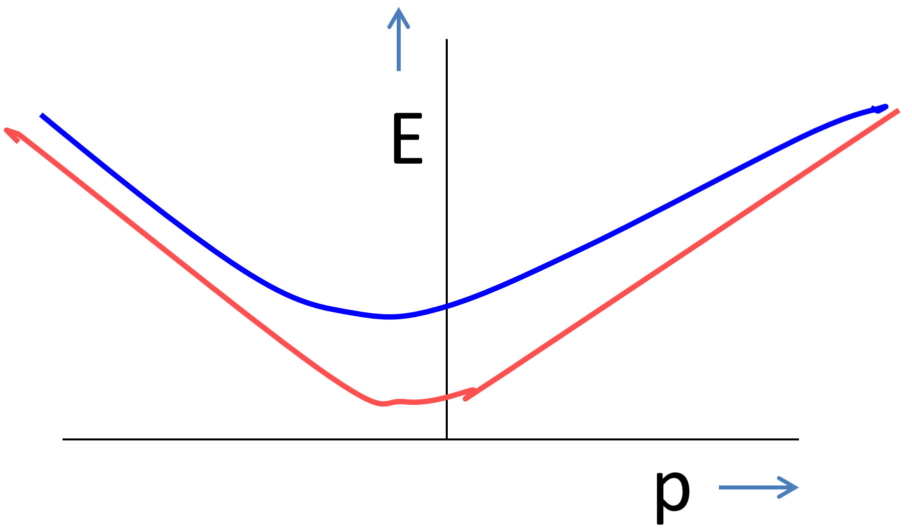
* Better = improved efficiency v mistag rate

SIMPLE EXAMPLE

Try to separate π and proton using E and p

π: $E^2 = p^2 + m_\pi^2$

P: $E^2 = p^2 + m_p^2$



Easy: p = 0 → 2 GeV

Harder: p = -4 → 4 GeV

Hardest: p_x, p_y, p_z = -4 → 4 GeV

More realistic: Add expt scatter of data wrt curves

PHYSICS EXAMPLE

Separate b-jets from light flavour, gluons, W, Z:

Input variables: Track IPs, SV mass, distance, quality, etc.

Output: 0 → 1

Issues:

Pre NN cuts

Training and testing samples (Where from? How many events? Ratios of different bgds,….)

How many inputs?

Network structure

How many networks?

Single output or several

Systematics (use different sets of testing events}

Stability wrt NN cut

NN Summary

- ADVANTAGES:

  Very flexible

  Correlations OK

  Tunable cut
- DISADVANTAGES

  Training takes time

  Tendency to include too many variables

  Treat as black box

* Past attitude: Need to convince colleagues NN is sensible

More recently: Why aren't you using NN?

Now/future: Why aren't you using a Deep Network?

Conclusions

Resources:

Software exists: e.g. RooStats, Combine

Books exist: Barlow, Cowan, James, Lista, Lyons, Roe,…..

`Data Analysis in HEP: A Practical Guide to Statistical Methods' , Behnke et al.

PDG sections on Prob, Statistics, Monte Carlo

CMS, ATLAS and LHCb have Statistics Committees (and BaBar and CDF earlier) – see their websites.

PHYSTAT Workshops: LHC, Neutrino, Dark Matter, Flavour Physics

Before re-inventing the wheel, try to see if Statisticians have already found a solution to your statistics analysis problem.

Don't use your square wheel if a circular one already exists.

"Good luck"


No thanks!
We are too busy