PRACTICAL STATISTICS #5

Hunter + Dog paradox

Bayes + Frequentism, contd

- 2 more Bayes examples
- Frequentist approach

See #4

Feldman - Cousins

- Ordering rule
- Flip - flop
- Gaussian + Poisson examples
- ν oscillations

Systematics

- Shift method
- Profile $\mathcal{L}$
- Bayes
- Frequentist
- Mixed: Cousins + Highland

Multivariate analysis

- Neural networks

BLUE combination technique

LOUIS LYONS
CDF

FELDMAN - COUSINS

WANT TO AVOID EMPTY CLASSICAL INTERVALS $\Rightarrow$

USE "$\mathcal{L}$ RATIO ORDERING PRINCIPLE"

TO RESOLVE AMBIGUITY ABOUT "WHICH 90% REGION?" $\Rightarrow$

[NEYMAN-PEARSON SAY $\mathcal{L}$ RATIO IS BEST FOR HYPOTHESIS TESTING]

NO FLIP-FLOP PROBLEM $\Rightarrow$

90% classical interval for Gaussian

$\sigma = 1$

$\mu \geq 0$

e.g. $m^2(\nu_e)$

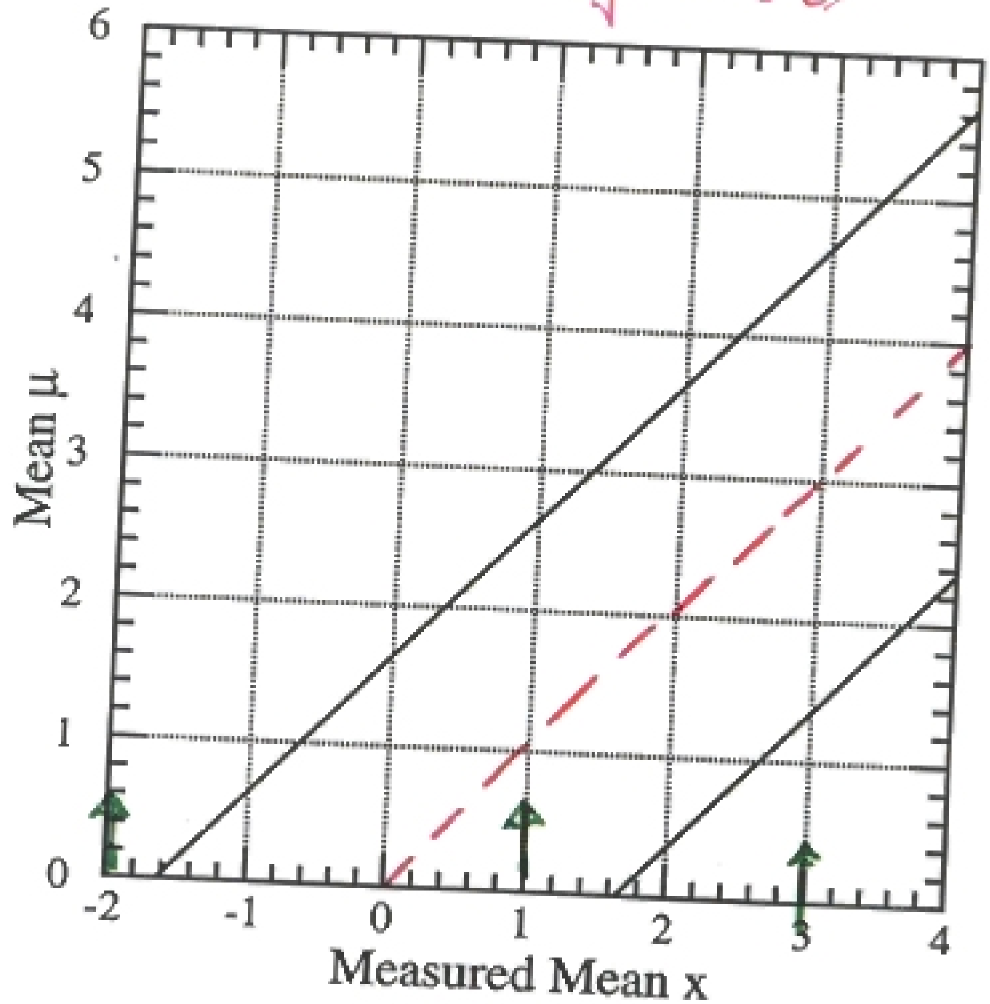


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

$x_{obs} = 3$ Two sided limit

$x_{obs} = 1$ Upper limit

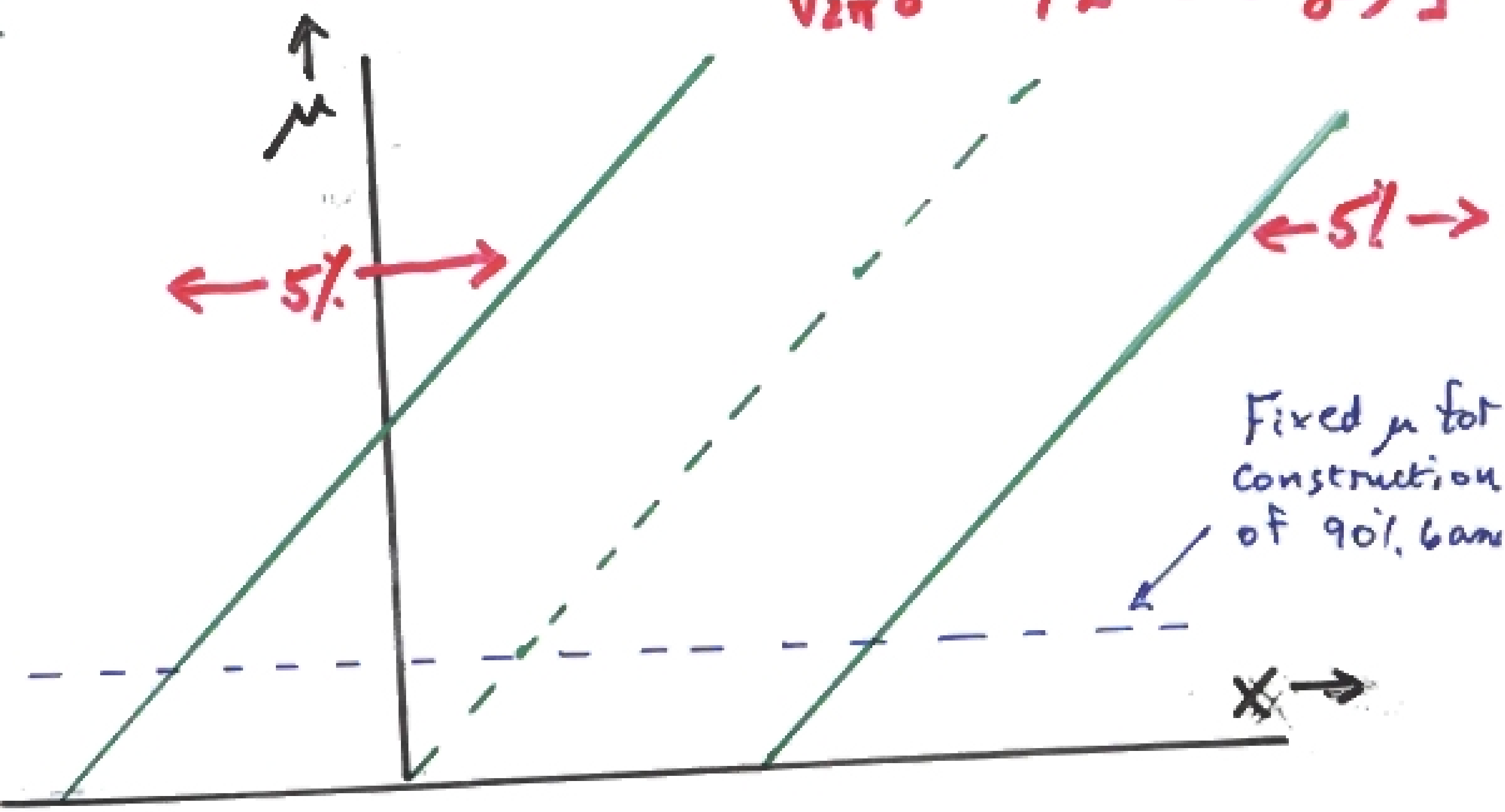
$x_{obs} = -2$ No region for μ

FELDMAN-COUSINS ORDERING RULE

$R = p(x,\mu)/p(x,\mu_{best})$ [Likelihood ratio ordering]

Gaussian example $p(x,\mu) = G(x,\mu,\sigma)$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

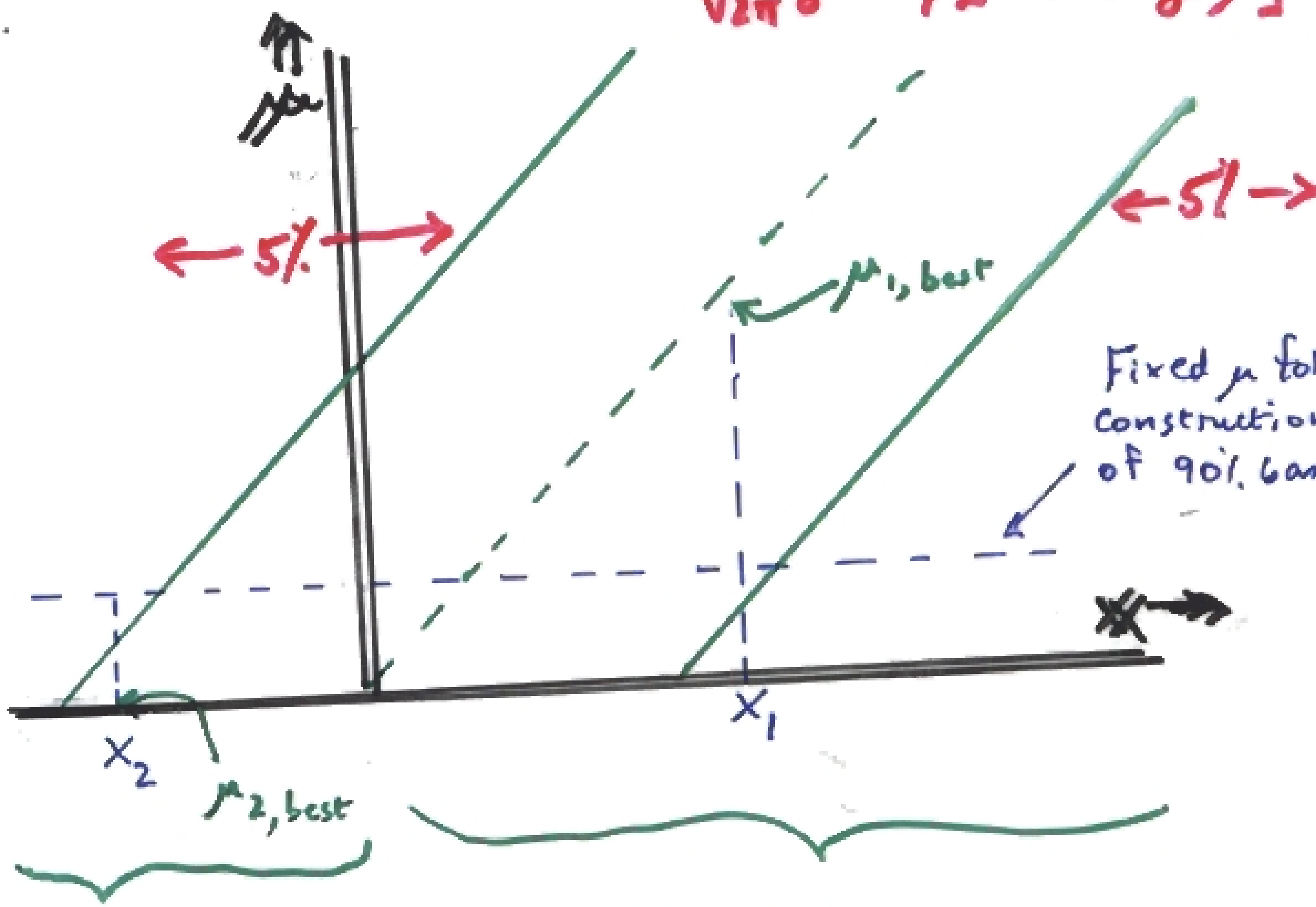


FELDMAN - COUSINS ORDERING RULE

$R = p(x,\mu) / p(x, \mu_{best})$ [Likelihood ratio ordering]

Gaussian example $p(x,\mu) = G(x,\mu,\sigma)$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$



$\mu_{best} = 0$

$p(x,\mu_b) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2}$

$\mu_{best} = x$

$p(x,\mu_b) = \frac{1}{\sqrt{2\pi}\,\sigma} = \text{const}$

$p(x_1,\mu) > p(x_2,\mu)$

Standard : Select x_1 before x_2

BUT $R(x_2,\mu) > R(x_1,\mu)$

F.C : Select x_2 before x_1

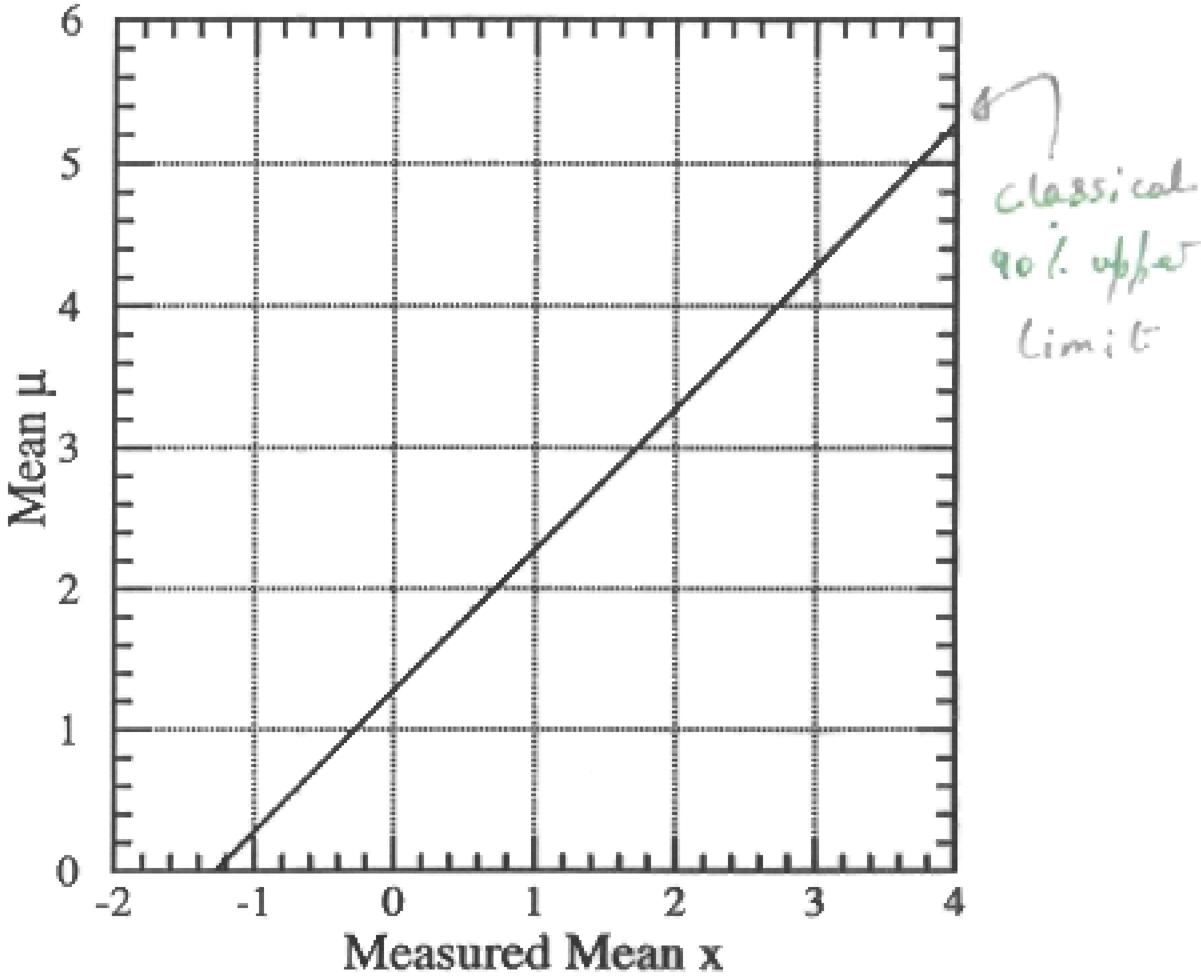


FIG. 2. Standard confidence belt for 90% C.L. upper limits for the mean of a Gaussian, in units of the rms deviation. The second line in the belt is at $x = +\infty$.

Feldman-Cousins 90% conf interval

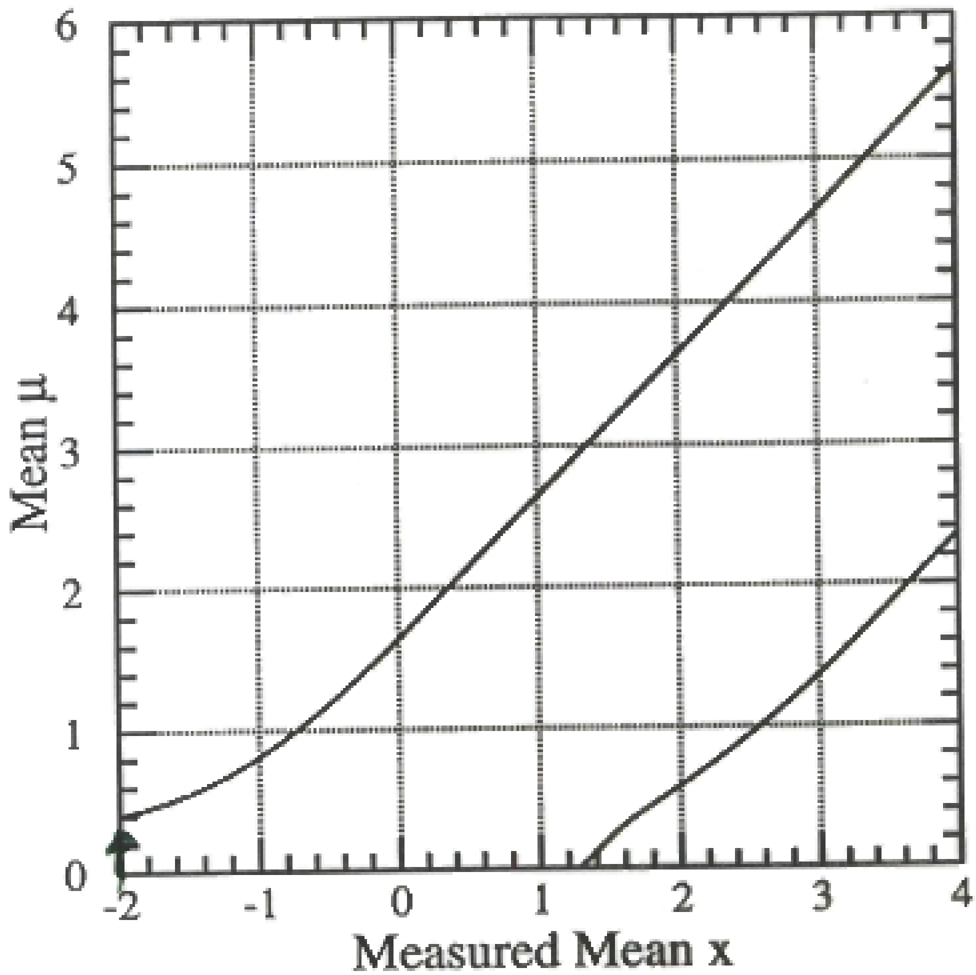


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

$x_{obs} = -2$ Now gives upper limit

FLIP - FLOP

90% upper limit for $x_{obs} \le 3$

90% 2-sided interval for $x_{obs} > 3$

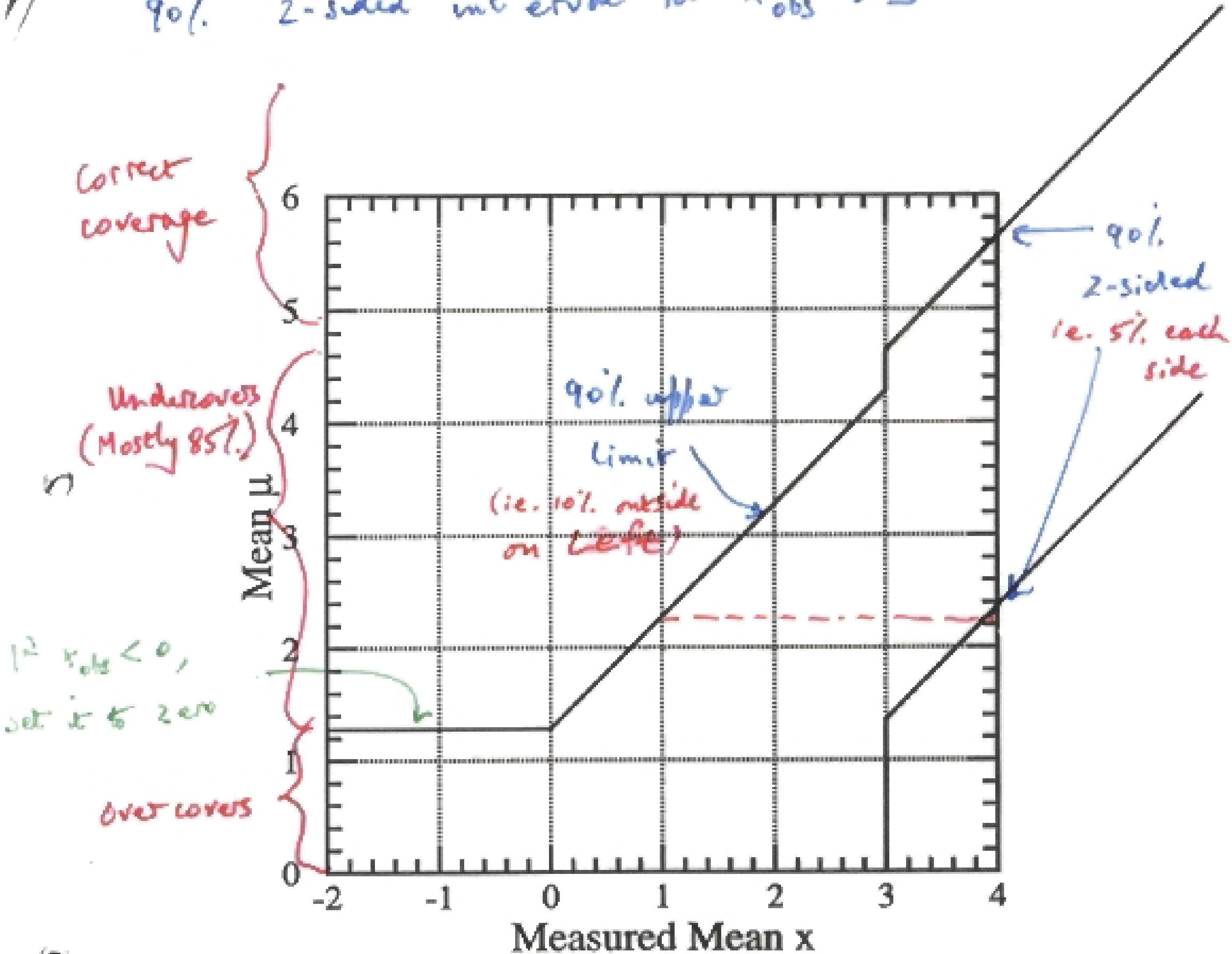


FIG. 4. Plot of confidence belts implicitly used for 90% C.L. confidence intervals (vertical intervals between the belts) quoted by flip-flopping Physicist X, described in the text. They are not valid confidence belts, since they can cover the true value at a frequency less than the stated confidence level. For $1.36 < \mu < 4.28$, the coverage (probability contained in the horizontal acceptance interval) is 85%.

Not good to let x_{obs} determine how result will be presented

F-C goes smoothly from 1-sided → 2-sided

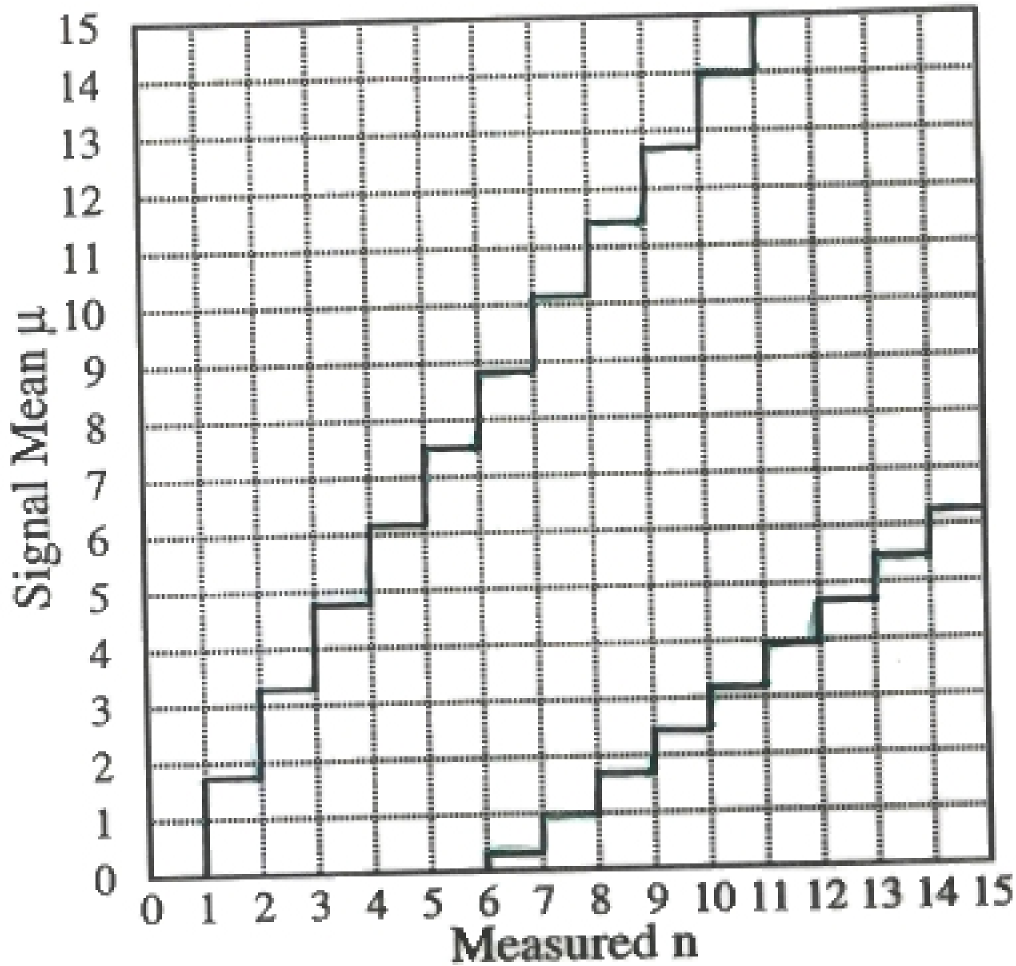


FIG. 6. Standard confidence belt for 90% C.L. central confidence intervals, for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

Standard Frequentist
for Poisson mean μ

FELDMAN & COUSINS FOR POISSON MEAN μ

90% conf

$b = 3.0$

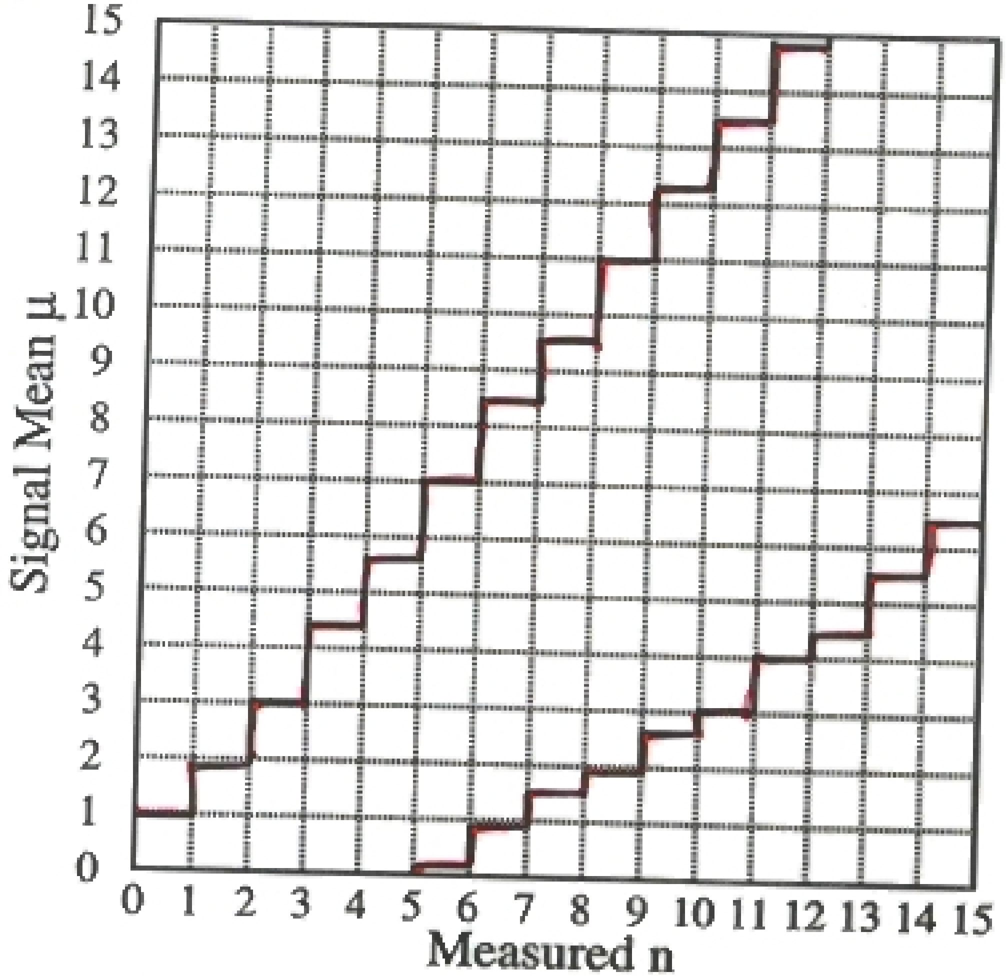


FIG. 7. Confidence belt based on our ordering principle, for 90% C.L. confidence intervals for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

FREQUENTIST POISSON C.B. CONSTRN.

< 10%

< 5%

TABLES

TABLE I. Illustrative calculations in the confidence belt construction for signal mean μ in the presence of known mean background $b = 3.0$. Here we find the acceptance interval for $\mu = 0.5$.

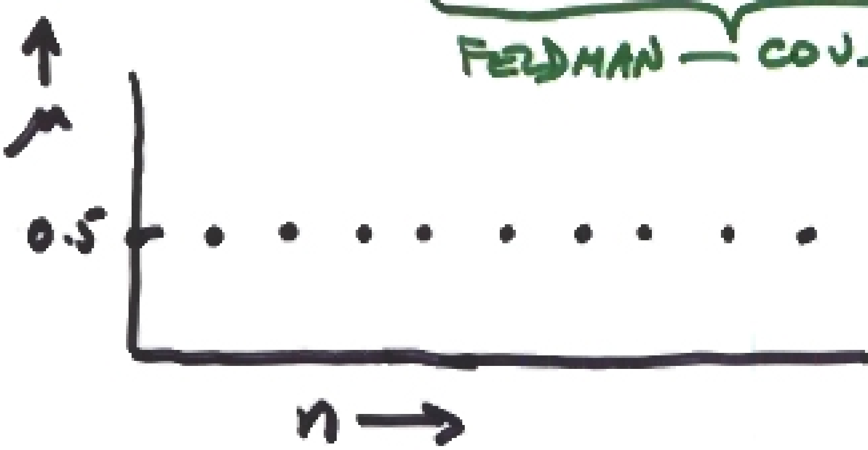
Prob ordering

n	$P(n\|\mu)$	$\mu_{\rm best}$	$P(n\|\mu_{\rm best})$	R	rank	U.L.	central
0	0.030	0.	0.050	0.607	6		
1	0.106	0.	0.149	0.708	5	✓	✓
2	0.185	0.	0.224	0.826	3	✓	✓
3	0.216	0.	0.224	0.963	2	✓	✓
4	0.189	1.	0.195	0.966	1	✓	✓
5	0.132	2.	0.175	0.753	4	✓	✓
6	0.077	3.	0.161	0.480	7	✓	✓
7	0.039	4.	0.149	0.259		✓	✓
8	0.017	5.	0.140	0.121		✓	✓
9	0.007	6.	0.132	0.050		✓	
10	0.002	7.	0.125	0.018		✓	
11	0.001	8	0.119	0.006		✓	

5
3
1
2
4
6
7

FELDMAN — COUSINS

< 5%



FEATURES OF F+C

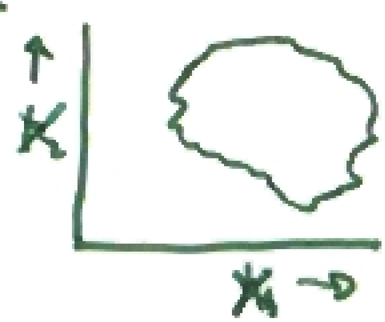
REDUCES EMPTY INTERVALS

UNIFIED 1-SIDED + 2-SIDED INTERVALS

ELIMINATES FLIP-FLOP

NO ARBITRARINESS OF INTERVAL

"READILY" EXTENDS TO SEVERAL DIMENSIONS



LESS OVERCOVERAGE THAN "5% AT ENDS"

MAX PROB DENSITY?

5% AT ENDS ?

NEYMAN CONSTRUCTION ⇒ CPU-INTENSIVE (ESP IN SEVERAL DIMENSIONS)

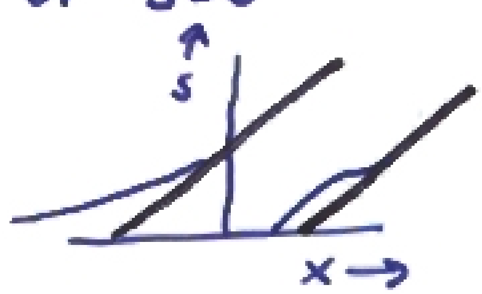
MINOR PATHOLOGIES : DISJOINT INTERVALS

WRONG BEHAVIOUR WRT BGD

TIGHT LIMITS FOR $b > n_{obs}$ e.g.

n_{obs}	bgd	90% Limit
0	3.0	1.08
0	0	2.44

UNIFIED ⇒ QUICKER EXCLUSION OF $s=0$

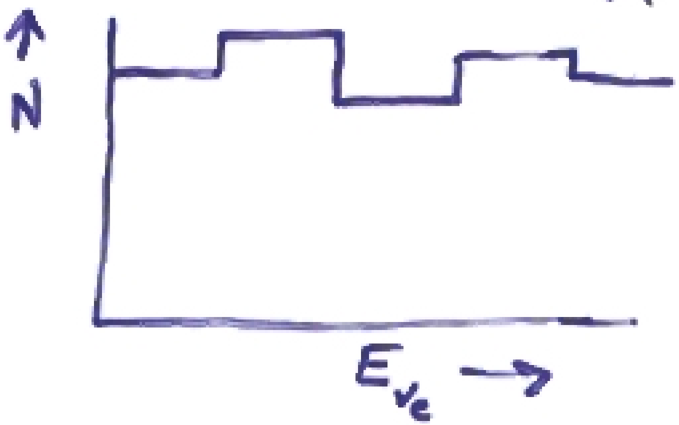


NEUTRINO OSCILLATIONS

$$P(\nu_\mu \to \nu_e) = \sin^2 2\theta \, \sin^2\left[\frac{1.27\,\Delta m^2 L}{E}\right]$$

eV² GeV km

Data = "ν_e" energy spectrum



Bgd = 100 ev/bin

Signal = 10,000 ev/bin if P = 1

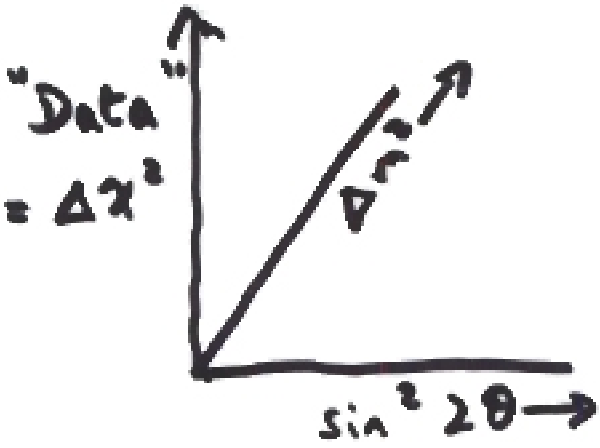
Compare data & prediction via

$$\Delta\chi^2 = \sum \left(\frac{n_i - b_i - \mu_i}{\sigma_i}\right)^2 - \left(\frac{n_i - b_i - (\mu_{best})_i}{\sigma_i}\right)^2$$

$$\text{OR} \quad 2\sum\left\{\mu_i - (\mu_{best})_i + n_i \frac{\mu_{best\,i} + b_i}{\mu_i + b_i}\right\}$$

(Ln [Likelihood ratio])

N.B. $\Delta\chi^2$ is more than just one piece of data



FIND ACCEPTANCE REGION FOR "DATA" BY M.C.

i.e. HOW BIG SHOULD $\Delta\chi^2_{cut}$ BE FOR 90% ACCEPTANCE?

[NOT STANDARD 4.61 for χ^2_2

BECAUSE a) EFFECT OF BOUNDARIES

b) WRONG OVERALL MINIMUM

c) POISSON $\not\equiv$ GAUSSIAN

d) $\Delta\chi^2 \not\equiv \chi^2_2$

e) '1-D REGIONS AT LOW Δm^2

$[P \sim \sin^2 2\theta . (\Delta m^2)^2]$

$\Delta\chi^2_{cut} = 2.4 - 6.6$

FINALLY, USE DATA $\Rightarrow$ $\Delta\chi^2$ AT EACH $(\sin^2 2\theta, \Delta m^2)$ + COMPARE WITH $\Delta\chi^2_{cut}(\sin^2 2\theta, \Delta m^2)$ TO FIND ACCEPTABLE REGION IN $(\sin^2 2\theta, \Delta m^2)$

VERY MUCH BETTER THAN "RASTER SCAN"

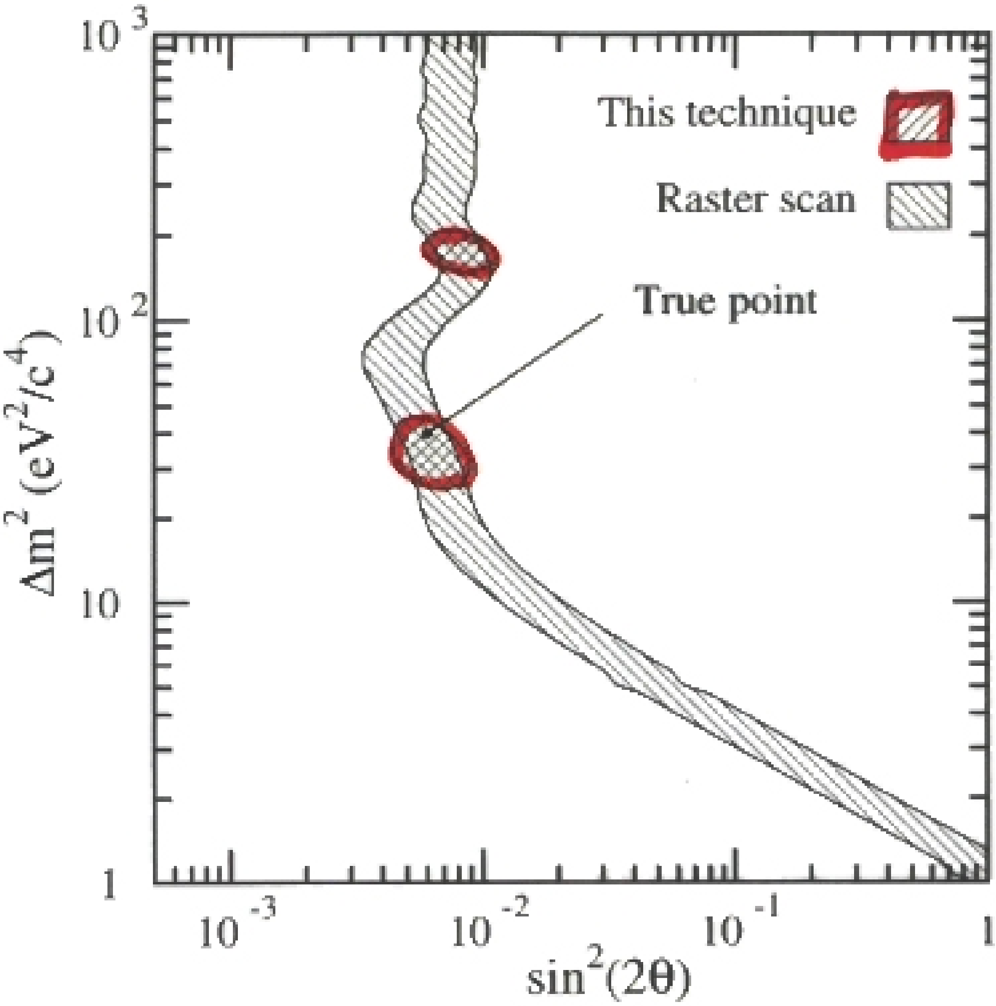


FIG. 12. Calculation of the confidence regions for an example of the toy model in which $\Delta m^2 = 40$ $(\mathrm{eV}/c^2)^2$ and $\sin^2(2\theta) = 0.006$, as evaluated by the proposed technique and the Raster Scan.

i.e. FELDMAN – COUSINS IS MUCH BETTER THAN RASTER SCAN

(cf $B-\bar{B}$ OSCILLATIONS)

SENSITIVITY

(indep of actual data)

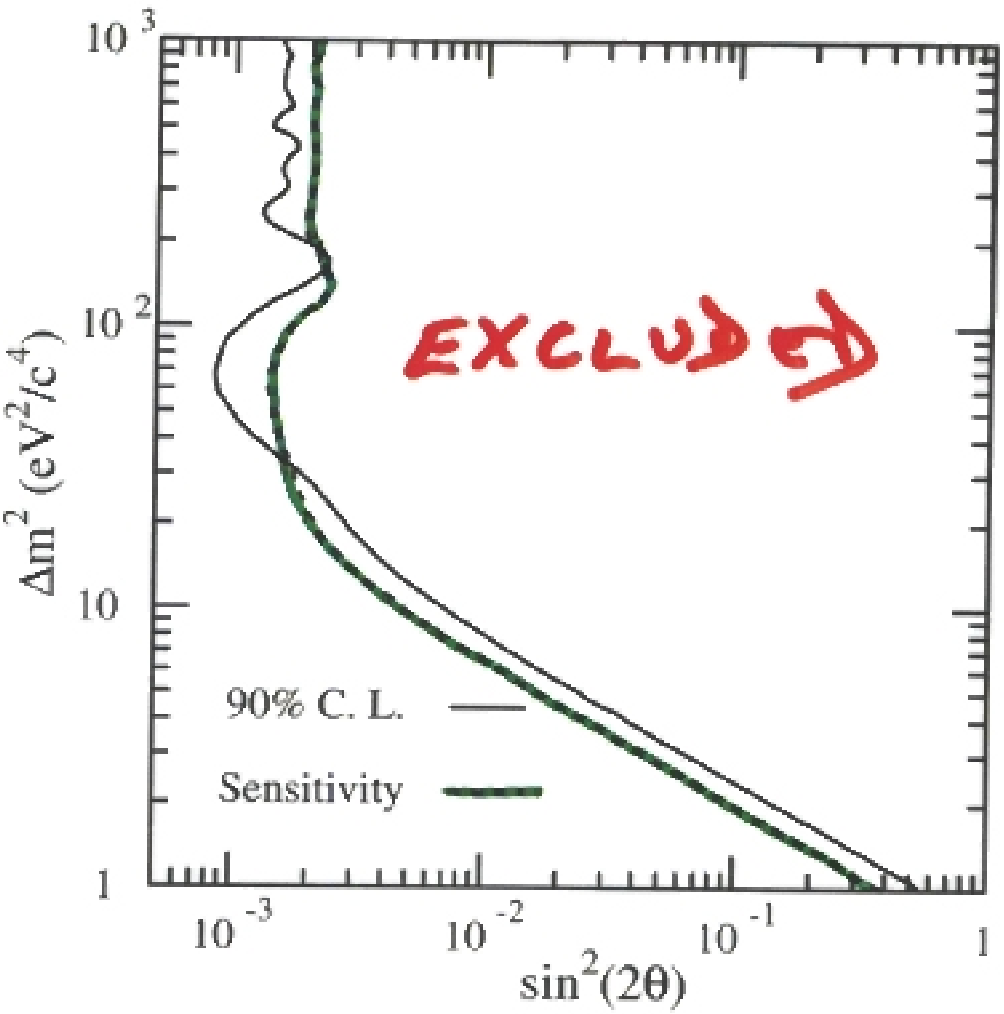
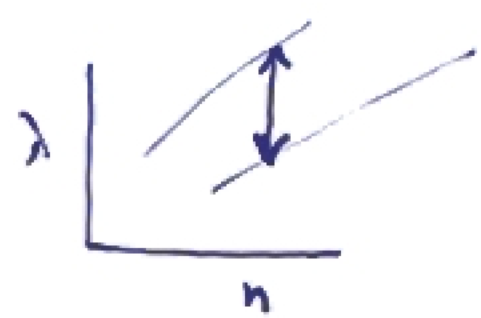
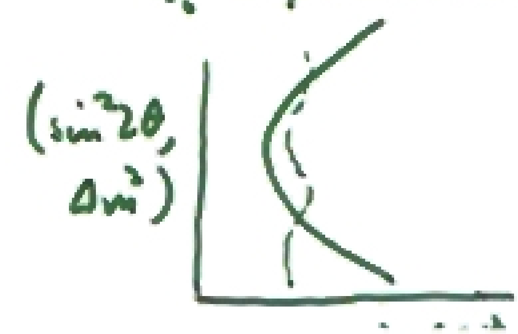


FIG. 15. Comparision of the confidence region for an example of the toy model in which $\sin^2(2\theta) = 0$ and the sensitivity of the experiment, as defined in the text.

	1-D POISSON	ν OSCILLATIONS
PARAM	λ	$\sin^2 2\theta$, Δm^2
DATA	n	$n_i\ (E_\nu)$
ORDERING RULE	$\ln[\mathcal{L}_{RATIO}]$	$\ln[\mathcal{L}_{RATIO}] \sim \Delta\chi^2$
HOW MANY?	At each λ, $\sum P(n\mid\lambda) = 0.9$	At each $(\sin^2 2\theta, \Delta m^2)$, include first 90% of $\Delta\chi^2$
ACCEPTANCE REGION	USE OBSERVED n $\Rightarrow \lambda$ RANGE	USE OBSERVED $n_i(E_\nu) \Rightarrow \Delta\chi^2(\sin^2 2\theta, \Delta m^2)$ $\Rightarrow (\sin^2 2\theta, \Delta m^2)$ REGION
	[REGION WHERE $\lambda \approx n$]	[REGION WHERE $n_i \approx$ prediction for $(\sin^2 2\theta, \Delta m^2)$]





SYSTEMATICS

For example

$$N_{events} = \sigma\, LA + b$$

Observed ↑ $N \pm \sqrt{N}$ for statistical errors

Physics parameter

we need to know these, probably from other measurements (and/or theory)

Uncertainties → error in σ

Some are arguably statistical errors

$$LA = LA_{\circ} \pm \sigma_{LA}$$

$$b = b_{\circ} \pm \sigma_{b}$$

Shift Central Value

Bayesian

Frequentist

Mixed

Profile Likelihood

$$N_{events} = \sigma\, LA + b$$

Simplest Method

Evaluate σ_0 using LA_0 and b_0

Move nuisance parameters (one at a time) by their errors → $\delta\sigma_{LA}$ & $\delta\sigma_{b}$

If nuisance parameters are uncorrelated

Combine these contributions in quadrature

→ total systematic

PROFILE $\mathcal{L}$

Rolke, Lopez, Conrad + James

"Limits + Confidence Intervals in the presence of Nuisance Parameters"

$$p\mathcal{L}(\mu|\text{data}) = \mathcal{L}(\mu, b_{best}|\text{data})$$

$$\Delta \ln p\mathcal{L} = 0.5$$

Coverage much smoother (as fn of μ) than for standard Bayesian without nuisance parameters

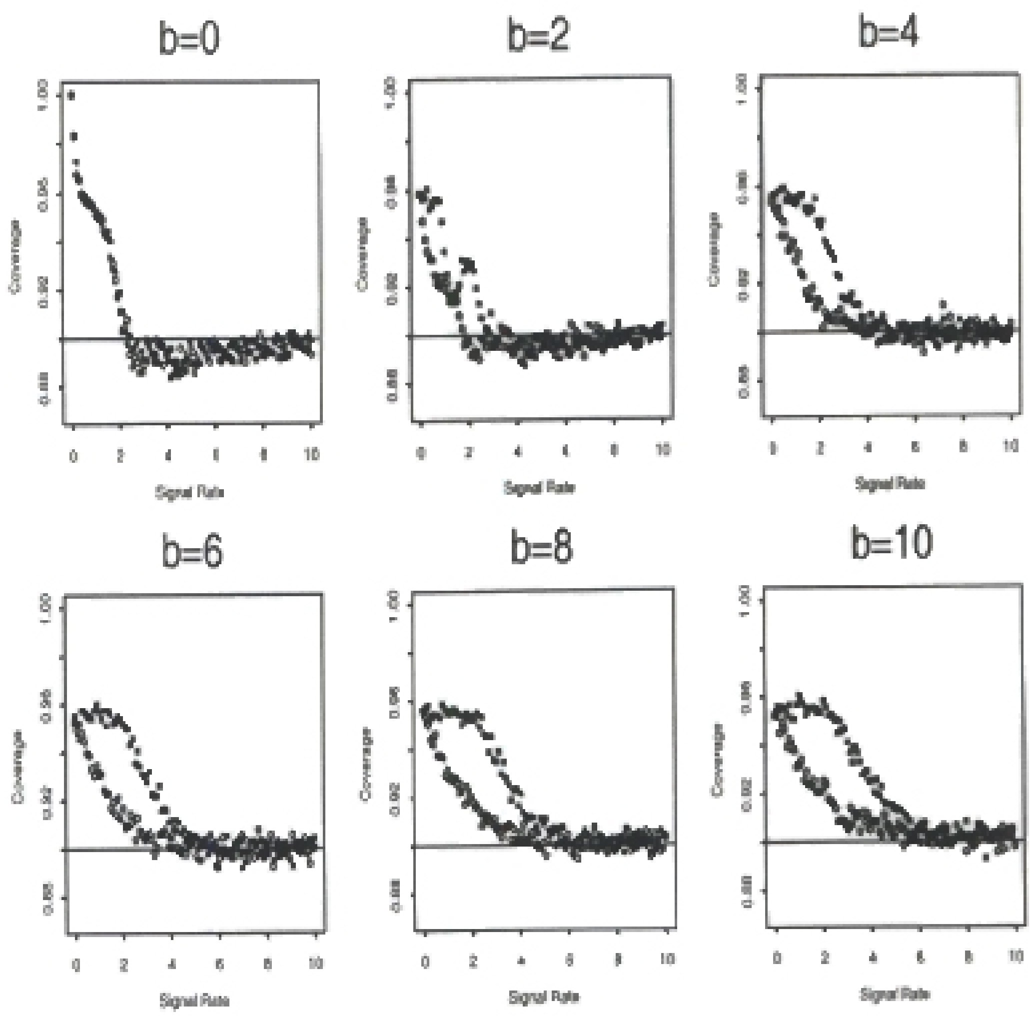


FIG. 3: 90% coverage graphs when the signal and the background are modeled as Poisson and the efficiency is modeled as a Binomial. We have $\tau = 3.5$, $\epsilon = 0.85$ and $m = 100$. The empty circles show the coverage using the unbounded likelihood method and the solid squares show the coverage using the bounded likelihood method.

Rolke et al

Profile $\mathcal{L}$

Bayesian

Without systematics $$p(\sigma;N) \propto p(N;\sigma)\Pi(\sigma)$$

prior

With systematics

$$p(\sigma, LA, b; N) \propto p(N;\sigma, LA, b)\Pi(\sigma, LA, b)$$

$$\sim \Pi_1(\sigma)\Pi_2(LA)\Pi_3(b)$$

Then integrate over LA and b

$$p(\sigma;N) = \iint p(\sigma, LA, b; N)\, dLA\, db$$

$$p(\sigma; N) = \iint p(\sigma, LA, b; N)\, dLA\, db$$

If $\Pi_1(\sigma)$ = constant and $\Pi_2(\mathrm{LA})$ = truncated Gaussian TROUBLE!

Upper limit on σ from $\int p(\sigma, N)\, \mathrm{d}\sigma$

Significance from likelihood ratio for $\sigma = 0$ and σ_{max}

BAYES 90% UPPER LIMITS

	$\epsilon = 1.0 \pm 0.1$		$\epsilon = 1$ exactly	
Bgd	0	3	0	3
n_{obs}				
0	2.35 indep of b		2.30 indep of b	
1	3.99	2.90	3.89	2.84
2	5.47	3.60	5.32	3.52
3	6.87	4.46	6.68	4.36
4	8.24	5.48	7.99	5.34
⋮	⋮	⋮	⋮	⋮
20	28.3	25.04	27.05	24.04

Less than 10% bigger than for $\epsilon = 1$ exactly

$\Delta = 0$ for $b = 0$

$\Delta = 3$ for large b

$\sim n + k\sqrt{n}$

JOEL HEINRICH et al (CDF St. Cttee)

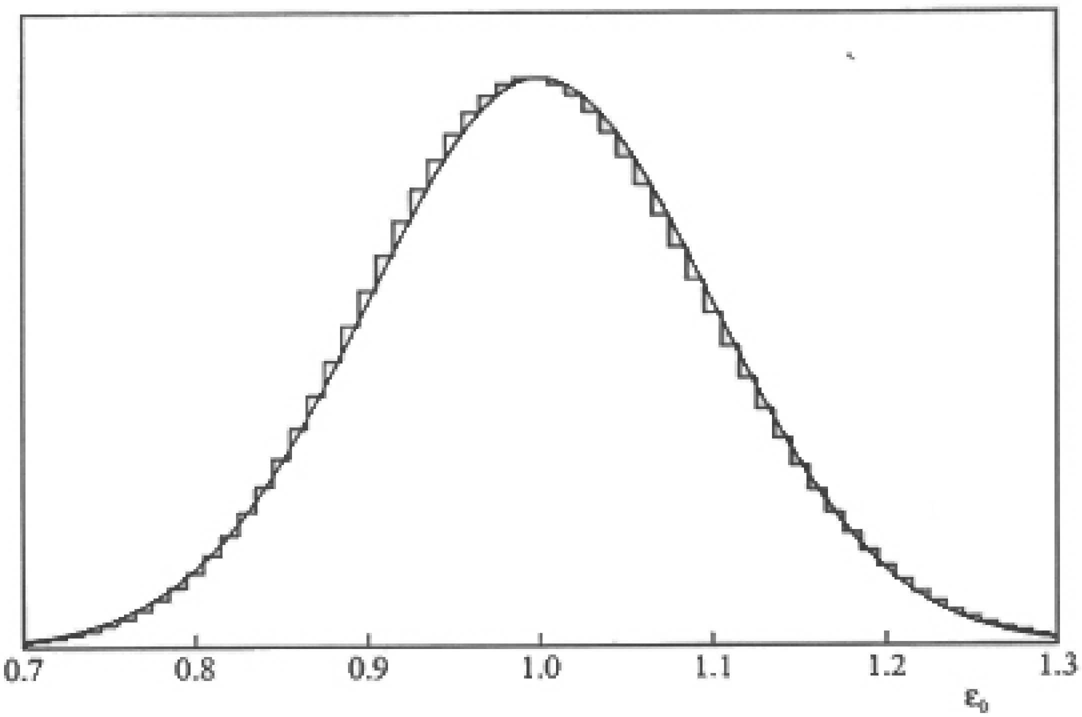


Figure 2: Comparison of our discrete probability for ϵ_0 (shown as a histogram, see eqn (11)) and Gaussian (continuous curve) for the case $\epsilon = 1 \pm 0.1$.

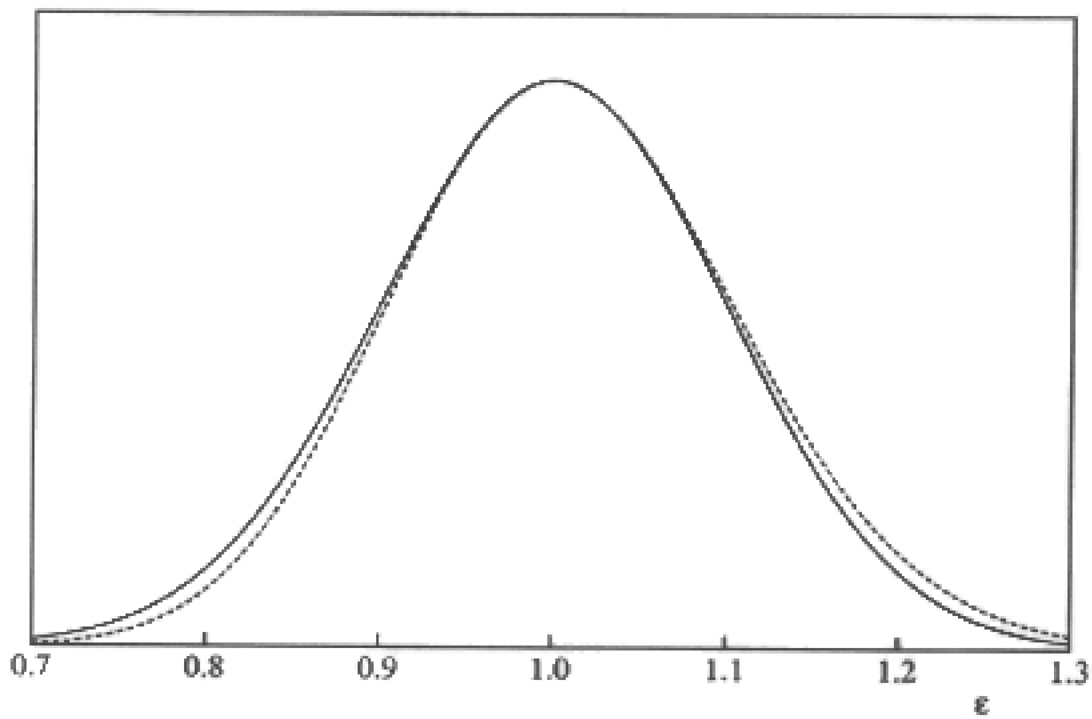


Figure 3: Comparison of our likelihood (dashed, see eqn (12)) and Gaussian (solid) for the case $\epsilon = 1 \pm 0.1$.

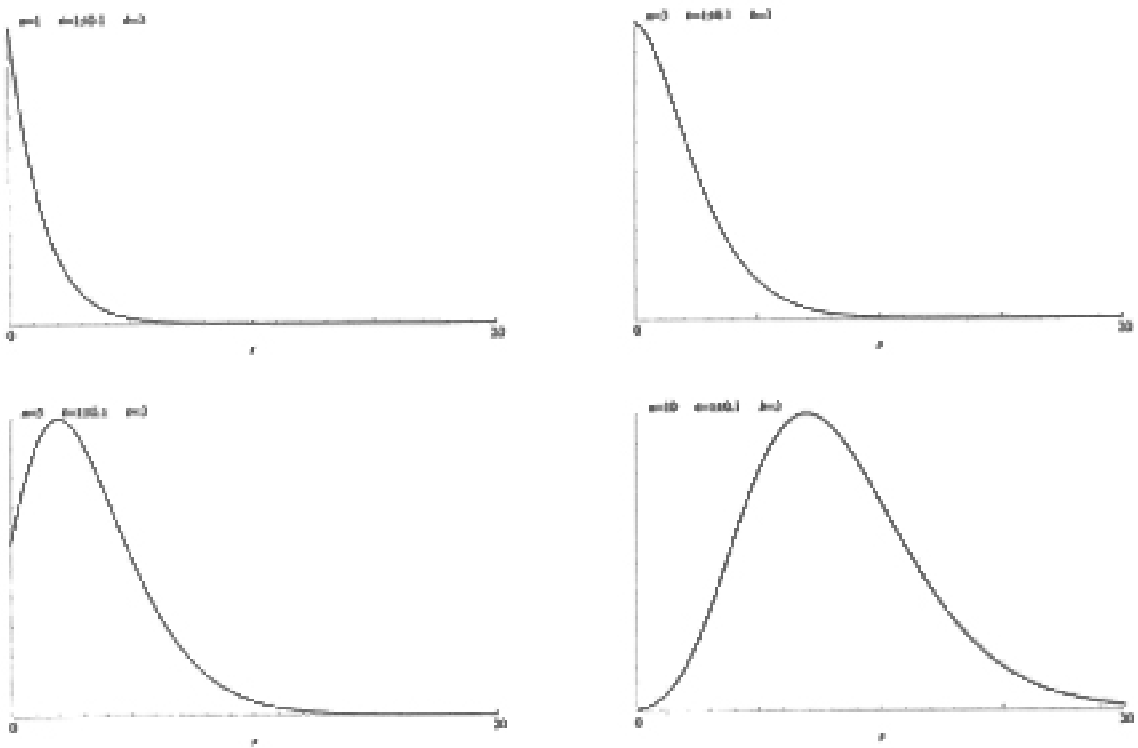


Figure 4: Posterior densities $p(s|n, b)$ vs s for $n = 1$, 3, 5, 10. In each case, $b = 3$ and $\epsilon = 1 \pm 0.1$ (i.e. $\kappa = 100$ and m=99).

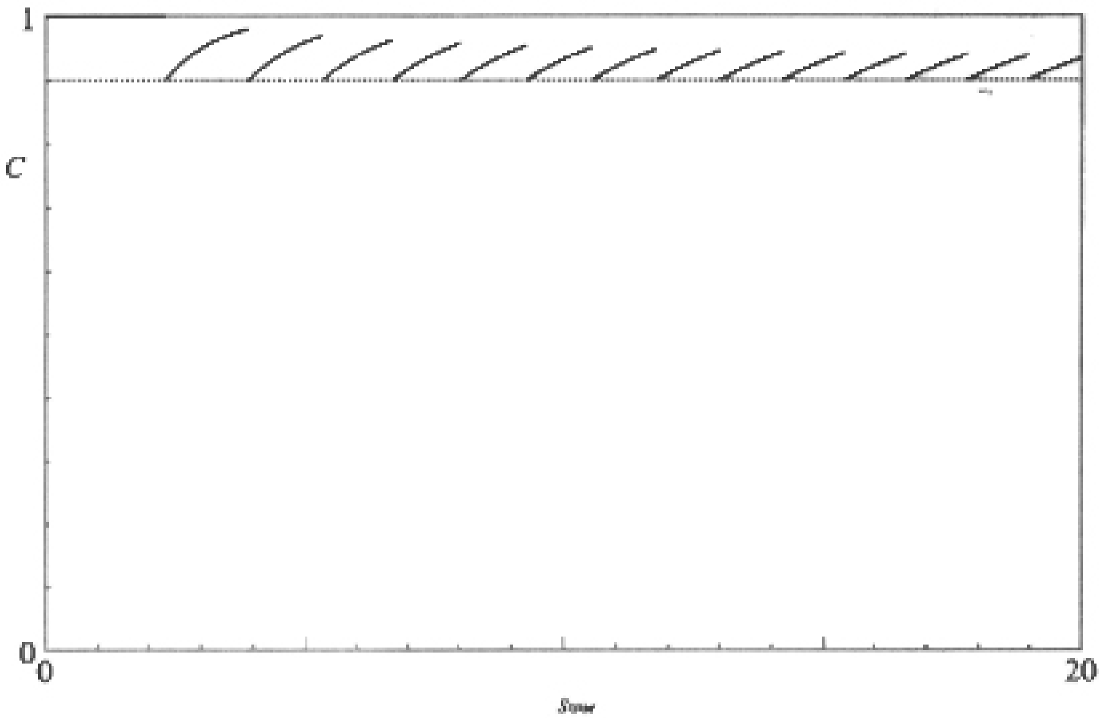


Figure 1: Coverage as a function of the true signal rate s for Bayes 90% limits, for the simple case of no background and no uncertainty on $\epsilon = 1$.

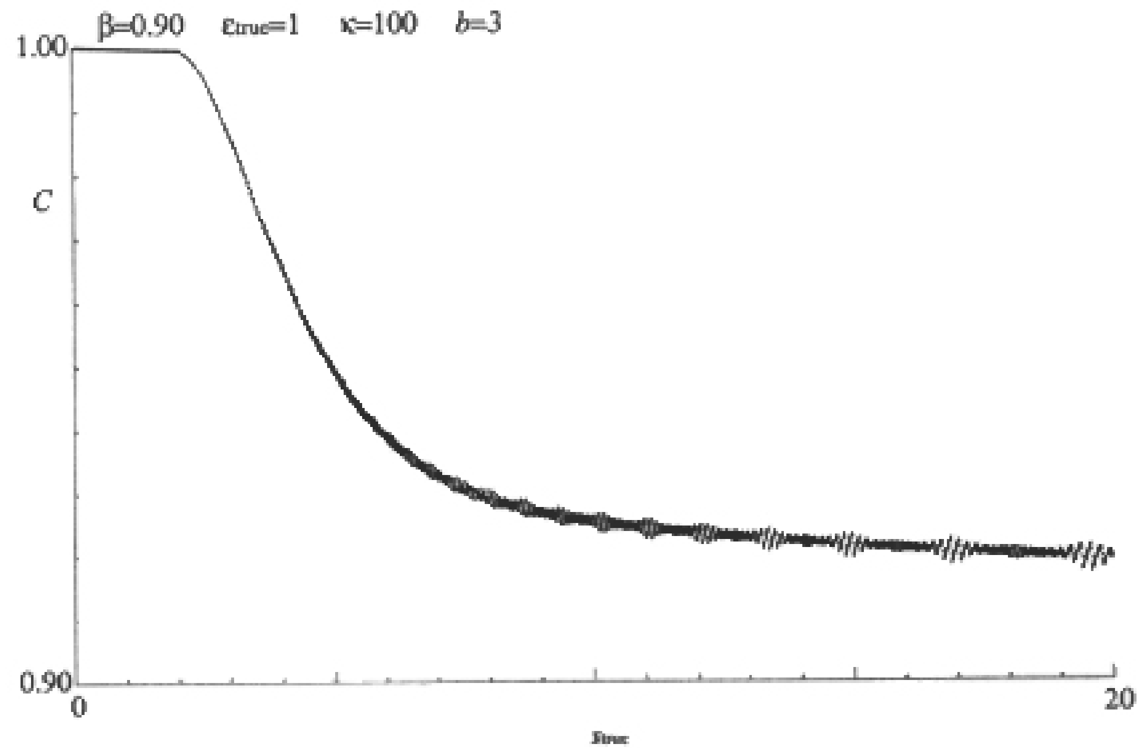


Figure 5: Coverage of 90% upper limits as a function of $s_{\rm true}$ for $\epsilon_{\rm true} = 1$, nominal 10% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

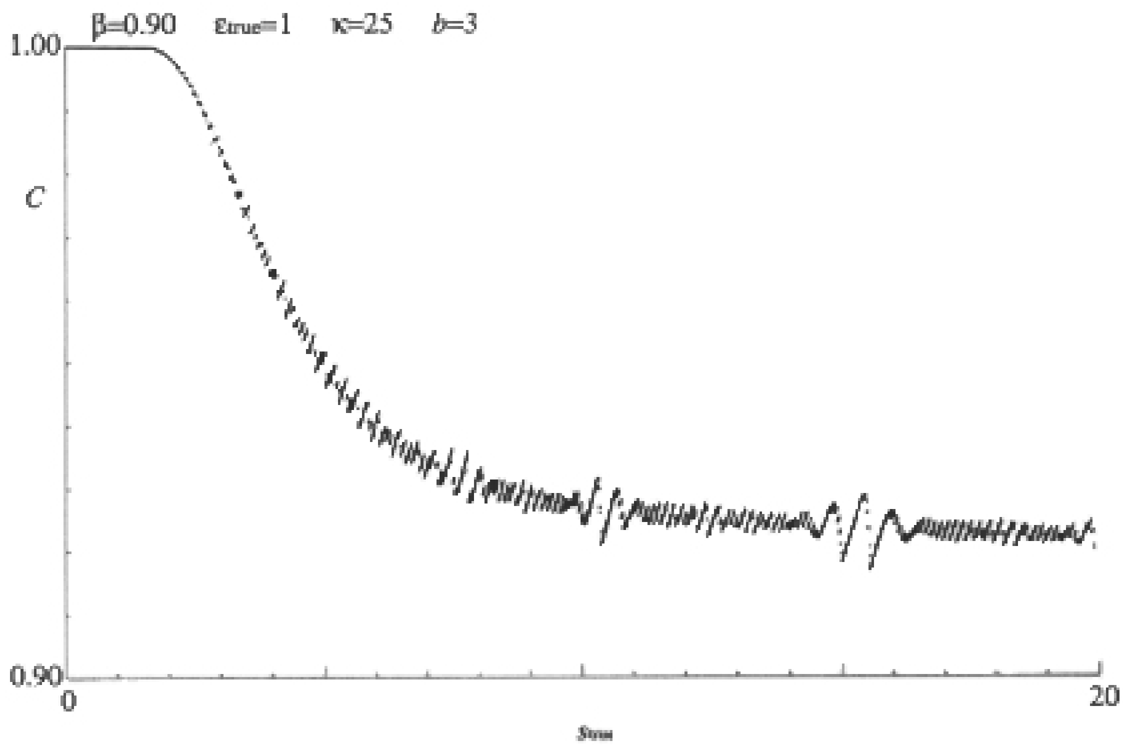


Figure 6: Coverage of 90% upper limits as a function of s_{true} for $\epsilon_{\text{true}} = 1$, nominal 20% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

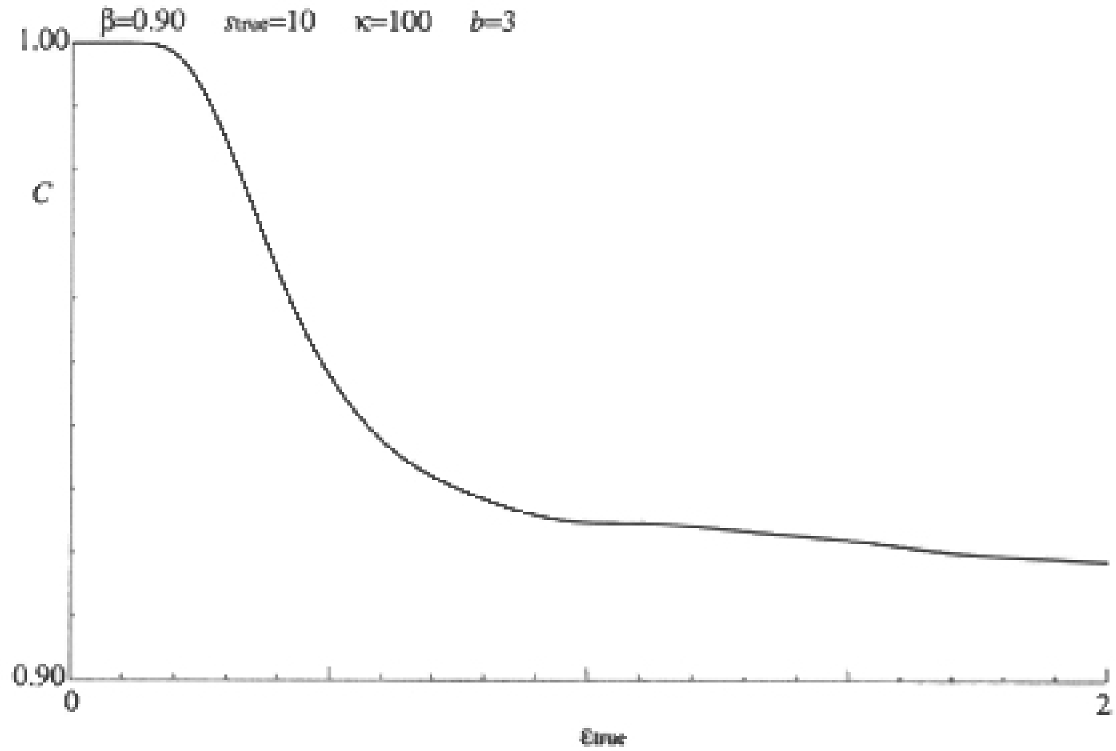


Figure 7: Coverage of 90% upper limits as a function of ϵ_{true} for $s_{\text{true}} = 10$, nominal 10% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

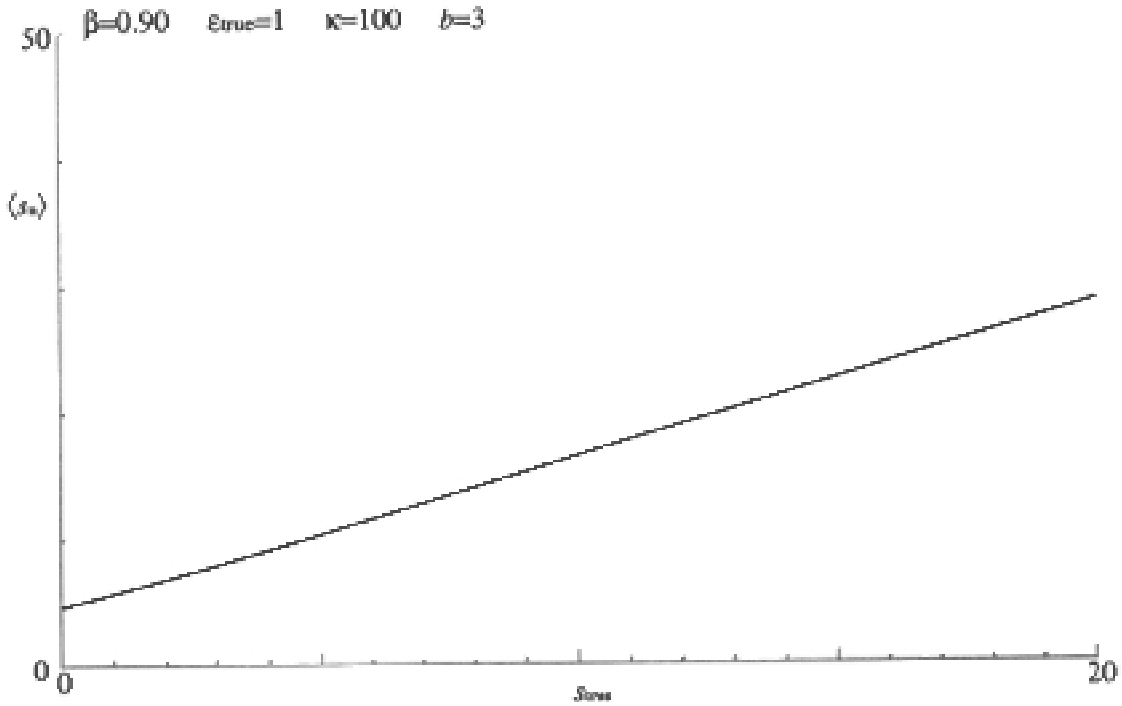


Figure 8: Sensitivity of 90% upper limits as a function of $s_{\rm true}$ for $\epsilon_{\rm true} = 1$, nominal 10% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

Frequentist

Full Method

Imagine just 2 parameters σ and LA

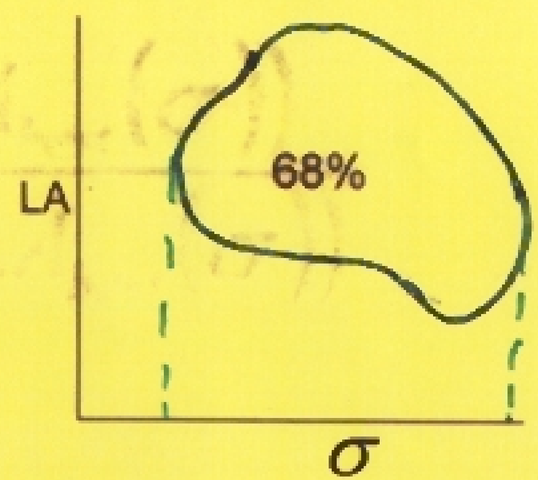
and 2 measurements N and M

$\uparrow$ Physics (N) $\quad$ $\uparrow$ Nuisance (M)

Do Neyman construction in 4-D

Use observed **N** and **M**, to give

Confidence Region



Then project onto σ axis

This results in OVERCOVERAGE

Aim to get better shaped region, by suitable choice of ordering rule

Example: Profile likelihood ordering

$$\frac{L\left(N_0 M_0; \sigma, LA_{best}(\sigma)\right)}{L\left(N_0 M_0; \sigma_{best}, LA_{best}(\sigma)\right)}$$

Full frequentist method hard to apply in several dimensions

Used in ≤ 3 parameters

For example: Neutrino oscillations (CHOOZ)

$$\sin^2 2\theta \, , \Delta m^2$$

Normalisation of data

Use approximate frequentist methods that reduce dimensions to just physics parameters

e.g. <u>Profile</u> pdf

i.e. $pdf_{profile}(N;\sigma) = pdf(N, M_0; \sigma, LA_{best})$

Contrast Bayes marginalisation

Distinguish "profile <u>ordering</u>"

Properties being studied by Giovanni Punzi

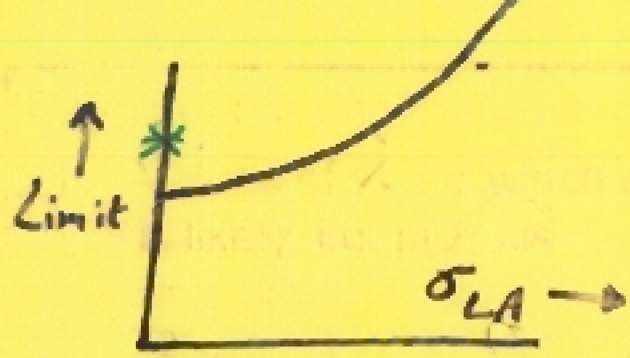
Method: <u>Mixed Frequentist - Bayesian</u>

Bayesian for nuisance parameters and

Frequentist to extract range

Philosophical/aesthetic problems?

Highland and Cousins



(Motivation was paradoxical behavior of Poisson limit when LA not known exactly)

Coverage studied by Tegenfeldt + Conrad

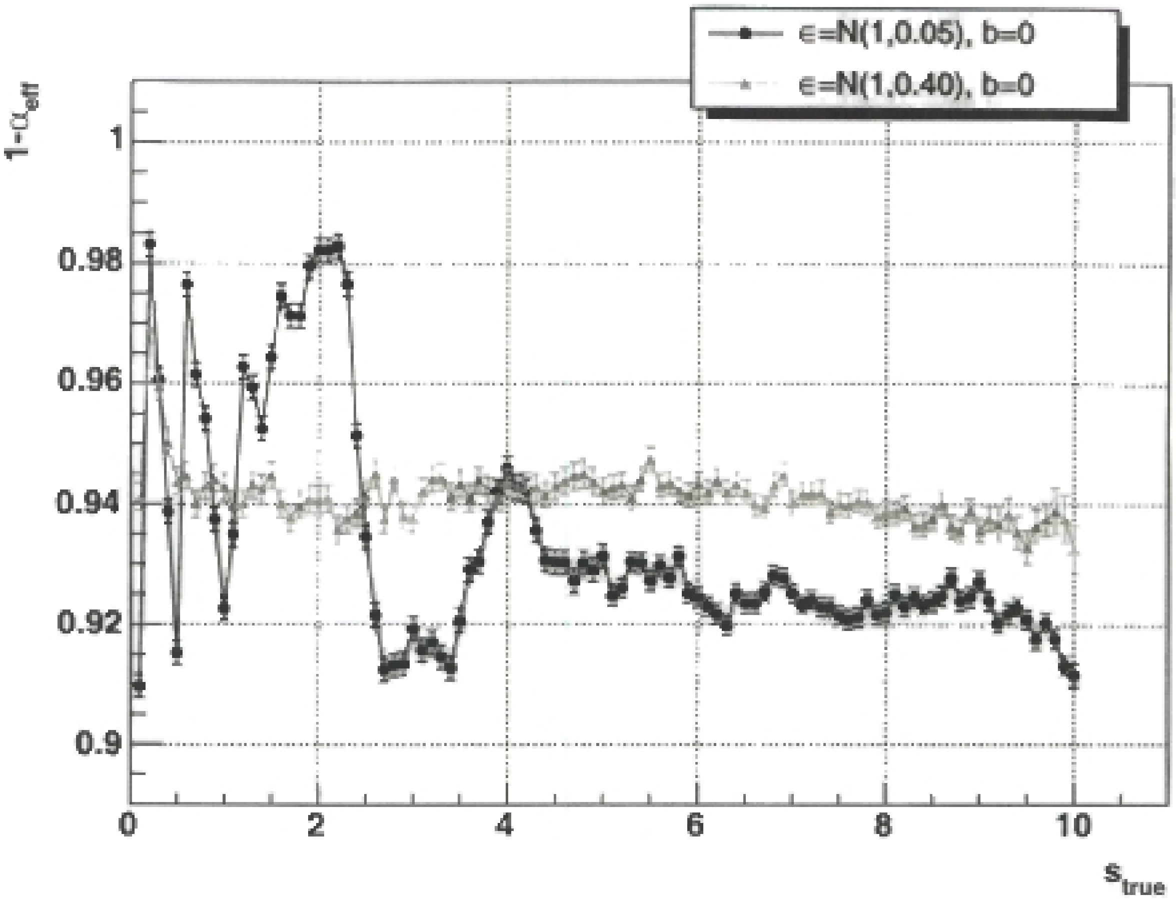


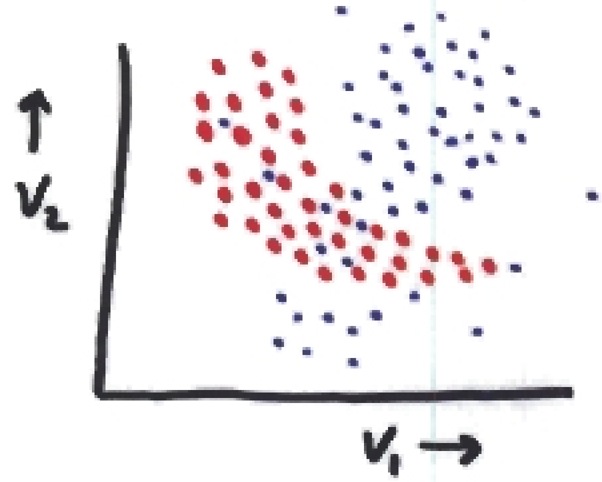
Fig. 1. Calculated coverage as function of signal hypothesis. Two case are shown: 5 % and 40 % Gaussian uncertainties in the signal efficiency. The nominal coverage was 90%.

Tegenfeldt + Conrad ✓

Feldman - Cousins + Bayes ✓

MULTIVARIATE ANALYSIS

AIM TO SEPARATE SIGNAL FROM BACKGROUND



NEYMAN - PEARSON THEOREM

1) IMAGINE ALL POSSIBLE CONTOURS THAT SELECT SIGNAL WITH α% EFFICIENCY
(LOSS = ERROR OF 1st KIND)

2) BEST IS ONE CONTAINING MINIMAL AMOUNT OF BGD (CONTAM = ERROR OF 2nd KIND

EQUIVALENT TO ORDERING DATA BY

$$\mathcal{L}\text{-RATIO} = \frac{\mathcal{L}_S(v_1, v_2 \dots)}{\mathcal{L}_B(v_1, v_2 \dots)}$$

IF VARIABLES INDEPENDENT,

$$\rightarrow \frac{\mathcal{L}_S(v_1)}{\mathcal{L}_S(v_1)} \times \frac{\mathcal{L}_S(v_2)}{\mathcal{L}_B(v_2)} \times \dots\dots$$

PROBLEM: DON'T KNOW $\mathcal{L}$ - RATIO EXACTLY BECAUSE :-

1) GENERATED BY M.C. WITH FINITE STATISTICS

2) UNCERTAIN PARAMETERS (NUISANCE PARAMS, SYSTEMATICS)

3) NEGLECTED SOURCES OF BGD

4) HARD TO IMPLEMENT N-P IN MANY DIMENSIONS

METHODS : CUTS

FISHER DISCRIMINANT

PRINCIPAL COMPONENT ANALYSIS

INDEPENDENT COMP. ANALYSIS

BOOSTED TREE METHODS

KERNEL DENSITY ESTIMATION

NEURAL NETS ✡ ✡

SUPPORT VECTOR MACHINES

⋮

USEFUL REFERENCES

H. PROSPER : 'MULTIVARIATE ANALYSIS' (DURHAM)

J. FRIEDMAN : 'PREDICTIVE MACHINE LEARNING' (PHYSTAT 2003)

R. BOCK : 'MULTIDIM. EVENT CLASSIFICATION FOR γ RAY SHOWERS (DURHAM)

FNAL AAG http://projects.fnal.gov/run2aag/

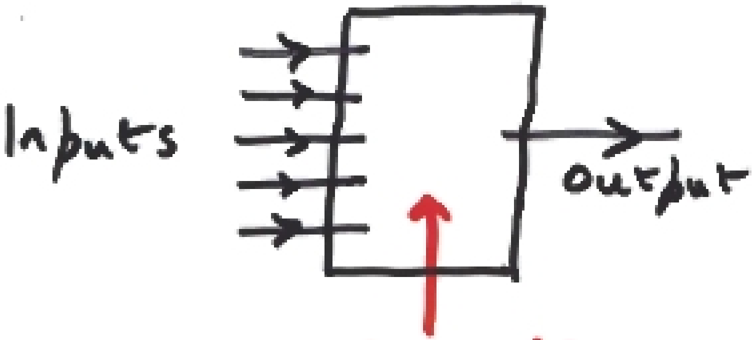
S. TOWERS : i) PROB DENSITY ESTIMATION.
ii) REDUCE NUMBER OF VARIABLES
(BOTH AT DURHAM)

A VAICIULIS : SUPPORT VECTOR MACHINES (DURHAM)

NEURAL NETWORKS

TYPICAL APPLICATION:

CLASSIFY EVENTS AS $\begin{cases} \text{SIGNAL} \\ \text{BACKGROUND} \end{cases}$



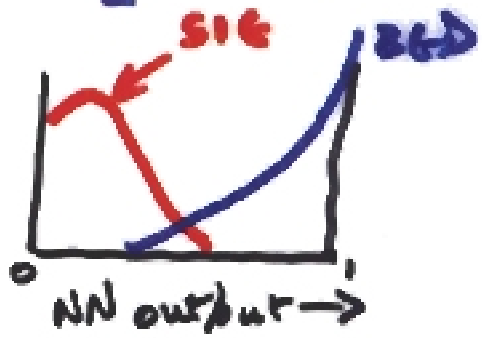
$\begin{cases} \sim 1 \text{ FOR SIGNAL} \\ \sim 0 \text{ FOR BGD} \end{cases}$

Adjustable params (WEIGHTS + THRESHOLDS)

1) LEARNING PROCESS:

INPUT = $\begin{cases} \text{KNOWN SIGNAL} \\ \text{KNOWN BGD} \end{cases}$ M.C?

ADJUST PARAMS ⇒ 'BEST' OUTPUT



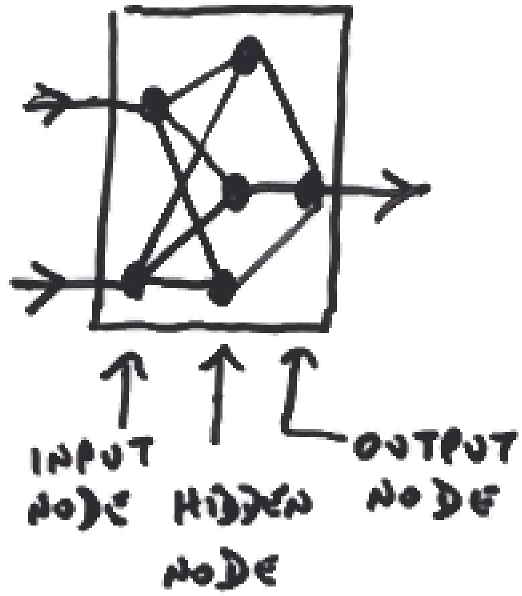
2) TESTING PROCESS

MAKE SURE NO 'OVERTRAINING'

3) USE TRAINED NET ON ACTUAL DATA.

CLASSIFY AS SIGNAL IF NN OUTPUT > Cut

HOW DOES IT WORK?



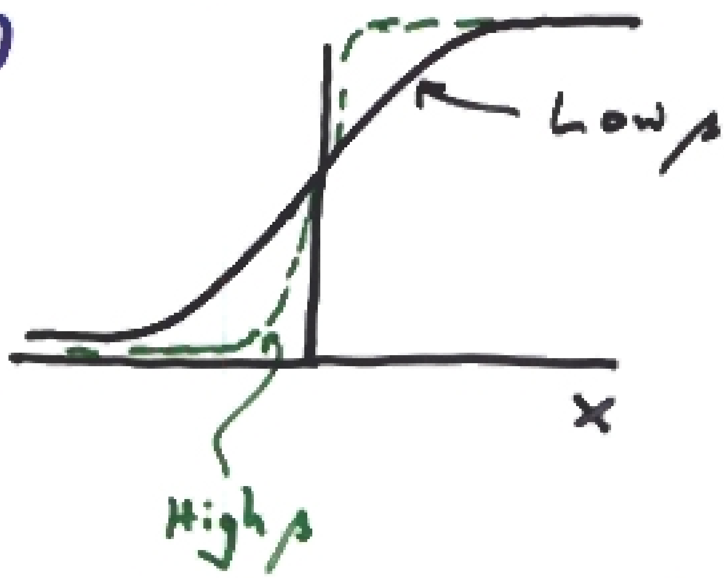
FOR EACH NODE

$$\text{Output} = F\left[\sum \text{Input}_i \times W_i + T\right]$$

PARAMS

Typical F(x)

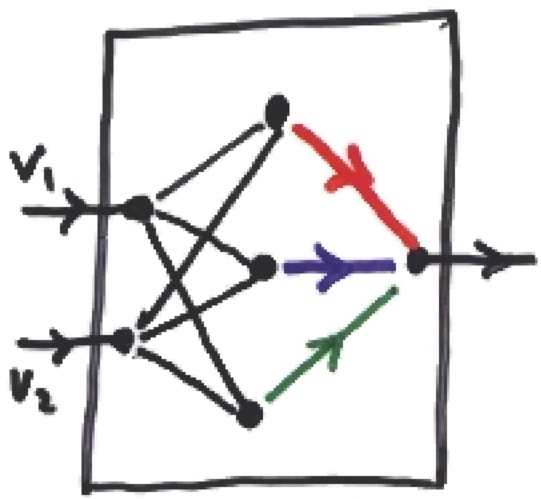
$$\frac{1}{1+e^{-\beta x}}$$

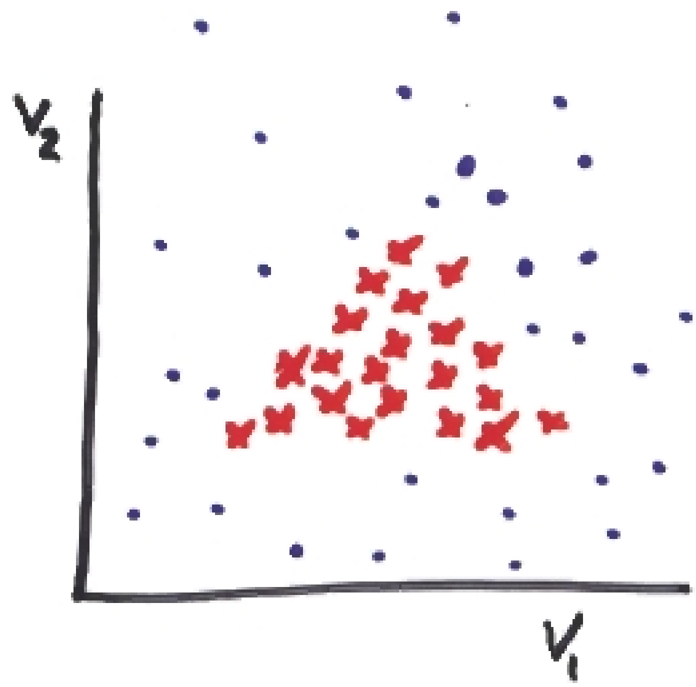


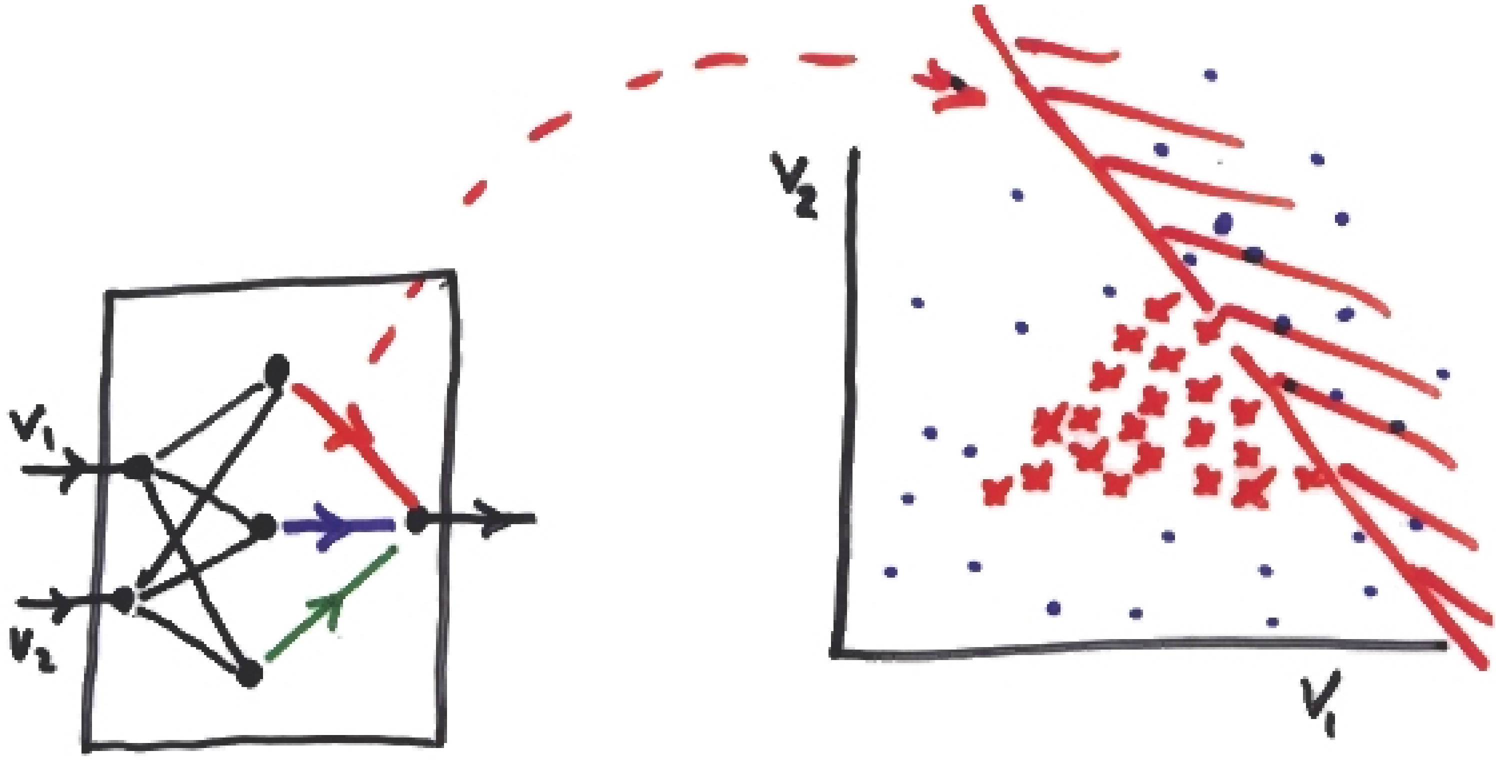
$\Rightarrow$ Output of node in "ON"

if $\sum I_i W_i + T > 0$

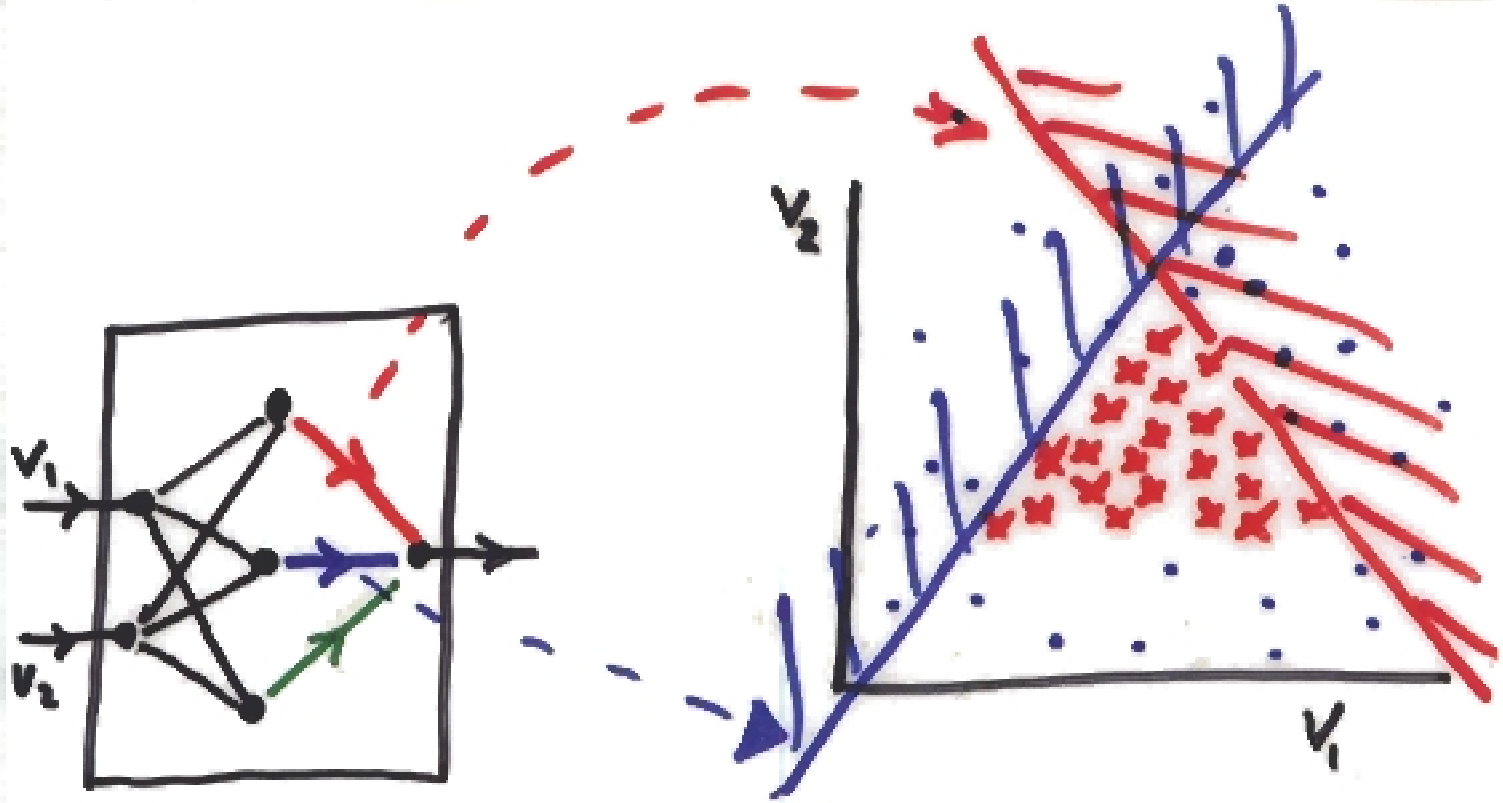
This is "hyper-plane" in I. space


V_1
V_2

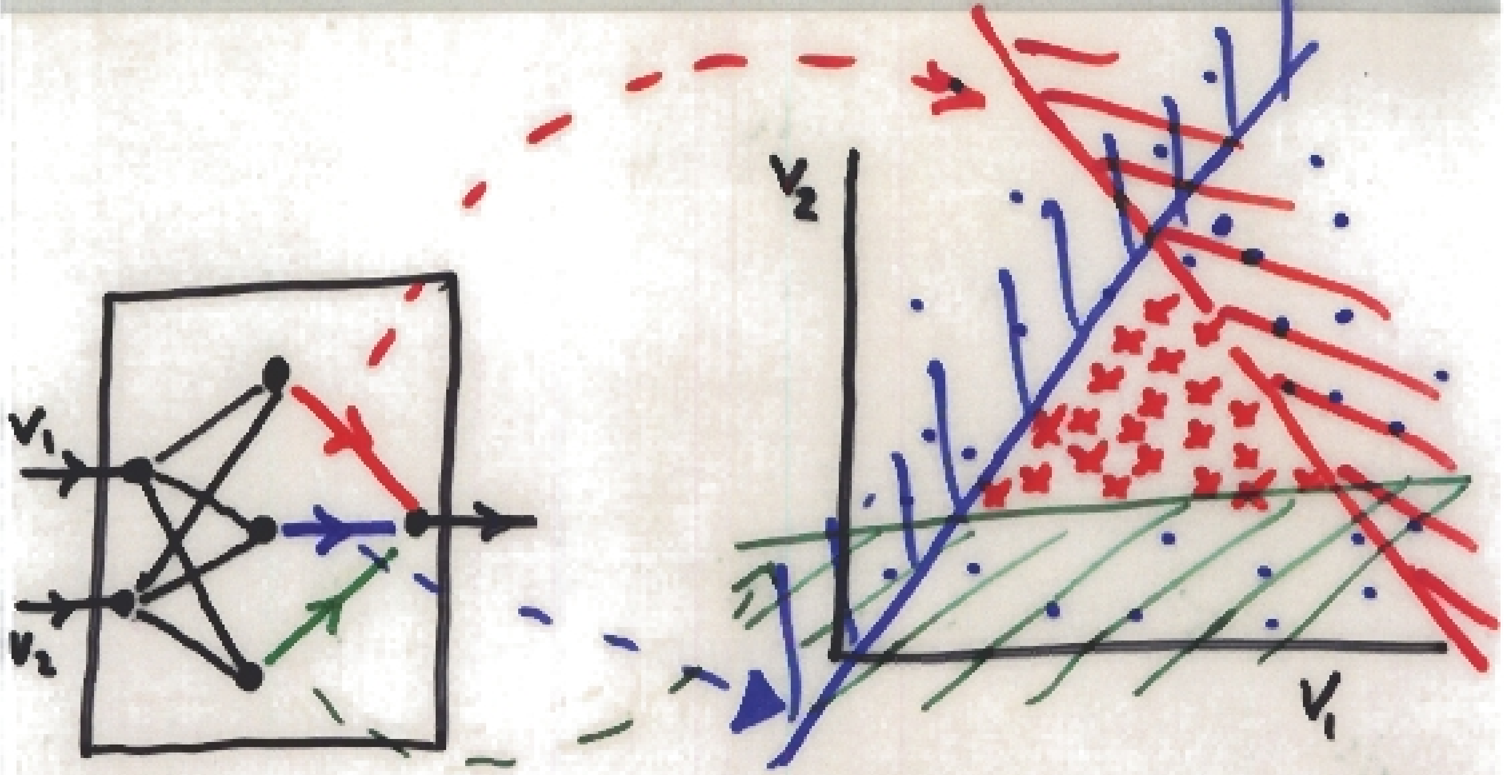

V_2
V_1



V_1
V_2
V_2
V_1



V_1
V_2
V_2
V_1



$$\text{Output} = F\left[0.4H_1 + 0.4H_2 + 0.4H_3 - 1.0\right]$$

Output = "ON" only if H_1 H_2 H_3 all are "ON"

N.B.

1) Complexity of final region depends on number of hidden nodes

2) Finite $\beta \Rightarrow$ rounded edges for selected region.

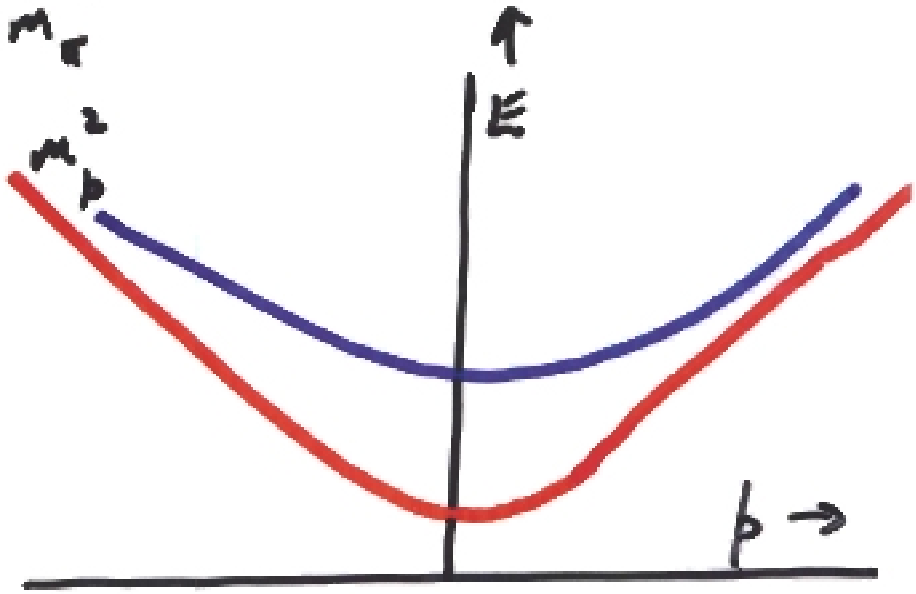
[Output contour lines in V_1–V_2 plane]

TOY EXAMPLE

Try to separate $\left\{ \begin{matrix} \pi \\ p \end{matrix} \right\}$ using p & E

π: $E^2 = p^2 + m_\pi^2$

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



Easy : $p = 0 \rightarrow 2$ GeV/c

Harder : $p = -4 \rightarrow +4$ GeV/c

Hardest $\left. \begin{matrix} p_x \\ p_y \\ p_z \end{matrix} \right\} = -4 \rightarrow +4$ GeV/c

More realistic : Add scatter of data about curves

Is NN better than simple cuts?

In principle, NO

[Can cut on complicated variable
e.g. NN output]

In practice, YES (usually)

But better NN performance
$\Rightarrow$ motivation to improve cuts.

PHYSICS EXAMPLE

Separate $e^+e^- \to c\bar{c}$ ..

from ; $b\bar{b}$, $q\bar{q}$, W^+W^-, ZZ

at LEP

Input variables: "Lifetime"

Track rapidities

Secondary vertex mass

quality

etc.

ISSUES: PRE N-N. CUTS

MISSING VARIABLES

WHERE TO GET TRAINING/TESTING EVENTS

HOW MANY NN'S.

HOW MANY INPUT VARIABLES

HOW MANY HIDDEN NODES/LAYERS

SINGLE OUTPUT OR SEVERAL

RATIO OF $c\bar{c}$, $b\bar{b}$, $q\bar{q}$,... TRAINING EVENTS

SYSTEMATICS [USE DIFFERENT SETS OF <u>TESTING</u> EVENTS]

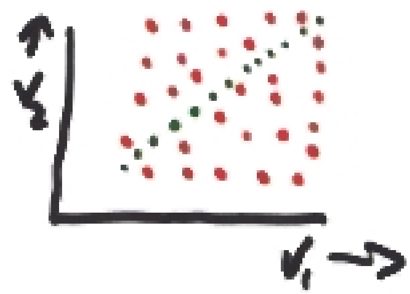
STABILITY W.R.T. NN CUT

NN SUMMARY

ADVANTAGES:

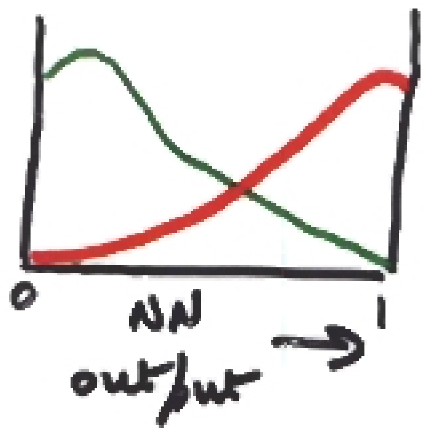
VERY FLEXIBLE

CORRELATIONS O.K.



TUNABLE CUT

e.g. minimum σ



DISADVANTAGES

TRAINING TAKES TIME

TENDENCY TO INCLUDE TOO MANY VARIABLES

TREAT AS BLACK BOX

OVER LAST FEW YEARS, CHANGE IN ATTITUDE FROM:

CONVINCE COLLEAGUES NN IS SENSIBLE

TO

"WHY DON'T YOU USE NN?"

BLUE

Best Linear Unbiassed Estimate

$x_i \pm \sigma_i$, possibly correlated

$\hat{x} = \sum \alpha_i x_i \qquad \sum \alpha_i = 1$

$\sigma_{\hat{x}}^2 = \text{minimum}$

LL , Duncan Gibault & Peter Clifford
NIM A270 (1988) 110

For contemplation:

Given x_1 and x_2 (& error matrix),

Can $\hat{x}$ lie outside range $x_1 \to x_2$?