How Statistics can Improve your Experiment

Jim Linnemann
Michigan State University

University of Virginia HEP Seminar

Feb 6, 2008

Two topics

Event weighting: competitive with ML
  and less computation

Evaluating Systematic Errors
  usual methods don't get all variation

Phystat.org conferences every 2y or so

site map accessibility contact

Phystat

Phystat Physics Statistics Code Repository

log in

An open, loosely moderated repository for code, tools, and documents relevant to statistics in physics applications. Search and download access is universal; package submission is loosely moderated for suitability.

Using the Site

- Lists of packages
- Search for a package
- Submit a Package
- Comment on a package (not yet available)

About the Repository

- Repository Policies and Procdures
- The Phystat Repository Steering Committee
- Comment on the repository site or policies

PHYSTAT Conference Links

- PHYSTAT 07 (CERN) 05 (Oxford) 03 (SLAC) 02 (Durham)
- Phystat Workshops: 06 (BIRS/Banff) 00 (Fermilab) 00 (CERN)
- More Conferences and Workshops ...

Lists and Statistics Resources

- The R Project For Statistical Computing StatCodes (Center for Astrostatistics) Statistical Software Resources on the Web
- More resources ...

Event Weighting: The Context

Milagro cosmic γ ray experiment

2630 m altitude = 750 g/cm^2 (of 1030) overburden

H_2O Cherenkov pond (+ tank surface array) =

calorimeter after 20.5 Xo, 8.3λ

Task: tell if hadron or γ started the shower

AND: most cosmic rays are hadron-initiated (p, He,…)

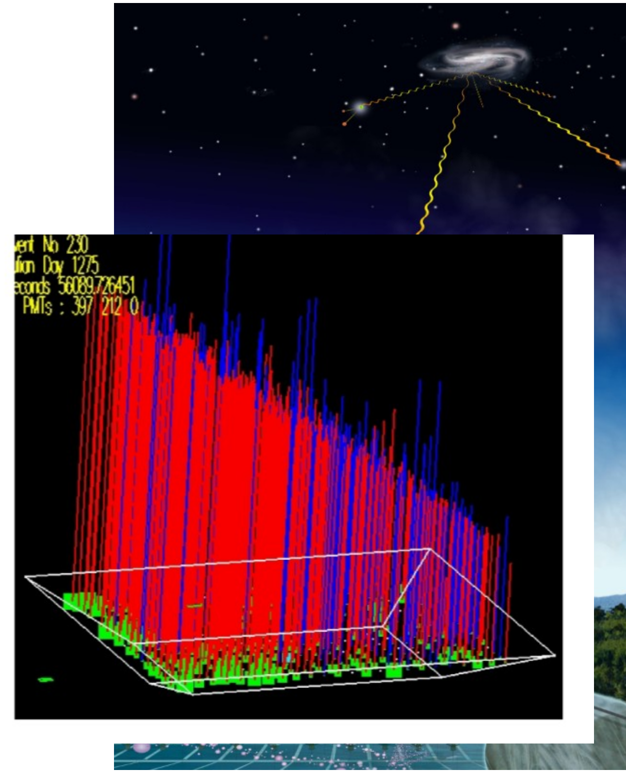
No big surprise that $\langle B \rangle \approx 10^3 \langle S \rangle$

Milagro Gamma Ray Observatory @ 8600’ altitude near Los Alamos, NM

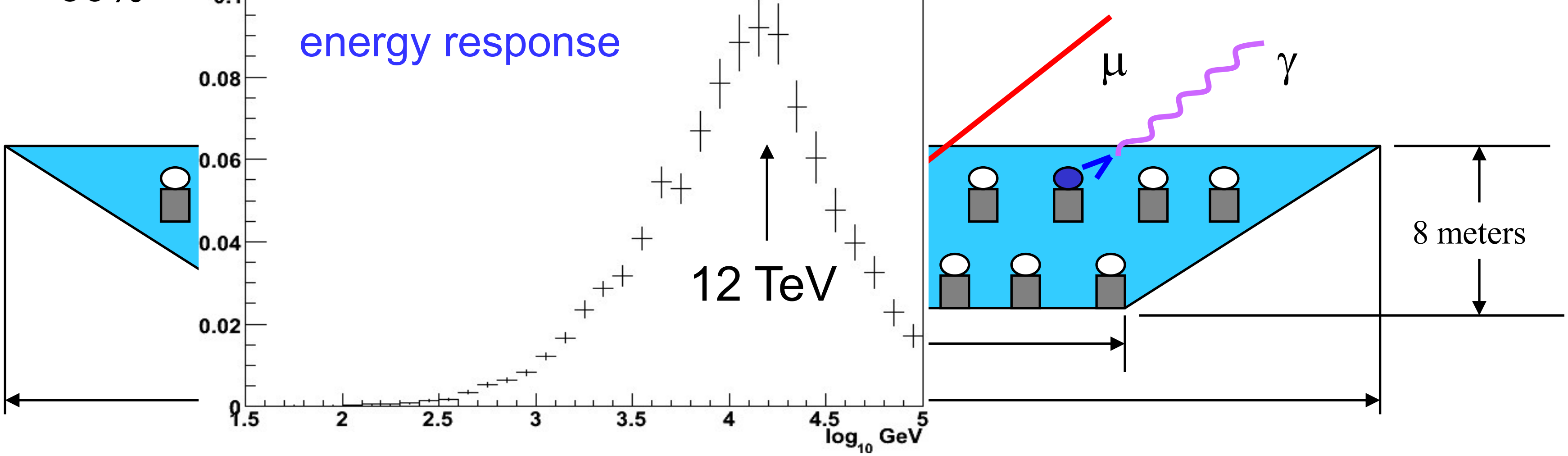
A. Abdo, B. Allen, D. Berley, T. DeYoung,B.L. Dingus, R.W. Ellsworth, M.M. Gonzalez, J.A. Goodman, C.M. Hoffman,P. Huntemeyer, B. Kolterman, C.P. Lansdell, J.T. Linnemann, J.E. McEnery, A.I. Mincer, P. Nemethy, J. Pretz, J.M. Ryan, P.M. Saz Parkinson, A. Shoup, G. Sinnis, A.J. Smith, G.W. Sullivan, D.A. Williams, V. Vasileiou, G.B. Yodh

Los Alamos NATIONAL LABORATORY
UC SANTA CRUZ
UNIVERSITY OF MARYLAND
MICHIGAN STATE UNIVERSITY
UCIrvine University of California, Irvine
New York University
NEW HAMPSHIRE

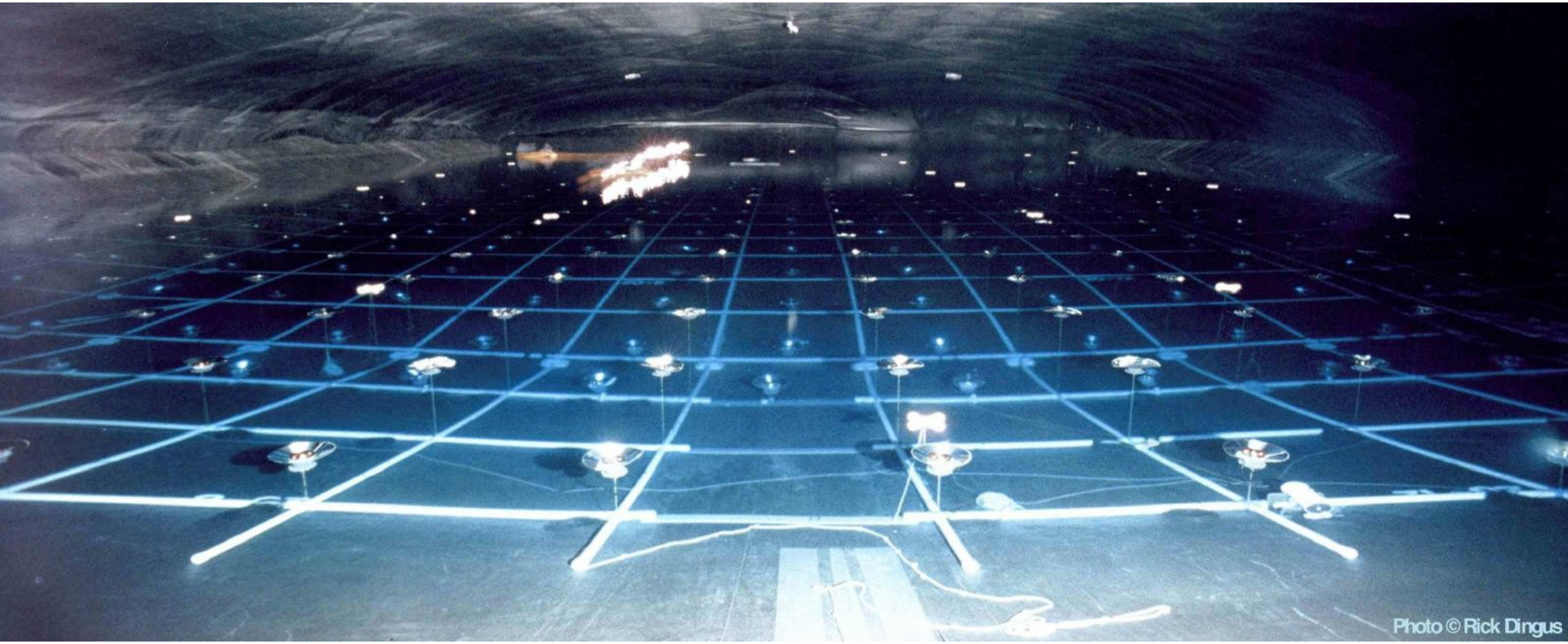
How Does Milagro Work?



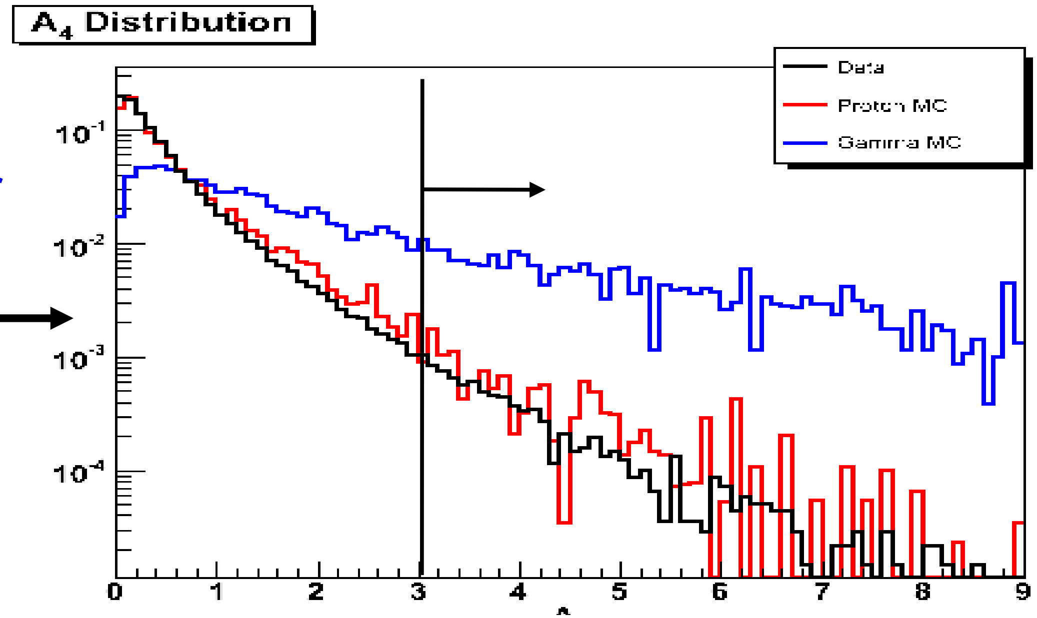
- Detect Particles in Extensive Air Showers from Cherenkov light created in 60m x 80 m x 8m pond containing filtered water
- Reconstruct shower direction to ~0.5° from the time different PMTs are hit
- 1700 Hz trigger rate mostly due to Extensive Air Showers created by cosmic rays
- Field of view is ~2 sr and the average duty factor is >90%



Inside the Milagro Detector



Milagro Background Rejection



Background Rejection Parameter

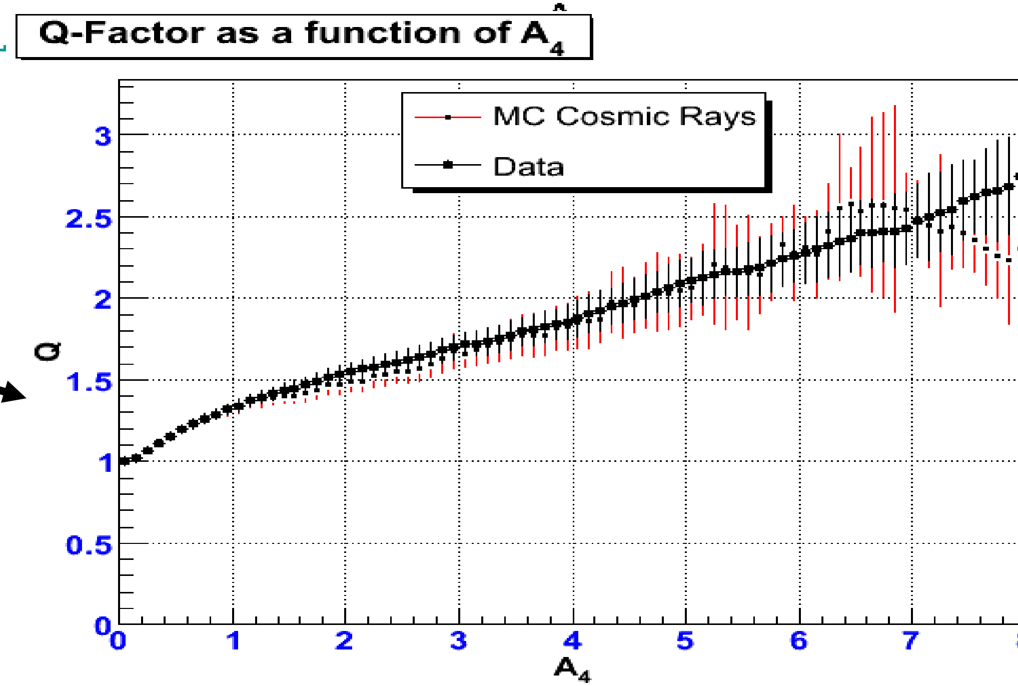
$$A_4 = \frac{(fTop + fOut) * nFit}{mxPE}$$

mxPE: maximum # PEs in bottom layer PMT

fTop: fraction of hit PMTs in Top layer

fOut: fraction of hit PMTs in Outriggers

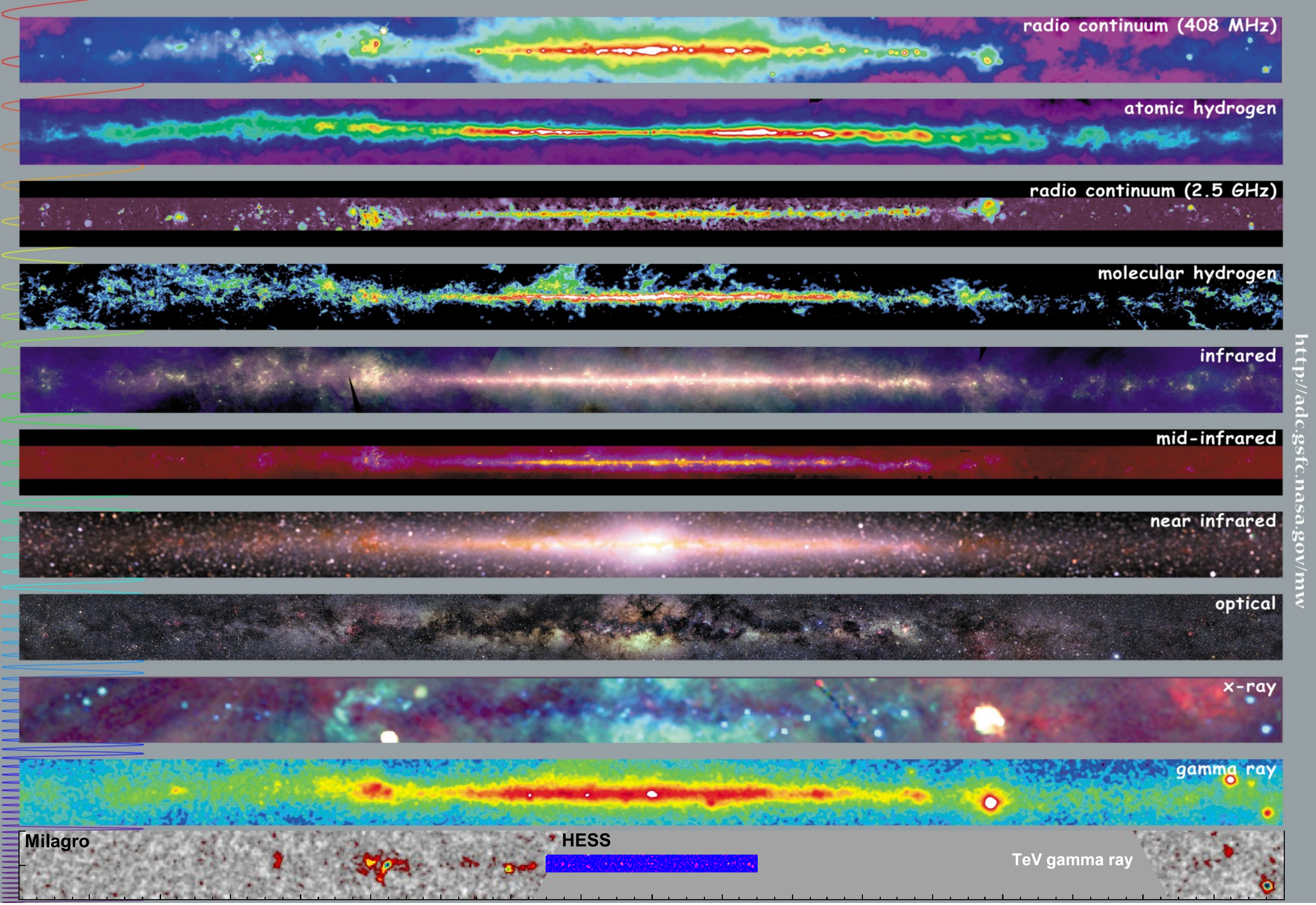
nFit: # PMTs used in the angle reconstruction



S/B increases with increasing A_4 so analysis weights events by S/B as determined by the A_4 value of the event

HAWC site is Sierra Negra, Mexico

- 4100 m above sea level
- Latitude of 19 deg N
- Easy Access
  - 2 hr drive from Puebla
  - 4 hr drive from Mexico City
- Existing Infrastructure
  - Few km from the US/Mexico Large Millimeter Telescope
  - Power, Internet, Roads
  - Sierra Negra Scientific Consortium of ~7 projects
- Excellent Mexican Collaborators
  - ~15 Faculty at 7 institutions have submitted proposal to CONACYT for HAWC
  - Experience in HEP, Auger, and astrophysics (including TeV)



TeV γ-rays: A New Window on the Sky

Background Subtraction

To see a signal, must subtract background with 10^{-3} precision

We do this: use nearby sky (“sideband”)

$$m = n - \hat{B}$$

Consider as a model for large-background LHC signal

Let's talk statistics

$\hat{\theta}$ Estimate of parameter

$E[\theta]$ Expected value

Gaussian Significance etc.

$$Z = m / \delta m = m / \sqrt{Var(m)}$$

$$1/Z = \text{fractional error} = \sigma / \mu = \text{Coeff. Variation}$$

$$\mathrm{N_e} = Z^2$$ ***Poisson Events w/o bkg, with same σ/μ***

Ne < *m*, *B*; typical: *m*~1000, Ne~100

Significance Improvement

Let x be a discriminator variable (possibly n-dim)

so pdf's s(x) and b(x) are different

Suppose I selected on $x > x_c$

Define $Q = Z(x > x_c) / Z(\text{no cut})$

A good cut has $Q > 1$

Suppose background is well known:

$$\delta m \approx \sqrt{\langle B \rangle} \quad \text{Then } Q = \varepsilon_s / \sqrt{\varepsilon_b}$$

More stringent than $\varepsilon_s > \varepsilon_b$

I've seen HEP cuts which fail this

Event Weighting

My colleague (Andy Smith of U Md) says I should weight
m(x) (***background subtracted*** data)

with $$\boxed{w(x) = \langle S(x) \rangle \;/\; \langle B(x) \rangle}$$

$$= s(x) / b(x) \quad \text{(withina constant)}$$

event weights defined only to within a constant
constant cancels in wtd averages and Ne

$$\bar{f} = \left(\sum f \bullet w\right) / \left(\sum w\right); \qquad N_e = \left(\sum w\right)^2 / \sum w^2$$

Cheating? Already subtracted B(x)!

But he's right!

Want estimate of M = true photons (Signal mean)

Naïve:

$$\hat{M}_1 = \sum m_i$$

$$Var(\hat{M}_1) = \sum Var(m_i) = \left(\sum V_i\right)$$

Sum: over bins of x for example; or integ. over all x

Better: if know s(x) = shape of x distribution

each bin m_i is an estimate of M

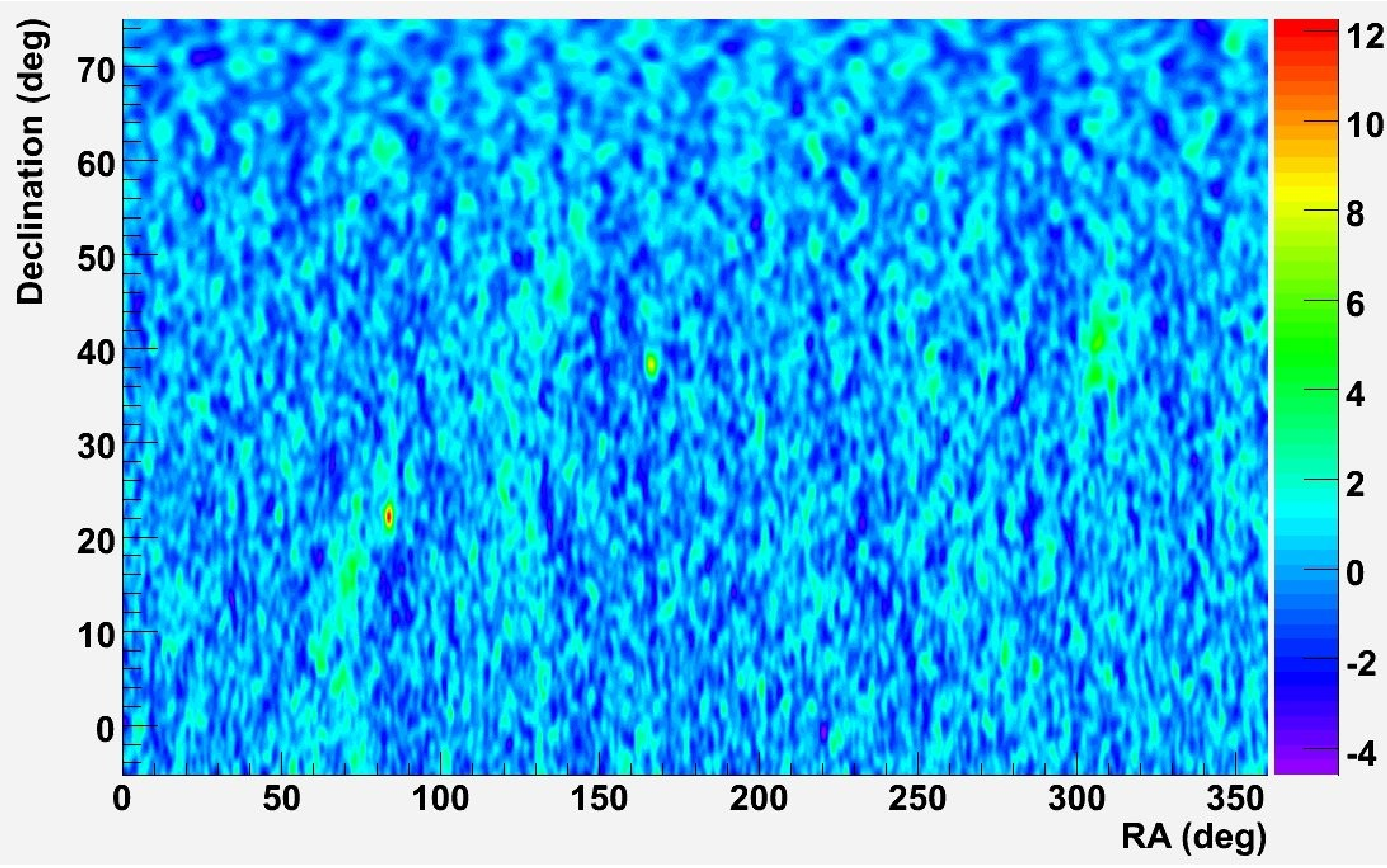
BLUE (Best Linear Unbiased Estimator)

Seek minimum variance estimator of M

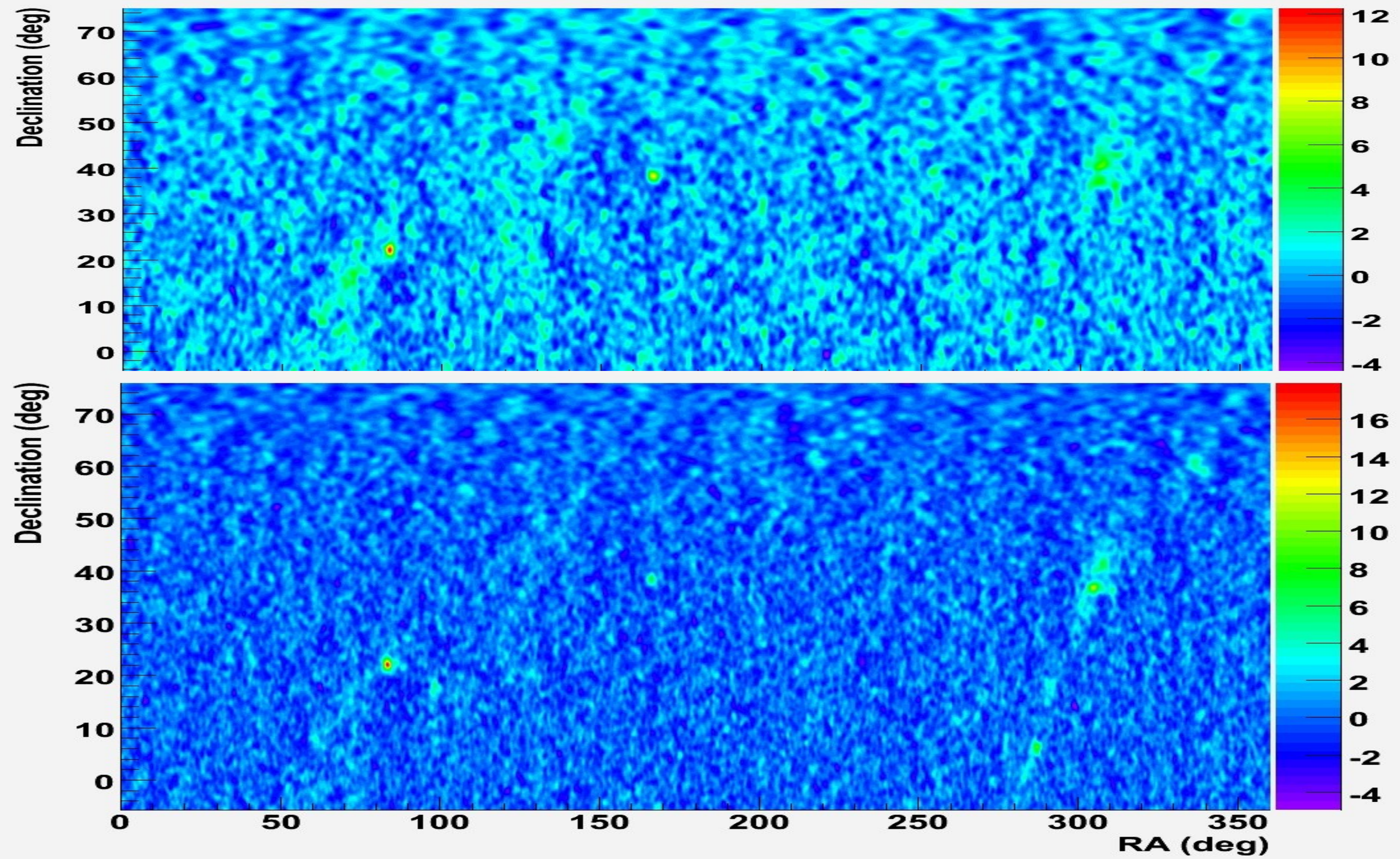
Equivalently, χ^2 fit for normalization multiplier

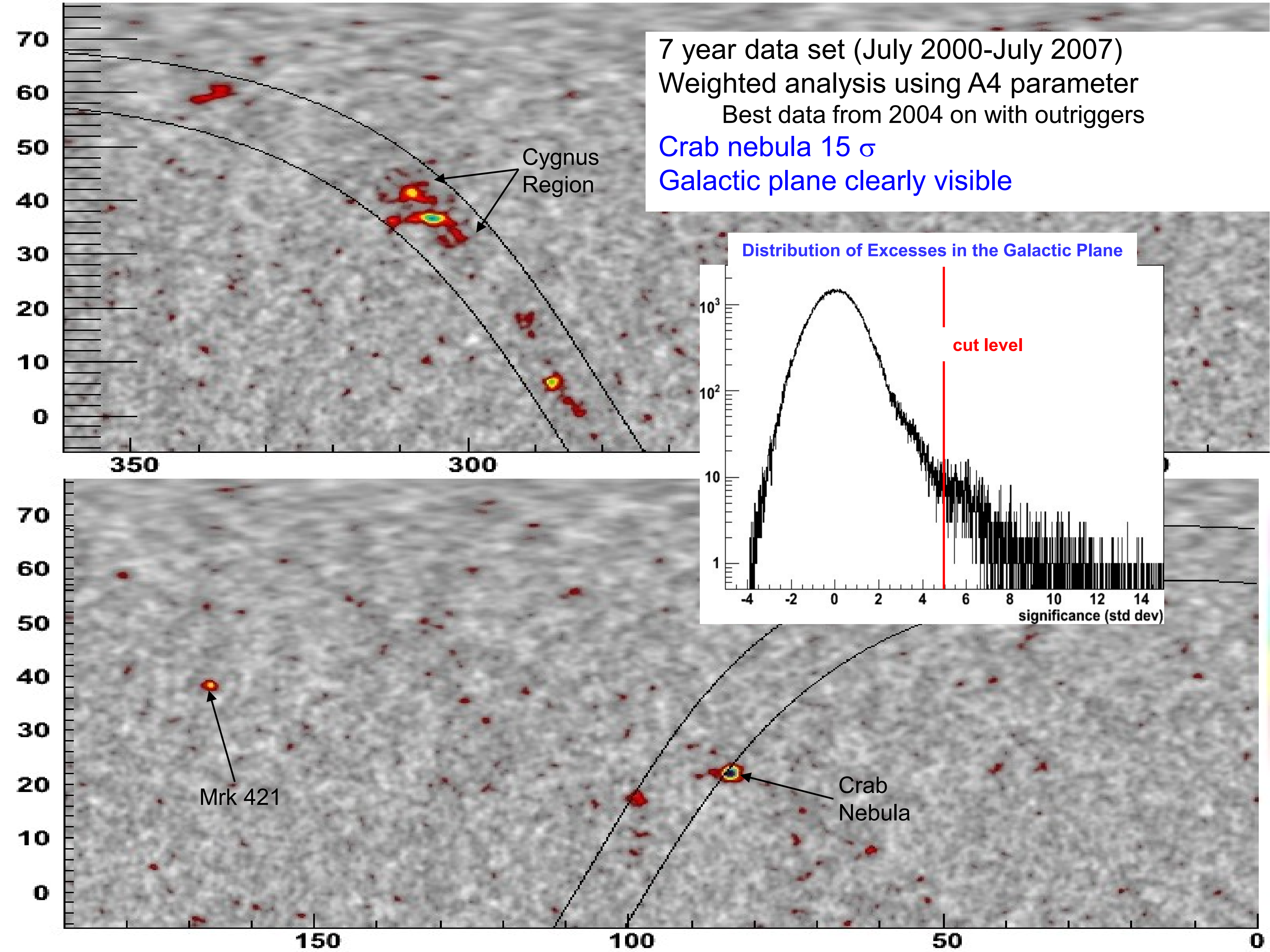
over bins of x

TeV Gamma Ray Sky: Before Weighting



TeV Gamma Ray Sky: Before and After Weighting





7 year data set (July 2000-July 2007)
Weighted analysis using A4 parameter
Best data from 2004 on with outriggers
Crab nebula 15 σ
Galactic plane clearly visible
Cygnus Region
Distribution of Excesses in the Galactic Plane
cut level
significance (std dev)
Mrk 421
Crab Nebula

BLUE treatment

Bin contents linear in parameter M:

$$\langle m_i \rangle = M s_i$$

Could have generalized with $s_i \rightarrow c_i\, s_i$

Gauss Markov: best estimator wtd by 1/variance:

$$\hat{M}_i = m_i / s_i; \quad Var(\hat{m}_i) = V_i / s_i^{\,2}; \quad w_i = 1/Var(\hat{m}_i)$$

$$\hat{M} = (\sum \hat{M}_i w_i) / \sum w_i = (\sum m_i s_i / V_i) / \sum s_i^{\,2} / V_i$$

Best = min variance among linear estimators

Using expected variance, not just estimated…

Chi-squared Treatment

Define and minimize a fit to the histogram of x:

$$\chi^2 = \sum \frac{(m_i - Ms_i)^2}{V_i}; \frac{\partial \chi^2}{\partial M} = 0 \text{ for } \hat{M}$$

$$\hat{M} = (\sum m_i s_i / V_i) / (\sum s_i^2 / V_i)$$

Bins could also be x bins over different data sets

BLUE = LLSQ

V_i = Variance of m_i (Careful: use true variance)

s(x) expected normalized signal distribution

$\Sigma s_i = 1$ (= $\int s(x)dx$) ; b(x) same for background

Then expected $m_i = M\, s_i$ and

$$\boxed{\hat{M} =} k \sum m_i \frac{s_i}{V_i} = \boxed{k \sum m_i u_i},$$

$$u_i = s_i / V_i; \quad 1/k = \sum \frac{s_i^{\,2}}{V_i}$$

Notice each m_i has a weight proportional to u_i

Can calculate M estimate just by accumulating weights!

Weight u_i

When $V_i \sim B_i$ (well-determined background)

and $B_i = B\ b_i$

$\boxed{u_i = s_i / b_i}$ in this limit

we have the advertised weight

(within a constant B, which doesn't matter)

When variance of mi and Bi estimated, use better Vi

V_i when B is uncertain

Reasonable: (assume Null Hyp for n in m=n-B; sidebands so $B = N_B/\tau$)

$V(m) \sim (n+N_B)/\tau$ (still close to B)

Better:

Calculate Z_{Bi} as in my PHYSTAT03 talk

Take $V \sim (m/Z_{Bi})^2$ (for m>0)

But: Careful: any variance small due to fluctuations should really use $m_i \rightarrow Ms_i$ (expected m_i) in calculations

(see Louis Lyons book)

Variance Improvement

$$Var(\hat{M}) = k^2 \sum Var(m_i)u_i^2 = k^2 \sum V_i u_i^2$$

$$= k^2 \sum (s_i^2 / V_i) = 1 / \sum (s_i^2 / V_i) = k$$

$$Var(\hat{M}_1) = \sum V_i \ \text{(larger)}$$

Cf. resistors: importance-weighted $R_{\|}$ vs. R_s
weighted variance ≤ unweighted

The variances are equal if all V_i, s_i equal

With optimum weights, approach Cramer-Rao min variance bound for enough data (Gauss-Markov theorem)

Cramer Rao Bound

As long as range of range of x indep of θ

And can swap derivative under integral sign

$$V[\hat{\vartheta}] \geq \left(1+\frac{\partial b}{\partial \theta}\right)^2 / E\left[-\frac{\partial^2 \ln L}{\theta^2}\right]$$

b=bias

Normal of known σ → V > σ^2

Efficient estimators when equality

ML whenever possible because:

If Efficient estimator exists, ML will find it

For large N, *always* efficient

Sensitivity to Assumptions

Since s and b normalized, indep. of absolute normalization assumptions.

However, sensitive to shape of s, b.

We know b accurately, fortunately:

b from data, so just use to check MC.

But s from MC: depends on

shower physics, and source energy spectrum

Test fit by χ^2 and pulls of fit of m’s to s, M.

A surprising application

Consider a map of counts vs. 2-d position xy: sky map.

Solve for sources by ML: consider all candidate positions, fit to photon excess * point spread function (angular resol)

many candidate pixels, events: ML infeasible

OR: weighting *all events* by

$$w(x) = s(xy)/(b(xy) + \alpha s(xy))$$

s(xy) = point spread function

b(xy) ~ flat; so w(xy) ~ s(xy) ~ 2d Gaussian (ideal)

So $\sum w$, $\sum w^2$ at each sky position (ideogram/kernel est.)

"ugh, you smeared the map" —but it approaches ML!

Modest (10%) gain in Z over "optimal" $s/\sqrt{b}$ bin size

BIG gains when 3d: {xy, z} where s(xy,z) varies with z

much more weight to events with good psf resolution!

General weighted event solution

Roger Barlow, J. Comp. Phys 72 (1987) p202

Write expected average weight in terms of parameter(s) and solve (Barlow):

$$p(x) = \alpha s(x) + (1-\alpha) b(x), \text{ so expect}$$

$$\overline{w}_d \equiv \frac{1}{N}\sum w = \alpha \overline{w}_s + (1-\alpha)\overline{w}_b; \text{ where}$$

$$\overline{w}_s = \int w(x) s(x) dx; \quad \overline{w}_b = \int w(x) b(x) dx$$

$$\text{solve for } \alpha \quad (\text{unbiased for any } w):$$

$$\hat{\alpha} = (\overline{w}_d - \overline{w}_b) / \left(\overline{w}_s - \overline{w}_b\right)$$

Why is weighting good?

Textbooks shows method of moments inefficient

ML typically has min var for parameters *a*

moments: generally above min var bound

A “moment” is just some weighting function whose data average you calculate

Then solve for the parameters *a* by equating to expected moments as f(*a*)

Typically weights not chosen optimally

$w(x) = x^k$ (classical moments)

say $x = \cos\theta$, expect $f(x) = 1 + \alpha x^2$; try k=2

solve $\langle x^2 \rangle = \langle x^2 (1 + \alpha x^2) \rangle$ for α

need not be good for estimating your parameters!

Barlow Optimal Weights

Calculated above ***unbiased*** solution for parameters for general weight function w(x), and its variance

Calculus of variations: find function w(x) giving minimum variance on parameter α (actually, on M)

Finds for large number of events, w(x) solution gives *same* variance as ML (*if* w(x) is close to optimal).

But: with weighting, unlike ML,

you do *NOT* need to iterate through all events!

Shows variance less than cut on ***same*** distribution w(x)

Comment: a fit to the distribution (histogram) of w(x) is also close to optimal

Barlow's Optimal solution:

$w(x) = \quad s(x) / (b(x) + \alpha_o\, s(x)), \qquad \alpha_o = M/B$

$= r(x) / (1 + \alpha\, r(x)) \; = \; 1/(\alpha + 1/ r(x)),$

where $\quad r(x) = s(x)/b(x)$

$w(x)\ \varepsilon\ [0,1];$ truly optimal if $\alpha_o = \alpha$

Cf. Neyman-Pearson best test variable:

$r(x) = s(x)/b(x)$

And discriminant variable

$d(x) =$ posterior prob$(s|x)$

$= s / (b + \alpha\, s), \; \alpha = \pi_s/(1 - \pi_s)$

What if weights are wrong?

Barlow: Near (quadratic) optimum, parameter **variance and Z estimates only slightly worse**
Note; MUST guess initial value for alpha, in order to estimate α: need α_o near true α

But: **wrong s or b => biased estimate of M**

you are fitting normalization to wrong shape

Relationship with BLUE

Barlow: knowing B reduces variance of M

Still: using same $w(x)$ is optimal.

Now compare with subtraction:

$w(x) = s(x)/(b(x) + \alpha s(x))$

When $\alpha \ll 1$, we recover our s/b above.

(i.e. for small α, s/b is near optimal)

F. Tkachov Optimal Weight

phyiscs/0001019=Part.Nucl.Lett.111(2002)28
physics/0604127

Elegant general principle for choosing w(x)

Again calculus of variations for minimum variance of parameter estimate

General:

$$w(x,a)_{\text{opt}} = C(a)\frac{\partial Ln[p(x;a)]}{\partial a} + D(a)$$

$$ML : \Sigma \frac{\partial Ln[p(x;a)]}{\partial a} = 0$$

$$\text{Let } p = (as+b)/(1+a)$$

$$w = s/(as+b) - 1/(1+a) \rightarrow s/(as+b)$$

Caution: He is “cavalier” with normalization of p(x)

Simpler ML/moments solution

Parameterize p = (as+b)/(1+a); a = (α/(1- α))

Then

$$\langle w_d \rangle = \int w(x)\,p(x)\,dx = \int \frac{(as+b)}{a+1} \frac{s\,dx}{(a_o s+b)} \approx 1/(a+1)$$

Compare ML Solution :

$$\sum w(x,a) = \frac{N}{a+1}$$

A Pitfall in Evaluating Systematic Errors

Ideal evaluation of Systematics?

Suppose know (Bayesian) pdf of systematic effects:

$\pi(\varphi) \rightarrow \pi(x,y)$ in 2d examples I'll use

e.g. {x,y} = {Jet Energy Scale factor, luminosity}

Let ***f(x,y)*** be what I am assessing systematic error of

single top cross section

Higgs mass

Upper Limit for SUSY in my channel

Nominal values for systematic params are at $\mathbf{x_o,y_o}$.

Redefine as (0,0), i.e. $(x,y) \rightarrow (x-x_o, y-y_o)$

Similarly, let **g(x,y)** = $f(x,y)-f(x_o,y_o)=f-f_o$ so **g(0,0)=0**

Systematic error = (not quite a variance—f_o not E[f])

$$V = \int dx\, dy\; g^2(x,y)\, \pi(x,y)$$

Instead: Do "Standard" Systematic Evaluation

You have a list of systematics; you ran MC at 0 point

Now run MC at + 1 σ for each systematic

Resulting changes are $d_i = f - f_o$

$$S^2 = \sum d_i{}^2$$

Report Systematic Error:

$$f_o \pm S$$

the "graduate student" solution?

What Justifies This?

1st order Variance Formula:

$$V = \sum_i \sum_j \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} Cov(x_i, x_j); \quad \text{eval } \frac{\partial f}{\partial x_i} \text{ at } \vec{x} = 0$$

Nice: avoid distribution assumptions on *π*, just Cov(x)

Claim can ignore cross terms:

$Cov(x_i,x_j) = 0$: systematics (usually) **uncorrelated**

What if your expt. contributes to PDF fits?

First order, so **good for linear** dependence of f on x

But we do a bit better:

finite differences to estimate partials (from MC…)

take into account some nonlinearity, right?

One Factor At A Time: OFAT

From my thesis advisor:

A physicist should be able to find and fix
any one single problem.

*It should take 2 things **both** wrong at the same time*
to confuse a physicist.

Corollary:

Changing more than one thing at a time
is asking for trouble.

V(exact) vs. S^2(OFAT): How well do we do?

Take $x_i \rightarrow z_i = x_i / \sigma_i$
consider $z_i = \pm 1$

Take $\pi(x,y) \sim N(0,a) \times N(0,b)$

$f=x+y$	Truly linear	
$V = a^2 + b^2$	$S^2=V$	OK as expect
$f = x^2 + y^2$	quadratic	
$V = 3a^4+2a^2b^2+3b^4$	$S^2=a^4+b^4$	**not so hot**
$f = xy$	bilinear	
$V = a^2b^2$ but	$S = 0$	**complete failure**

What went wrong?

Quadratic terms underestimated

finite diffs not enough to give effect on ***variance***

Covariance = 0 does not protect us from xy

xy and derivatives 0 on axes— as if ***f*** indep. of x,y

xy has twisting of *f* surface:

x derivatives depend on y and vice versa

Must consider off-axis points!

If you go to quadratic terms in Taylor series for V, need both xy and x^2, y^2 (consider rotations!)

Barlow: *run at* $\pm 1\sigma$, $d_i = (f^+ - f^-)/2$

makes quadratic → 0 …if you are asleep

You should notice $(f^+ - f_o) \neq -(f^- - f_o)$

don't forget about the 0 point

"Postdoc Solution?"

You have a list of systematics; you ran MC at 0 point

You run MC at $\pm 1\ \sigma$ for each systematic

Resulting changes are $d_i^{\pm}$

Report Systematic Error:

$$S_u^2 = \Sigma \max\{d_i^+, d_i^-\}^2$$

$$S_d^2 = \Sigma \min\{d_i^+, d_i^-\}^2$$

Report: $$f_o {}^{+Su}_{-Sd}$$

Here we can check for or even account for asymmetry of uncertainties on effects of systematics; should at least notice quadratic, but still BLIND to xy.

DOE
Design of Experiments
not your funding agency

OFAT is not a statistician's term of endearment. They wish your thesis advisor had talked to them first:

Always change more than one at a time

Assume each run long enough to measure effects of interesting size

Search for effects in order of likely importance

all linear (main effects)

then bilinear (2^{nd} order interactions)

then 3fold etc

Typically a few effects dominate

One expects "interactions" to be small if *each* main effect of interaction is small (i.e. bare xy term rare)

Interaction: twisting in response plane, i.e. *slope* wrt a variable depends on value of another variable

Typical Goals of DOE

1) Optimization/search
   - Best pattern of points for searching for
     - best yield for curing tracker epoxy
     - least variance of mass vs. cuts
   - Look for pattern to find a hilltop
     - which direction, if any, uphill from here?
     - i.e. good point set for numerical derivatives

2) Robustification (Taguchi)
   - Look for max or min (stationary)
   - worry about simultaneously maximizing multiple objectives
   - Look for ridge (separate important from unimportant params)
   - strangely named metrics to optimize

Response surface methodology: characterize shape of ***f***
- pattern of points for data to fit to 2nd degree curves
- geometry to characterize classes of curves:
  - hilltop, ridge, rising ridge…
- "composite designs" add points to basic design to better characterize area (e.g. near maxima)

Glossary

Factor x_i variable; systematic parameter
or from Analysis of Variance: linear combinations
Level values used: 2 level example $\pm 1\sigma$; 3 levels {+ 0 -}
Additive f linear in x_i's
Main Effects linear terms
Active factors main effects which are significant
Interaction multilinear terms $x_i x_j$ or trilinear or higher
Curvature Quadratic term
Respose Surface f(x,y,…)
Twisting of Response Surface $\partial_x f(x,y) \neq \partial_x f(x,0)$
Confounding Fractional Design can't Distinguish all interactions
can detect whether one of class active
ideally confound higher order with lower order
Factorial Design plan for sampling x_i space
Full: L^k all combinations of L levels of k factors
Fractional: L^{k-m} not all combinations
k has "subtracted" off m things confounded

OFAT vs. Design

OFAT advantages

Simpler to set up (fewer changes from nominal)

OK if main effects dominate

Easier to analyze w/o specialized software

One bad run loses less information

Can identify curvature if use 0

Design advantages

Can estimate interactions (or show negligible)

More important savings, the more variables

Less error (all runs contribute to each effect)

Can identify curvature if use 0

All DOE's change more than one factor at a time

2^2 full factorial design 2 levels +1, -1;

Zx	Zy
+1	+1
+1	-1
-1	+1
-1	-1

"Screening designs" in higher dimensions:

Not full 2^k combinations for 2 levels

See all main effects, and Groups of interactions

confound several low order, or low with high order

Calculating Main Effects and Interactions

Look at sign of factors in {x,y} runs:

Sgn {x,y}	++	+-	-+	--
Sgn (xy)	+	-	-	+
run	1	2	3	4

[(1 - 3) + (2 - 4)]/4 = main effect in x

compare the 2 terms for consistency: look for twisting

each term parallel to axes

rather than on axes like [(+0) – (-0)]/2

[(1 - 2) + (3 - 4)]/4 = main effect in y

[(1 - 2) + (4 - 3)]/4 = interaction xy

Or: fit Ax+By+Cxy to points

Sample calculations w/ DOE

without 0 point

$f=x+y$ no interactions

$V = a^2 + b^2$ $= S^2$ OK DOE=V

$f = xy$

$V = a^2b^2$ $S^2 = 0$ BAD DOE=V

$f = x^2 + y^2$

$V = 3a^4+2a^2b^2+3b^4$ $S^2 = a^4+b^4$ Ouch DOE= 0 Worse

DOE from sums of squares of main effects

Both need to explicitly look at 0 point to *notice* curvature
and can be extended to estimate effects better

OFAT CAN'T see xy even with 0 point added, but DOE can

Summary for Weighting

An optimal weight function can achieve ML accuracy

Weighting methods are powerful and simple

There is a rational scheme to choose optimal weight

Weighting (or fitting to weight distributions)
is more accurate than cuts

Summary for Systematics

- Even if your systematics *are* independent, your measurement probably correlates them for you
- If you worry about curvature (up-down asymmetry) you need to worry about xy too
- OFAT is blind to multi-linear (xy-like) effects
- You MUST leave OFAT to see xy-like terms
- OFAT evaluation of systematics misses some of nonlinear effects
- Don’t forget the point at nominal parameter values
- Statisticians have heard before from scientists who insist OFAT is the best/only way
- DOE might even help you—worth a think

References

My papers should appear soon at the phystat 07 web site
phystat.org | 07 | Proceedings

I'm in the process of putting them on the arxiv server…

Weighting

Books by Cowan and by Fred James

Papers by Barlow and by Tkachov

Design of Experiments

Nancy Reid's talk at Phystat 2007

B. Gunter, Computers In Physics 7 May (1993) (not online alas—complain to AIP)

Can look at NIST handbook or Wiki for definitions and some discussions

Box Hunter & Hunter "Statistics for Experimenters" good, but feels a bit wordy

Cox & Reid "Theory of D.O.E"

more compact but sometimes too terse