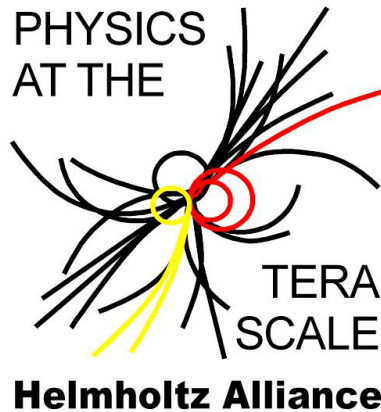


# Statistics for LHC Run II

<https://indico.desy.de/conferenceDisplay.py?confId=11244>



Terascale Statistics School  
DESY, Hamburg  
March 23-27, 2015



Glen Cowan  
Physics Department  
Royal Holloway, University of London  
[g.cowan@rhul.ac.uk](mailto:g.cowan@rhul.ac.uk)  
[www.pp.rhul.ac.uk/~cowan](http://www.pp.rhul.ac.uk/~cowan)

# Outline

Improving estimates of experimental sensitivity

Thoughts on multivariate methods

(Measuring distributions, unfolding)

# Recap of statistical tests

Consider test of a parameter  $\mu$ , e.g., proportional to cross section.

Result of measurement is a set of numbers  $\mathbf{x}$ .

To define test of  $\mu$ , specify *critical region*  $w_\mu$ , such that probability to find  $\mathbf{x} \in w_\mu$  is not greater than  $\alpha$  (the *size* or *significance level*):

$$P(\mathbf{x} \in w_\mu | \mu) \leq \alpha$$

(Must use inequality since  $\mathbf{x}$  may be discrete, so there may not exist a subset of the data space with probability of exactly  $\alpha$ .)

Equivalently define a  $p$ -value  $p_\mu$  such that the critical region corresponds to  $p_\mu < \alpha$ .

Often use, e.g.,  $\alpha = 0.05$ .

If observe  $\mathbf{x} \in w_\mu$ , reject  $\mu$ .

# Test statistics and $p$ -values

Often construct a test statistic,  $q_\mu$ , which reflects the level of agreement between the data and the hypothesized value  $\mu$ .

For examples of statistics based on the profile likelihood ratio, see, e.g., CCGV, EPJC 71 (2011) 1554; arXiv:1007.1727.

Usually define  $q_\mu$  such that higher values represent increasing incompatibility with the data, so that the  $p$ -value of  $\mu$  is:

$$p_\mu = \int_{q_{\mu,\text{obs}}}^{\infty} f(q_\mu|\mu) dq_\mu$$

observed value of  $q_\mu$

pdf of  $q_\mu$  assuming  $\mu$

Equivalent formulation of test: reject  $\mu$  if  $p_\mu < \alpha$ .

# Confidence interval from inversion of a test

Carry out a test of size  $\alpha$  for all values of  $\mu$ .

The values that are not rejected constitute a *confidence interval* for  $\mu$  at confidence level  $CL = 1 - \alpha$ .

The confidence interval will by construction contain the true value of  $\mu$  with probability of at least  $1 - \alpha$ .

The interval depends on the choice of the critical region of the test.

Put critical region where data are likely to be under assumption of the relevant alternative to the  $\mu$  that's being tested.

Test  $\mu = 0$ , alternative is  $\mu > 0$ : test for discovery.

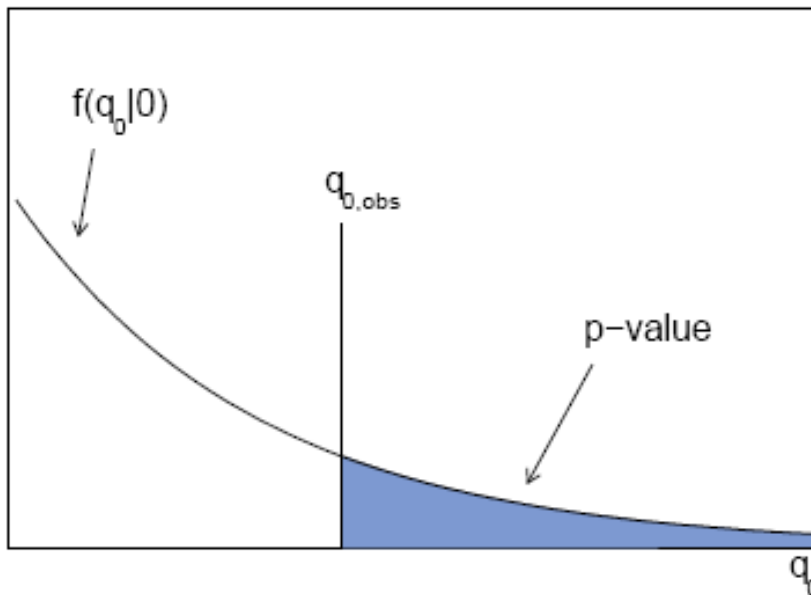
Test  $\mu = \mu_0$ , alternative is  $\mu \neq \mu_0$ : testing all  $\mu_0$  gives upper limit.

# $p$ -value for discovery

Large  $q_0$  means increasing incompatibility between the data and hypothesis, therefore  $p$ -value for an observed  $q_{0,\text{obs}}$  is

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|0) dq_0$$

will get formula for this later

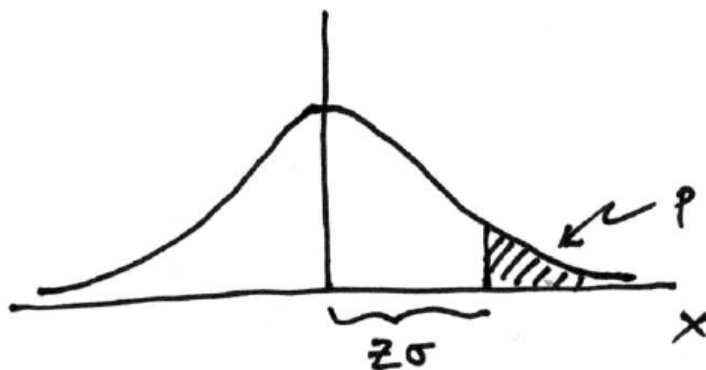


From  $p$ -value get equivalent significance,

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

# Significance from $p$ -value

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



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

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

# Prototype search analysis

Search for signal in a region of phase space; result is histogram of some variable  $x$  giving numbers:

$$\mathbf{n} = (n_1, \dots, n_N)$$

Assume the  $n_i$  are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s) dx, \quad b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b) dx.$$

signal

background



## Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \dots, m_M)$$

Assume the  $m_i$  are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

 nuisance parameters ( $\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\text{tot}}$ )

Likelihood function is

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

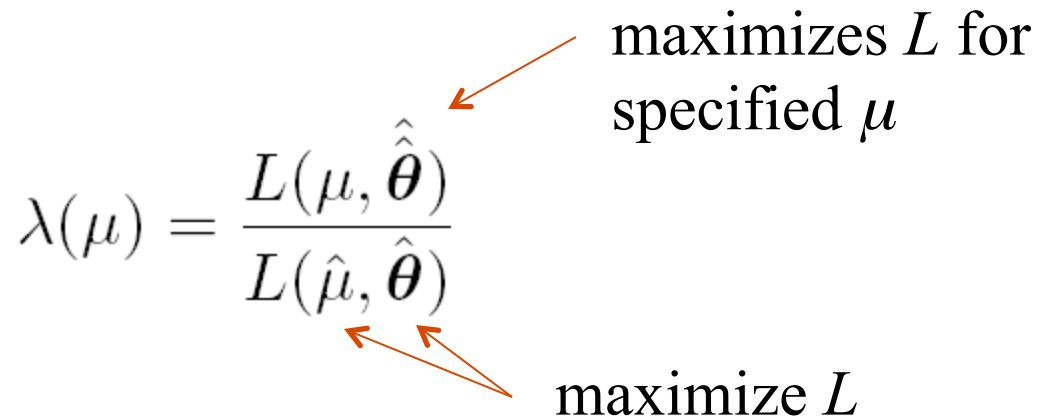
# The profile likelihood ratio

Base significance test on the profile likelihood ratio:

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

maximizes  $L$  for specified  $\mu$

maximize  $L$



The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

The profile LR should be near-optimal in present analysis with variable  $\mu$  and nuisance parameters  $\theta$ .

## Test statistic for discovery

Try to reject background-only ( $\mu = 0$ ) hypothesis using

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only regard upward fluctuation of data as evidence against the background-only hypothesis.

Note that even though here physically  $\mu \geq 0$ , we allow  $\hat{\mu}$  to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

## Distribution of $q_0$ in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of  $q_0$  as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right) \delta(q_0) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{q_0}} \exp\left[-\frac{1}{2} \left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case  $\mu' = 0$  is a “half chi-square” distribution:

$$f(q_0|0) = \frac{1}{2} \delta(q_0) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{q_0}} e^{-q_0/2}$$

In large sample limit,  $f(q_0|0)$  independent of nuisance parameters;  $f(q_0|\mu')$  depends on nuisance parameters through  $\sigma$ .

## Cumulative distribution of $q_0$ , significance

From the pdf, the cumulative distribution of  $q_0$  is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case  $\mu' = 0$  is

$$F(q_0|0) = \Phi(\sqrt{q_0})$$

The  $p$ -value of the  $\mu = 0$  hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance  $Z$  is simply

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$$

# Test statistic for upper limits

cf. Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554.

For purposes of setting an upper limit on  $\mu$  use

$$q_\mu = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \quad \text{where} \quad \lambda(\mu) = \frac{L(\mu, \hat{\boldsymbol{\theta}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized  $\mu$ :

From observed  $q_\mu$  find  $p$ -value: 
$$p_\mu = \int_{q_{\mu, \text{obs}}}^{\infty} f(q_\mu | \mu) dq_\mu$$

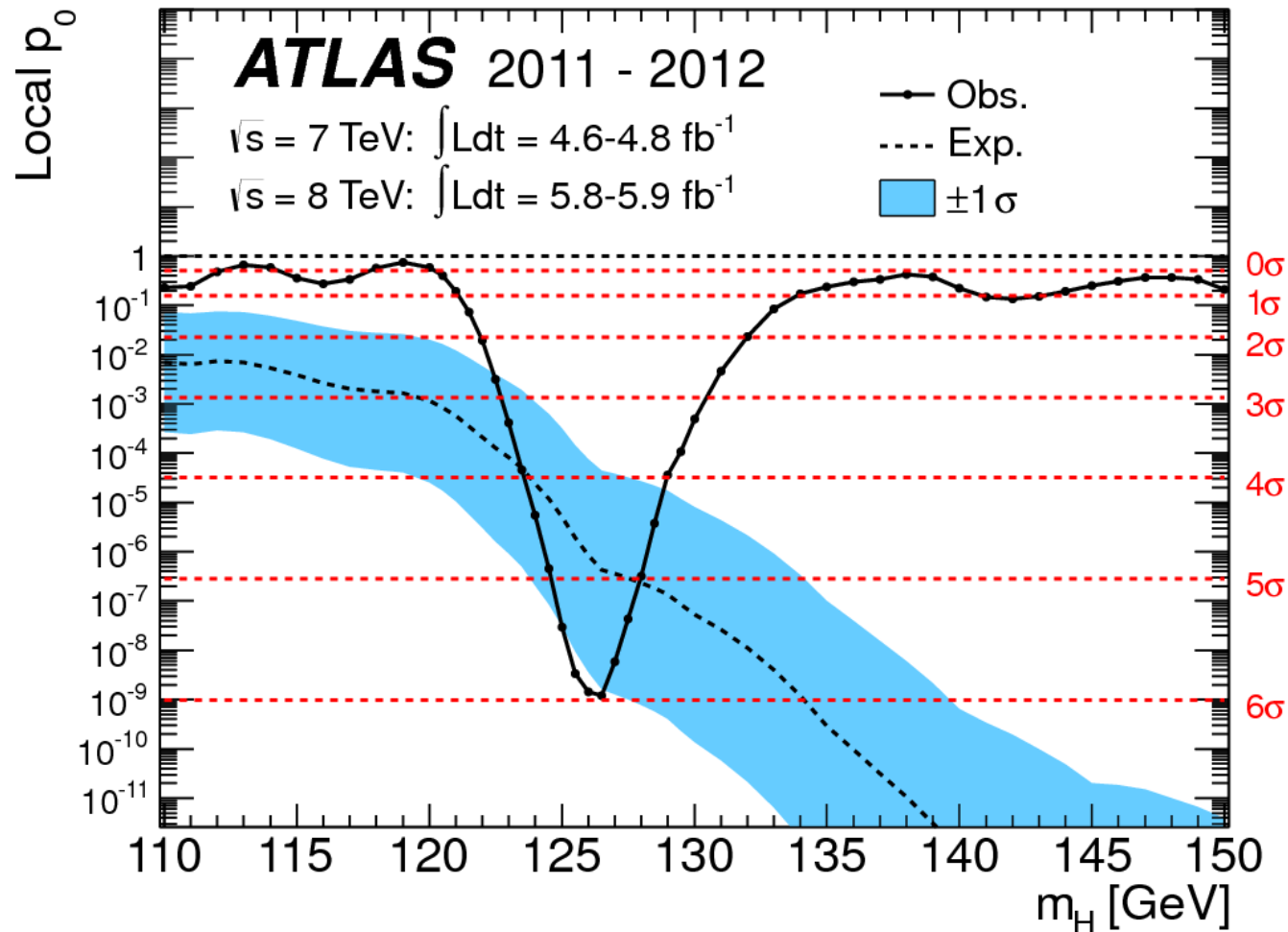
Large sample approximation:

$$p_\mu = 1 - \Phi(\sqrt{q_\mu})$$

95% CL upper limit on  $\mu$  is highest value for which  $p$ -value is not less than 0.05.

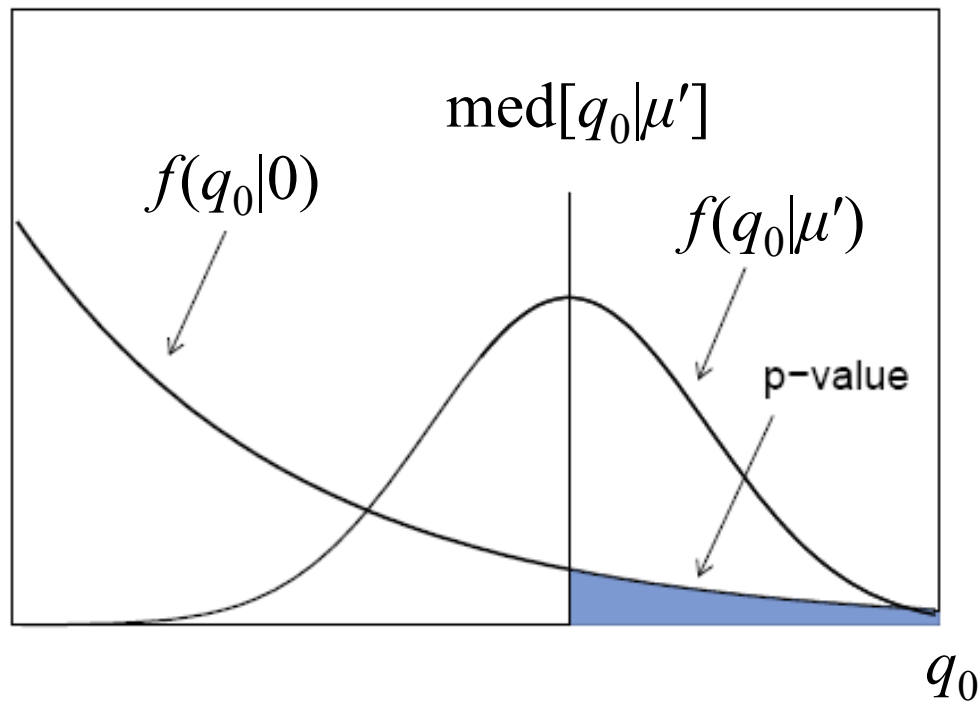
# Example of a $p$ -value

ATLAS, Phys. Lett. B 716 (2012) 1-29



# Expected (or median) significance / sensitivity

When planning the experiment, we want to quantify how sensitive we are to a potential discovery, e.g., by given median significance assuming some nonzero strength parameter  $\mu'$ .



So for  $p$ -value, need  $f(q_0|0)$ , for sensitivity, will need  $f(q_0|\mu')$ ,



# Expected discovery significance for counting experiment with background uncertainty

## I. Discovery sensitivity for counting experiment with $b$ known:

(a)  $\frac{s}{\sqrt{b}}$

(b) Profile likelihood ratio test & Asimov:  $\sqrt{2 \left( (s+b) \ln \left( 1 + \frac{s}{b} \right) - s \right)}$

## II. Discovery sensitivity with uncertainty in $b$ , $\sigma_b$ :

(a)  $\frac{s}{\sqrt{b + \sigma_b^2}}$

(b) Profile likelihood ratio test & Asimov:

$$\left[ 2 \left( (s+b) \ln \left[ \frac{(s+b)(b + \sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[ 1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

# Counting experiment with known background

Count a number of events  $n \sim \text{Poisson}(s+b)$ , where

$s$  = expected number of events from signal,

$b$  = expected number of background events.

To test for discovery of signal compute  $p$ -value of  $s = 0$  hypothesis,

$$p = P(n \geq n_{\text{obs}}|b) = \sum_{n=n_{\text{obs}}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; 2n_{\text{obs}})$$

Usually convert to equivalent significance:  $Z = \Phi^{-1}(1 - p)$   
where  $\Phi$  is the standard Gaussian cumulative distribution, e.g.,  
 $Z > 5$  (a 5 sigma effect) means  $p < 2.9 \times 10^{-7}$ .

To characterize sensitivity to discovery, give expected (mean or median)  $Z$  under assumption of a given  $s$ .

## $s/\sqrt{b}$ for expected discovery significance

For large  $s + b$ ,  $n \rightarrow x \sim \text{Gaussian}(\mu, \sigma)$ ,  $\mu = s + b$ ,  $\sigma = \sqrt{s + b}$ .

For observed value  $x_{\text{obs}}$ ,  $p$ -value of  $s = 0$  is  $\text{Prob}(x > x_{\text{obs}} \mid s = 0)$ ,:

$$p_0 = 1 - \Phi\left(\frac{x_{\text{obs}} - b}{\sqrt{b}}\right)$$

Significance for rejecting  $s = 0$  is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\text{obs}} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate  $s$  is

$$\text{median}[Z_0 \mid s + b] = \frac{s}{\sqrt{b}}$$

# Better approximation for significance

Poisson likelihood for parameter  $s$  is

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

For now  
no nuisance  
params.

To test for discovery use profile likelihood ratio:

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{s} \geq 0, \\ 0 & \hat{s} < 0. \end{cases} \quad \lambda(s) = \frac{L(s, \hat{\theta}(s))}{L(\hat{s}, \hat{\theta})}$$

So the likelihood ratio statistic for testing  $s = 0$  is

$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})} = 2 \left( n \ln \frac{n}{b} + b - n \right) \quad \text{for } n > b, \quad 0 \text{ otherwise}$$

# Approximate Poisson significance (continued)

For sufficiently large  $s + b$ , (use Wilks' theorem),

$$Z = \sqrt{2 \left( n \ln \frac{n}{b} + b - n \right)} \quad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

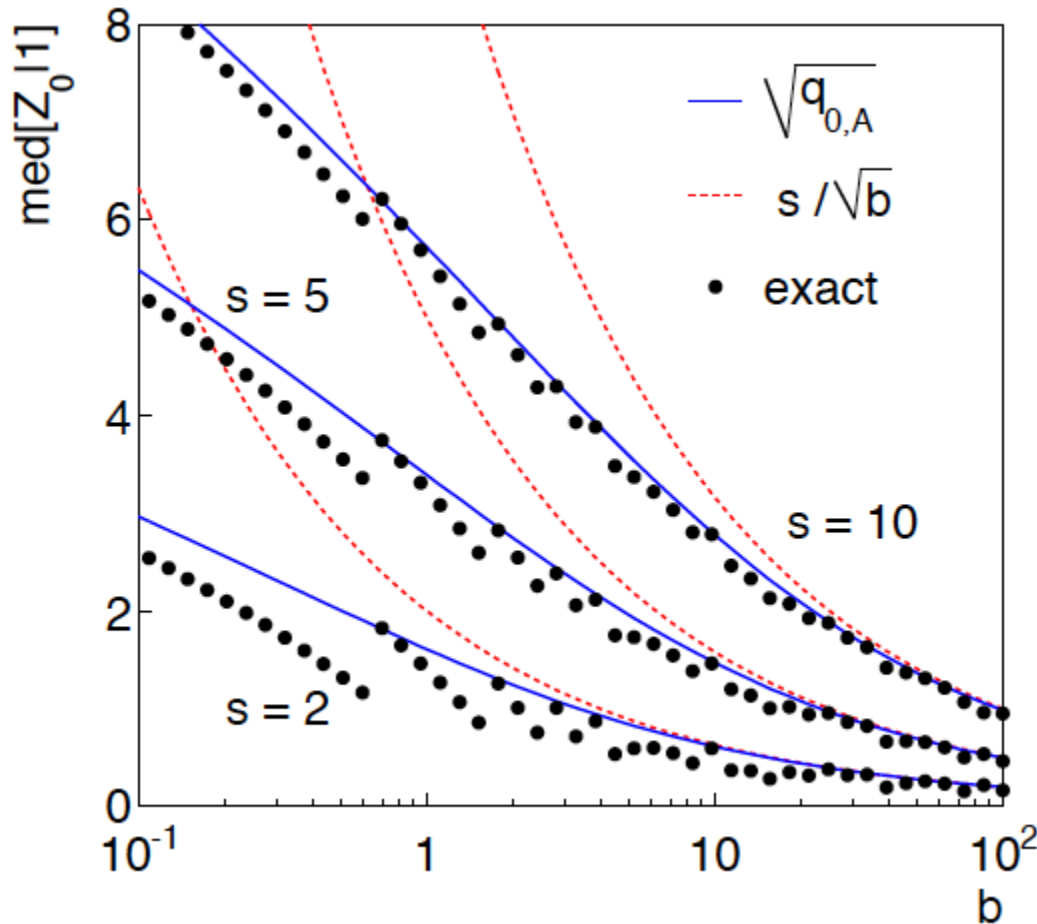
To find  $\text{median}[Z|s]$ , let  $n \rightarrow s + b$  (i.e., the Asimov data set):

$$Z_A = \sqrt{2 \left( (s + b) \ln \left( 1 + \frac{s}{b} \right) - s \right)}$$

This reduces to  $s/\sqrt{b}$  for  $s \ll b$ .

$n \sim \text{Poisson}(s+b)$ , median significance,  
assuming  $s$ , of the hypothesis  $s = 0$

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727



“Exact” values from MC,  
jumps due to discrete data.

Asimov  $\sqrt{q_{0,A}}$  good approx.  
for broad range of  $s, b$ .

$s/\sqrt{b}$  only good for  $s \ll b$ .

## Extending $s/\sqrt{b}$ to case where $b$ uncertain

The intuitive explanation of  $s/\sqrt{b}$  is that it compares the signal,  $s$ , to the standard deviation of  $n$  assuming no signal,  $\sqrt{b}$ .

Now suppose the value of  $b$  is uncertain, characterized by a standard deviation  $\sigma_b$ .

A reasonable guess is to replace  $\sqrt{b}$  by the quadratic sum of  $\sqrt{b}$  and  $\sigma_b$ , i.e.,

$$\text{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where  $\sigma_b$  cannot be neglected.

# Profile likelihood with $b$ uncertain

This is the well studied “on/off” problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...

Measure two Poisson distributed values:

$n \sim \text{Poisson}(s+b)$  (primary or “search” measurement)

$m \sim \text{Poisson}(\tau b)$  (control measurement,  $\tau$  known)

The likelihood function is

$$L(s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this to construct profile likelihood ratio ( $b$  is nuisance parameter):

$$\lambda(0) = \frac{L(0, \hat{\hat{b}}(0))}{L(\hat{s}, \hat{b})}$$



# Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\hat{s} = n - m/\tau ,$$

$$\hat{b} = m/\tau ,$$

$$\hat{\hat{b}}(s) = \frac{n + m - (1 + \tau)s + \sqrt{(n + m - (1 + \tau)s)^2 + 4(1 + \tau)sm}}{2(1 + \tau)} .$$

and in particular to test for discovery ( $s = 0$ ),

$$\hat{\hat{b}}(0) = \frac{n + m}{1 + \tau}$$

# Asymptotic significance

Use profile likelihood ratio for  $q_0$ , and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$Z = \sqrt{q_0};$$
$$= \left[ -2 \left( n \ln \left[ \frac{n+m}{(1+\tau)n} \right] + m \ln \left[ \frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2}$$

for  $n > \hat{b}$  and  $Z = 0$  otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

# Asimov approximation for median significance

To get median discovery significance, replace  $n$ ,  $m$  by their expectation values assuming background-plus-signal model:

$$n \rightarrow s + b$$

$$m \rightarrow \tau b$$

$$Z_A = \left[ -2 \left( (s + b) \ln \left[ \frac{s + (1 + \tau)b}{(1 + \tau)(s + b)} \right] + \tau b \ln \left[ 1 + \frac{s}{(1 + \tau)b} \right] \right) \right]^{1/2}$$

Or use the variance of  $\hat{b} = m/\tau$ ,  $V[\hat{b}] \equiv \sigma_b^2 = \frac{b}{\tau}$ , to eliminate  $\tau$ :

$$Z_A = \left[ 2 \left( (s + b) \ln \left[ \frac{(s + b)(b + \sigma_b^2)}{b^2 + (s + b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[ 1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

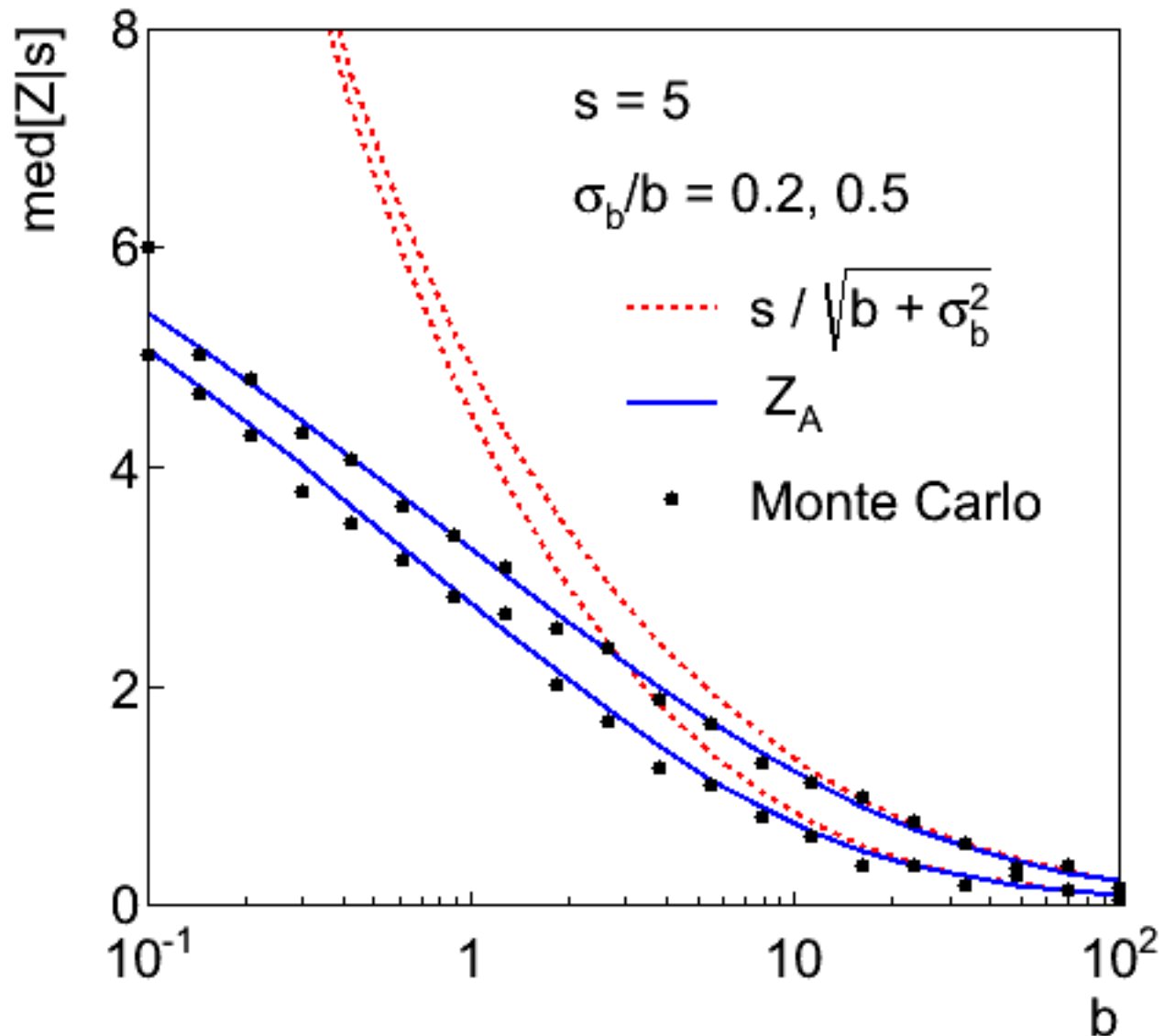
# Limiting cases

Expanding the Asimov formula in powers of  $s/b$  and  $\sigma_b^2/b$  ( $= 1/\tau$ ) gives

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}} \left( 1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right)$$

So the “intuitive” formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set.

## Testing the formulae: $s = 5$



# Using sensitivity to optimize a cut

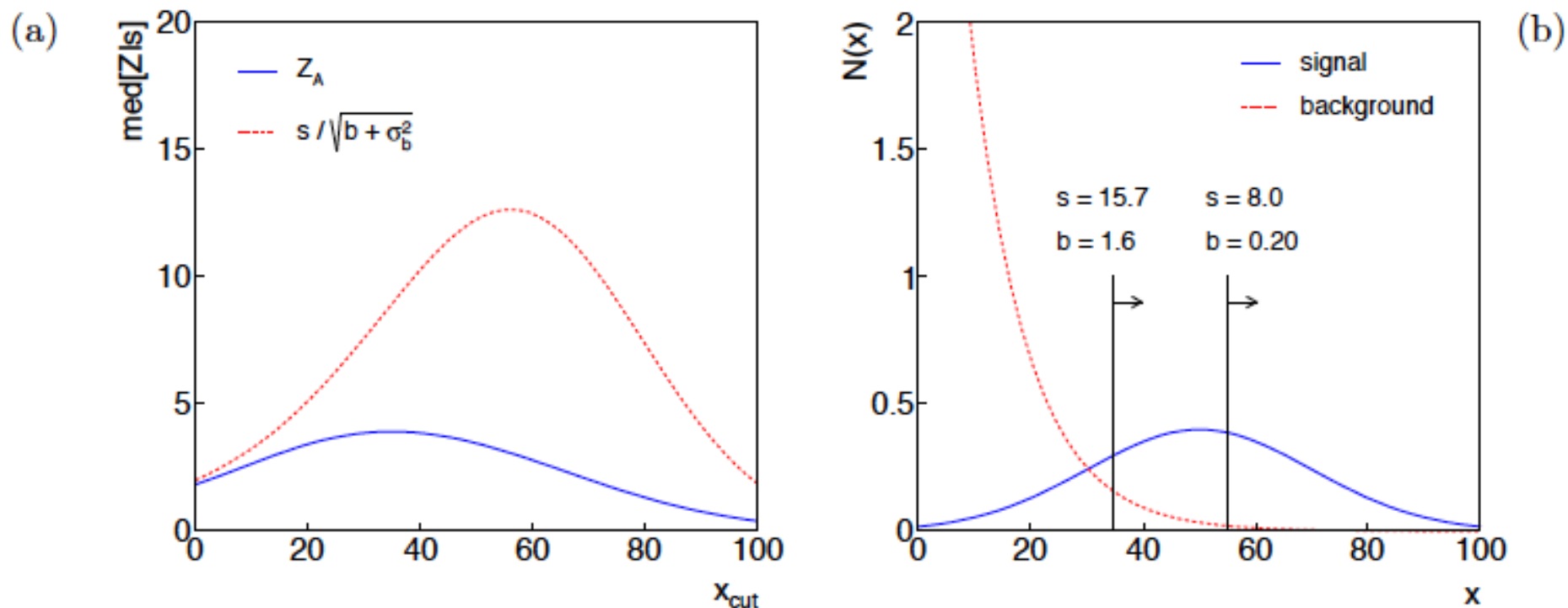


Figure 1: (a) The expected significance as a function of the cut value  $x_{\text{cut}}$ ; (b) the distributions of signal and background with the optimal cut value indicated.

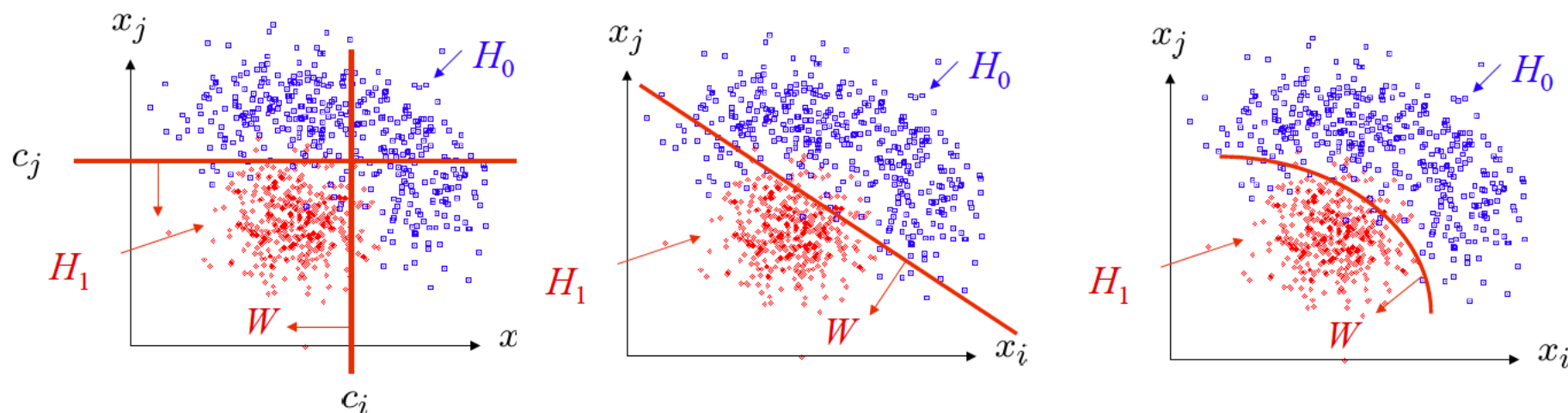
# Multivariate analysis in HEP

Each event yields a collection of numbers  $\vec{x} = (x_1, \dots, x_n)$

$x_1$  = number of muons,  $x_2 = p_t$  of jet, ...

$\vec{x}$  follows some  $n$ -dimensional joint pdf, which depends on the type of event produced, i.e., signal or background.

1) What kind of decision boundary best separates the two classes?



2) What is optimal test of hypothesis that event sample contains only background?

# Test statistics

The boundary of the critical region for an  $n$ -dimensional data space  $\mathbf{x} = (x_1, \dots, x_n)$  can be defined by an equation of the form

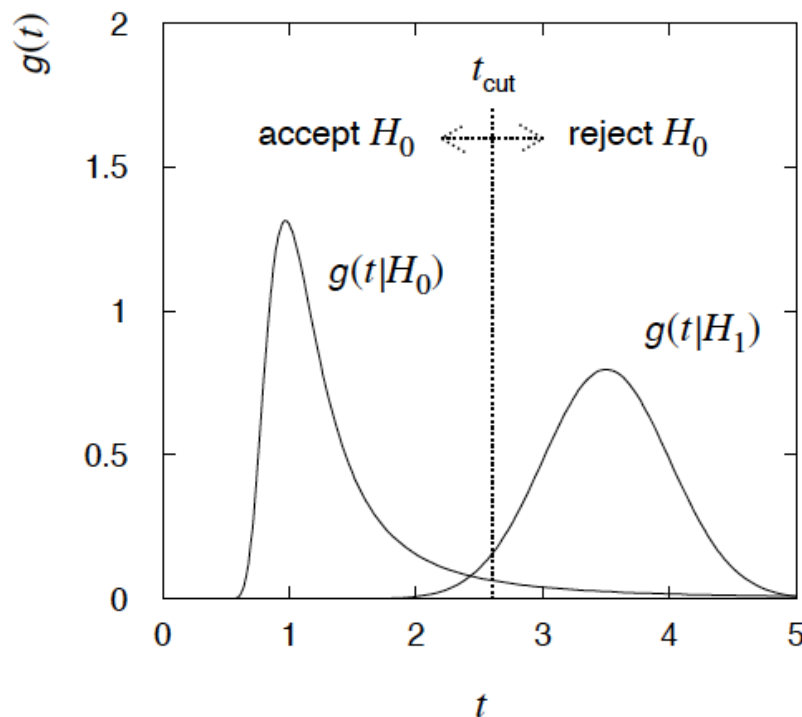
$$t(x_1, \dots, x_n) = t_{\text{cut}}$$

where  $t(x_1, \dots, x_n)$  is a scalar **test statistic**.

We can work out the pdfs  $g(t|H_0)$ ,  $g(t|H_1)$ ,  $\dots$

Decision boundary is now a single ‘cut’ on  $t$ , defining the critical region.

So for an  $n$ -dimensional problem we have a corresponding 1-d problem.





# Test statistic based on likelihood ratio

For multivariate data  $\mathbf{x}$ , not obvious how to construct best test.

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test of  $H_0$ , (background) versus  $H_1$ , (signal) the critical region should have

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

inside the region, and  $\leq c$  outside, where  $c$  is a constant which depends on the size of the test  $\alpha$ .

Equivalently, optimal scalar test statistic is

$$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

N.B. any monotonic function of this leads to the same test.

# Multivariate methods

In principle, likelihood ratio provides best classifier and leads also to the best test for the presence of signal in full sample.

But we usually don't have explicit formulae for  $f(\mathbf{x}|\mathbf{s})$ ,  $f(\mathbf{x}|\mathbf{b})$ ; we only have MC models from which we generate training data:

generate  $\mathbf{x} \sim f(\mathbf{x}|\mathbf{s}) \quad \rightarrow \quad \mathbf{x}_1, \dots, \mathbf{x}_N$

generate  $\mathbf{x} \sim f(\mathbf{x}|\mathbf{b}) \quad \rightarrow \quad \mathbf{x}_1, \dots, \mathbf{x}_N$

So instead of the likelihood ratio we try to construct a statistic that we can optimize using the training data.

Many new (and some old) methods:

Fisher discriminant, Neural networks, Kernel density methods Support Vector Machines, Decision trees with Boosting, Bagging, Deep Learning, ...

We continue to important new ideas from Machine Learning

# The Higgs Machine Learning Challenge

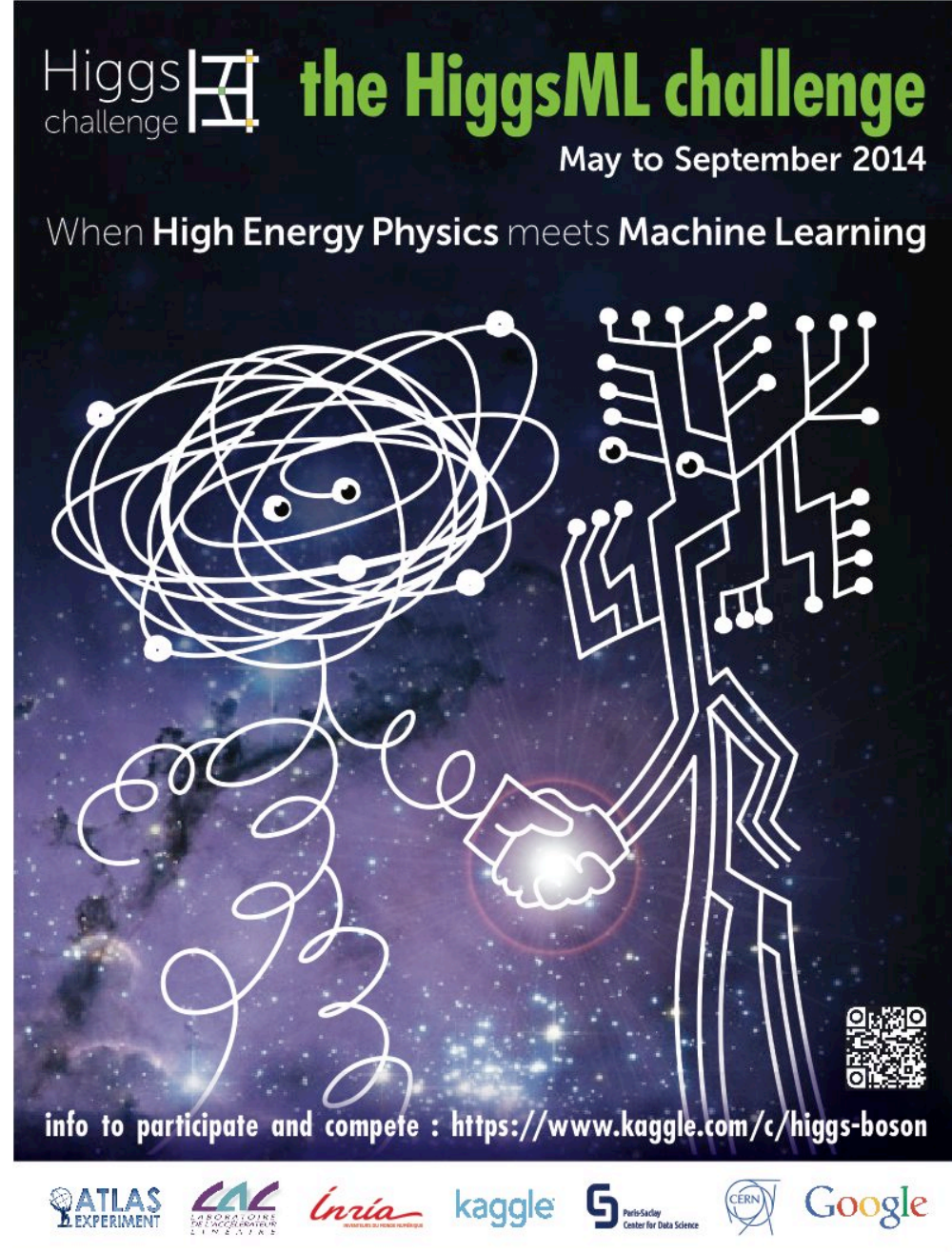
Samples of ATLAS MC data for  $H \rightarrow \tau\tau$  and backgrounds made publicly available through kaggle:

[www.kaggle.com/c/higgs-boson](http://www.kaggle.com/c/higgs-boson)

1785 teams (1942 people)  
from June-Sep 2014.

ML experts win easily:  
M. Gabor -- \$7000

Many new ideas e.g.,  
about Deep Learning,  
Cross Validation,...



The poster for the HiggsML challenge features a dark blue background with a starry space pattern. At the top, the text 'Higgs challenge' is in white, followed by a logo consisting of a square with a cross inside. To the right, 'the HiggsML challenge' is written in large green letters, with 'May to September 2014' below it. A central tagline reads 'When High Energy Physics meets Machine Learning'. The main illustration shows two figures shaking hands: one is a tangled white line representing a particle or a complex system, and the other is a white circuit board representing machine learning. A bright light emanates from their clasped hands. In the bottom right corner, there is a QR code. At the very bottom, a row of logos includes ATLAS EXPERIMENT, LAL (Laboratoire de l'Accélérateur Linéaire), Inria, kaggle, Paris-Saclay Center for Data Science, CERN, and Google.

Higgs challenge the HiggsML challenge  
May to September 2014  
When High Energy Physics meets Machine Learning  
info to participate and compete : <https://www.kaggle.com/c/higgs-boson>

ATLAS EXPERIMENT LAL Laboratoire de l'Accélérateur Linéaire Inria kaggle Paris-Saclay Center for Data Science CERN Google

## Organization committee

Balázs Kégl - *Appstat-LAL*  
Cécile Germain - *TAO-LRI*

David Rousseau - *Atlas-LAL*  
Glen Cowan - *Atlas-RHUL*

Isabelle Guyon - *Cholearn*  
Claire Adam-Bourdarios - *Atlas-LAL*

## Advisory committee

Thorsten Wengler - *Atlas-CERN*  
Andreas Hoecker - *Atlas-CERN*

Joerg Stelzer - *Atlas-CERN*  
Marc Schoenauer - *INRIA*

## A simple example (2D)

Consider two variables,  $x_1$  and  $x_2$ , and suppose we have formulas for the joint pdfs for both signal (s) and background (b) events (in real problems the formulas are usually not available).

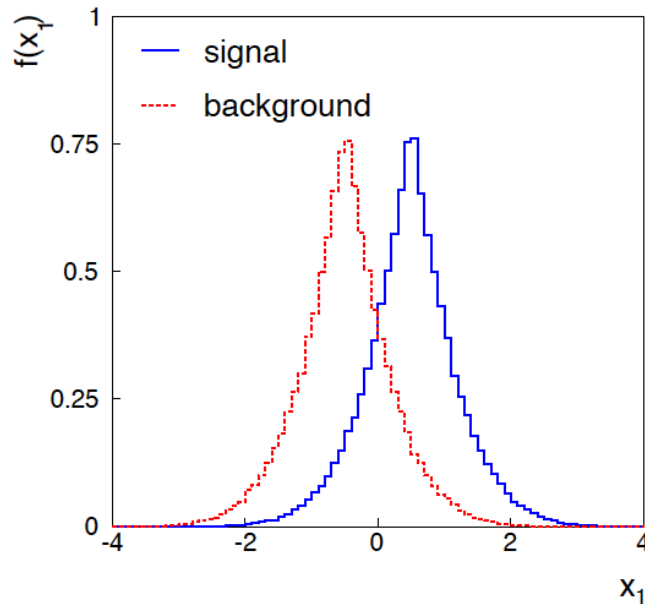
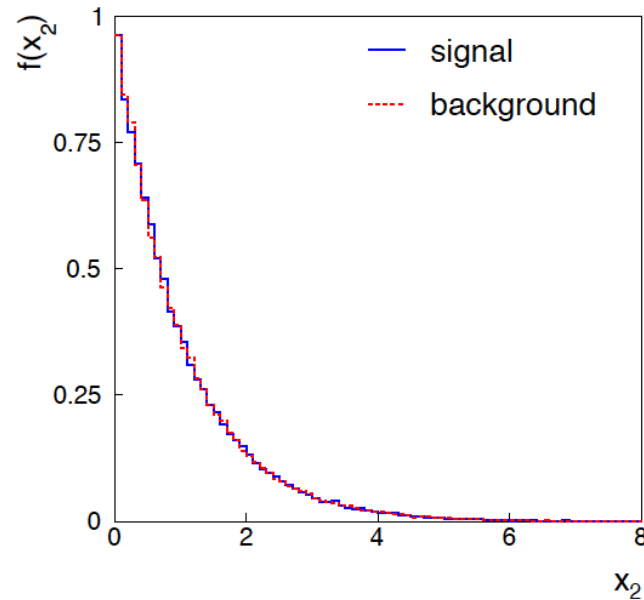
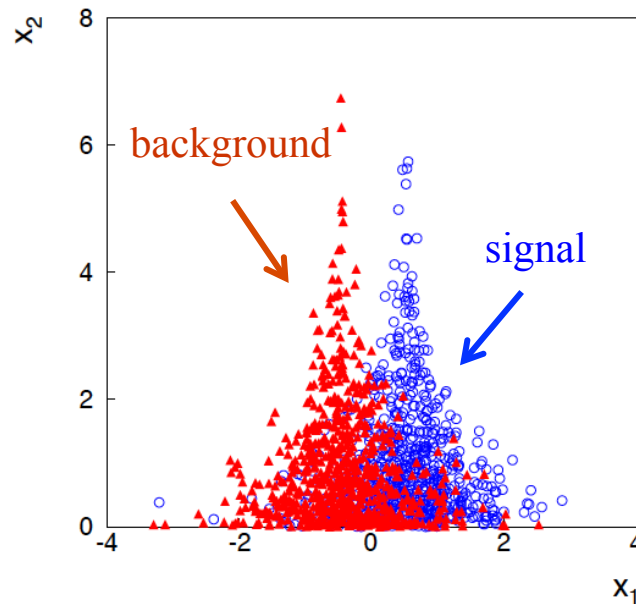
$f(x_1|x_2) \sim$  Gaussian, different means for s/b,  
Gaussians have same  $\sigma$ , which depends on  $x_2$ ,  
 $f(x_2) \sim$  exponential, same for both s and b,  
 $f(x_1, x_2) = f(x_1|x_2)f(x_2)$ :

$$f(x_1, x_2|s) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1 - \mu_s)^2 / 2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$

$$f(x_1, x_2|b) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1 - \mu_b)^2 / 2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$

$$\sigma(x_2) = \sigma_0 e^{-x_2/\xi}$$

# Joint and marginal distributions of $x_1$ , $x_2$



Distribution  $f(x_2)$  same for s, b.

So does  $x_2$  help discriminate between the two event types?

# Likelihood ratio for 2D example

Neyman-Pearson lemma says best critical region is determined by the likelihood ratio:

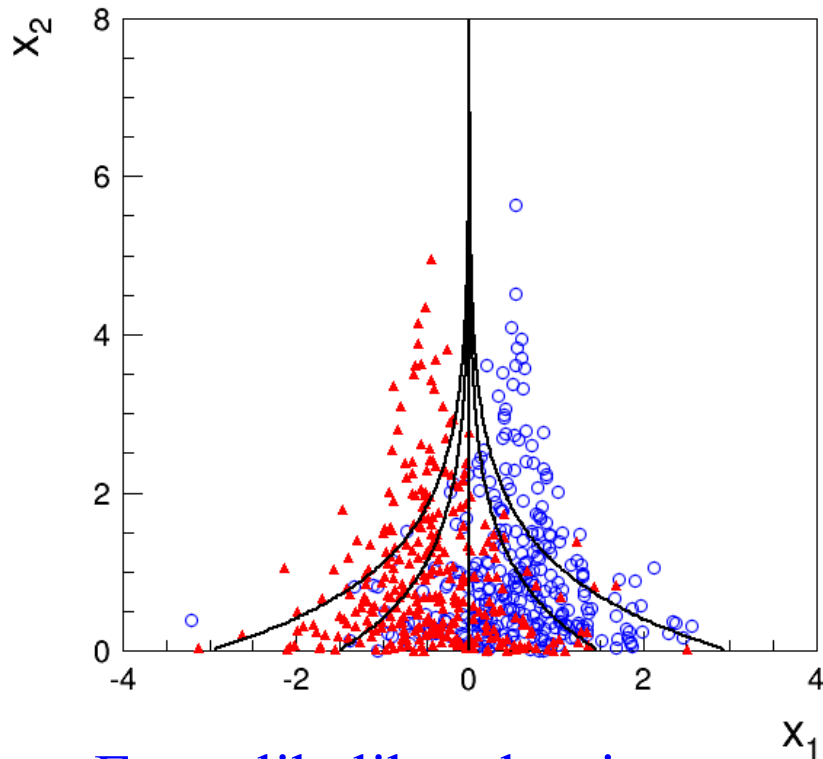
$$t(x_1, x_2) = \frac{f(x_1, x_2 | s)}{f(x_1, x_2 | b)}$$

Equivalently we can use any monotonic function of this as a test statistic, e.g.,

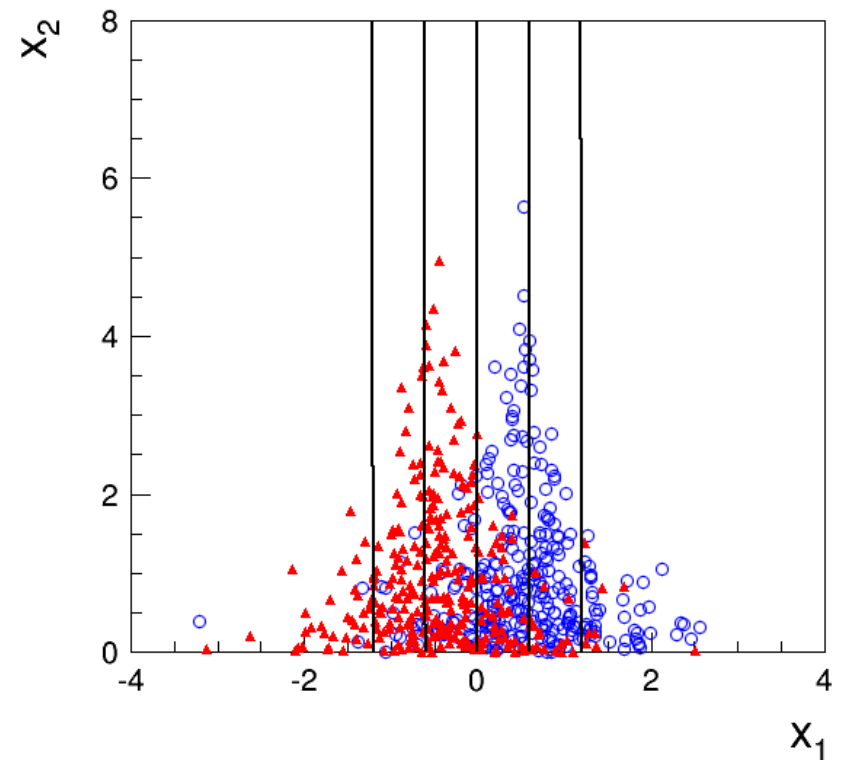
$$\ln t = \frac{\frac{1}{2}(\mu_b^2 - \mu_s^2) + (\mu_s - \mu_b)x_1}{\sigma_0^2 e^{-2x_2/\xi}}$$

Boundary of optimal critical region will be curve of constant  $\ln t$ , and this depends on  $x_2$ !

# Contours of constant MVA output



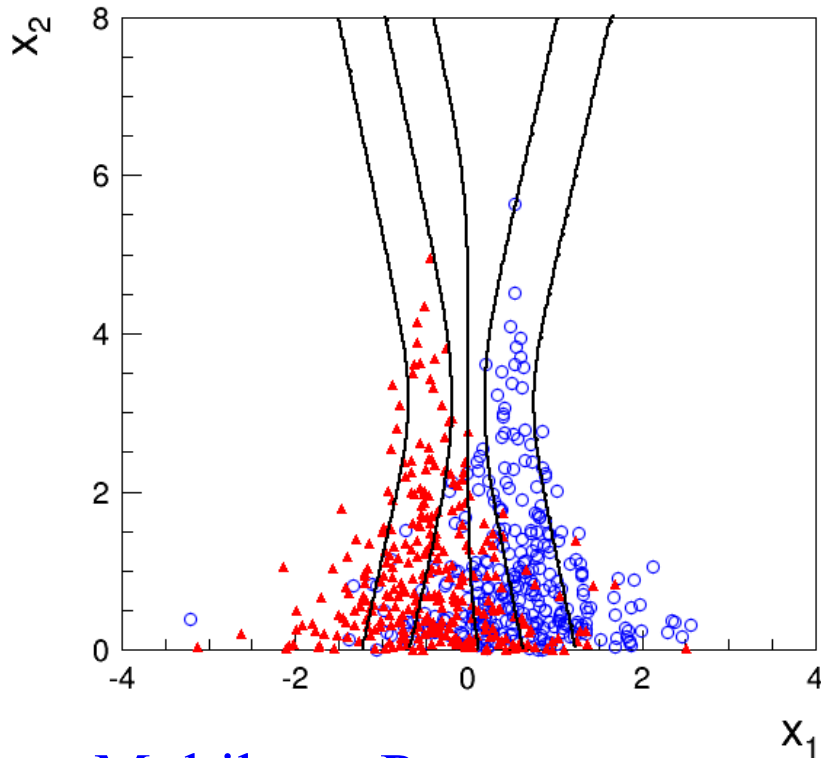
Exact likelihood ratio



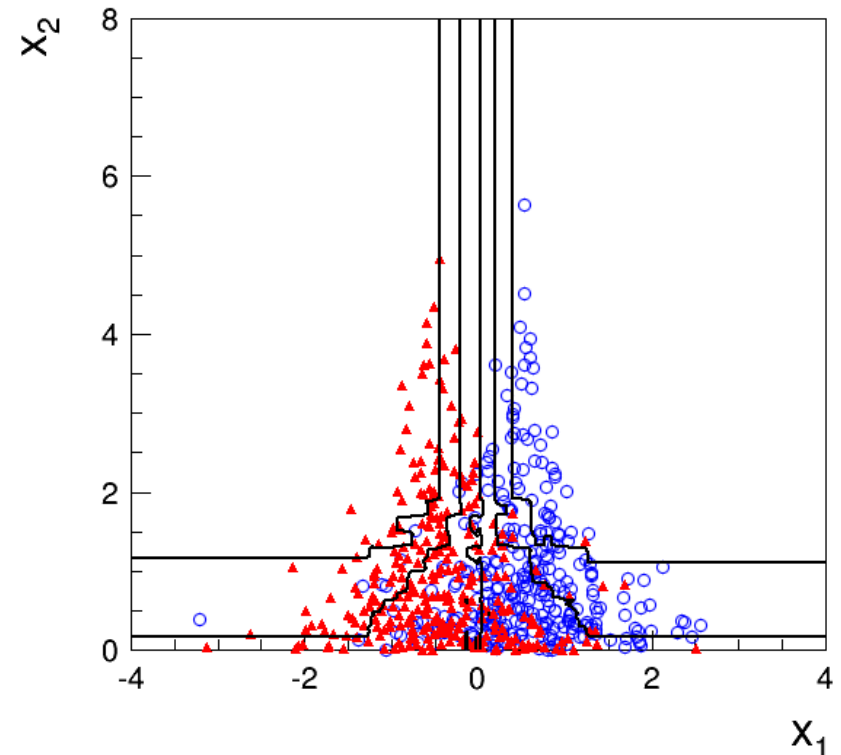
Fisher discriminant



# Contours of constant MVA output



Multilayer Perceptron  
1 hidden layer with 2 nodes

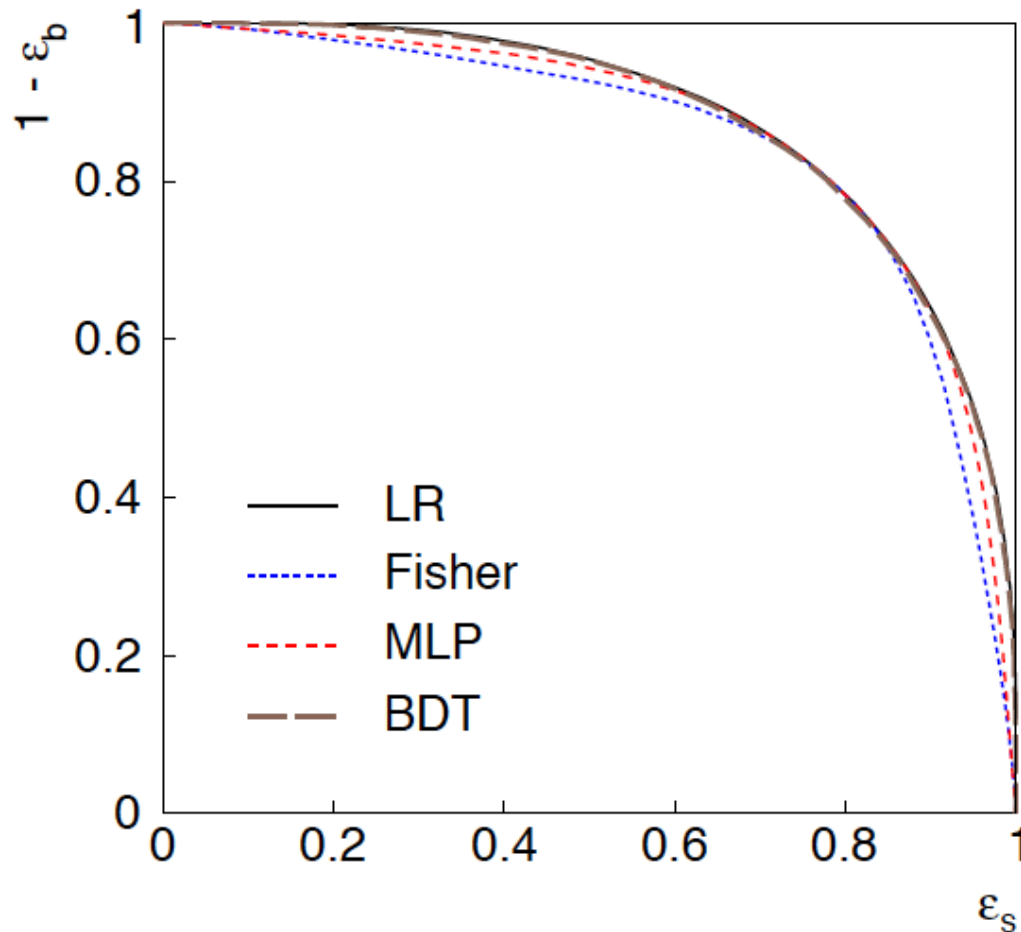


Boosted Decision Tree  
200 iterations (AdaBoost)

Training samples:  $10^5$  signal and  $10^5$  background events



# ROC curve



ROC = “receiver operating characteristic” (term from signal processing).

Shows (usually) background rejection ( $1 - \epsilon_b$ ) versus signal efficiency  $\epsilon_s$ .

Higher curve is better; usually analysis focused on a small part of the curve.

# Statistics in Run II

In the last decade there has been an increasing acceptance/popularity of multivariate methods, with many new developments entering from Machine Learning.

Run II will also face the same challenges as Run I, needing to quantify e.g. discovery significance and exclusion limits. Still need to construct accurate models and make accurate estimates of experimental sensitivity e.g. to optimize analyses.

There is a new particle to study! Having discovered the Higgs, people will now want to measure its properties, e.g., differential distributions ( $\rightarrow$  unfolding).

There is increased pressure/motivation to fully exploit the hard-won data, hence the need to report enough information to allow combinations and e.g. future refinements of theoretical uncertainties.

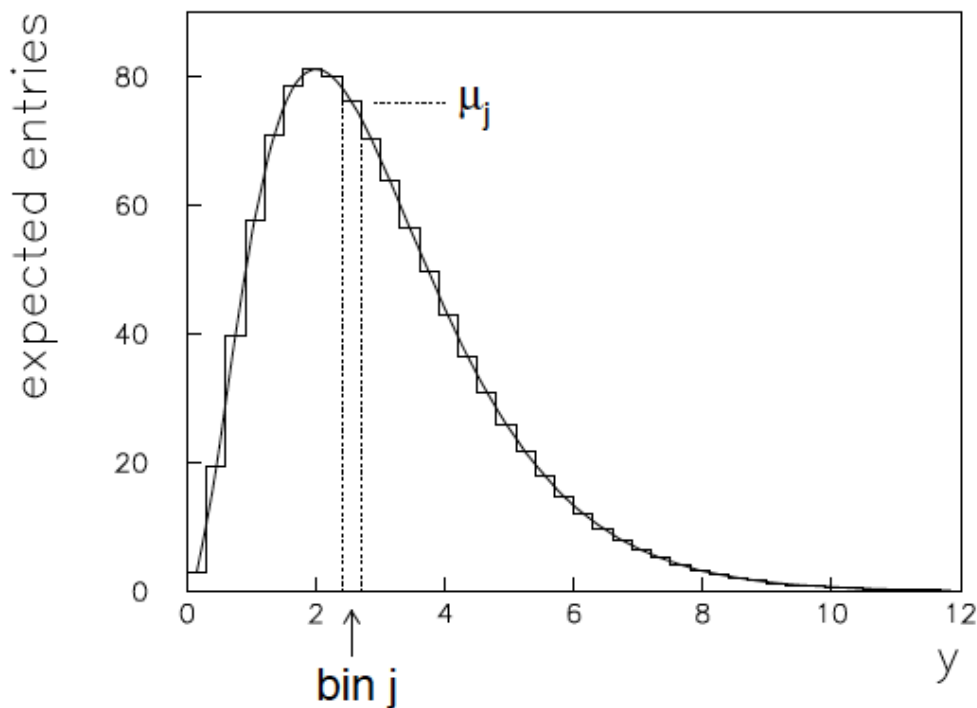
# Extra slides

# Unfolding

Consider a random variable  $y$ , goal is to determine pdf  $f(y)$ .

If parameterization  $f(y;\theta)$  known, find e.g. ML estimators  $\hat{\theta}$ .

If no parameterization available, construct histogram:



$$p_j = \int_{\text{bin } j} f(y) dy$$

$$\mu_j = \mu_{\text{tot}} p_j$$

↖ “true” histogram

New goal: construct estimators for the  $\mu_j$  (or  $p_j$ ).

# Migration

Effect of measurement errors:  $y$  = true value,  $x$  = observed value,  
migration of entries between bins,  
 $f(y)$  is ‘smeared out’, peaks broadened.

$$f_{\text{meas}}(x) = \int R(x|y) f_{\text{true}}(y) dy$$



discretize: data are  $\mathbf{n} = (n_1, \dots, n_N)$

$$\nu_i = E[n_i] = \sum_{j=1}^M R_{ij} \mu_j, \quad i = 1, \dots, N$$

$$R_{ij} = P(\text{observed in bin } i \mid \text{true in bin } j)$$

response  
matrix



Note  $\mu$ ,  $\nu$  are constants;  $\mathbf{n}$  subject to statistical fluctuations.

# Efficiency, background

Sometimes an event goes undetected:

$$\begin{aligned}\sum_{i=1}^N R_{ij} &= \sum_{i=1}^N P(\text{observed in bin } i \mid \text{true value in bin } j) \\ &= P(\text{observed anywhere} \mid \text{true value in bin } j) \\ &= \varepsilon_j \quad \longleftarrow \text{efficiency}\end{aligned}$$

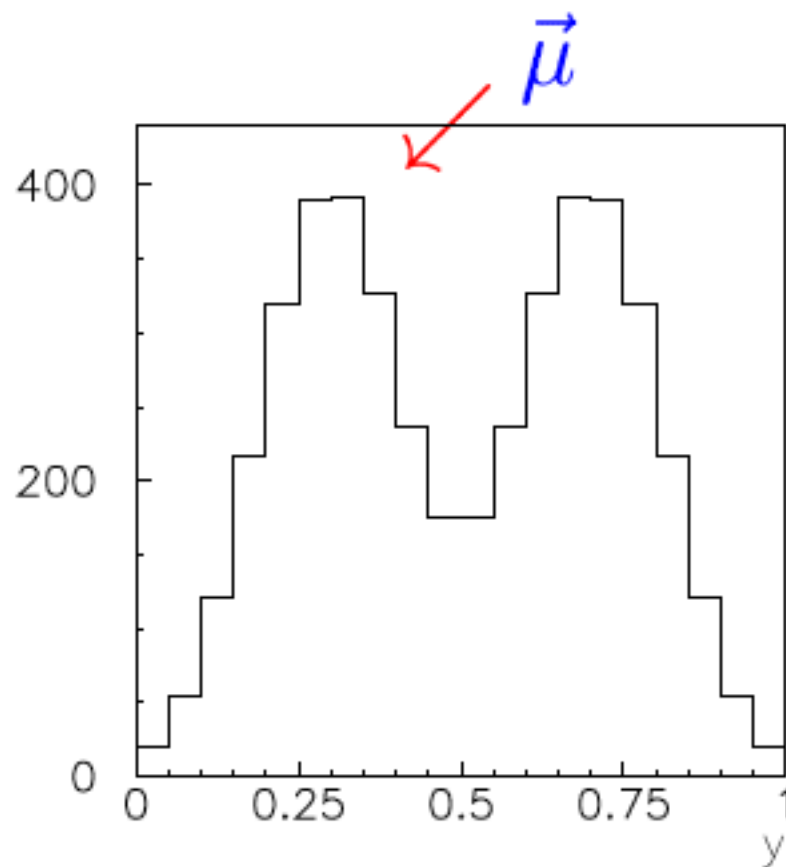
Sometimes an observed event is due to a background process:

$$\nu_i = \sum_{j=1}^M R_{ij} \mu_j + \beta_i$$

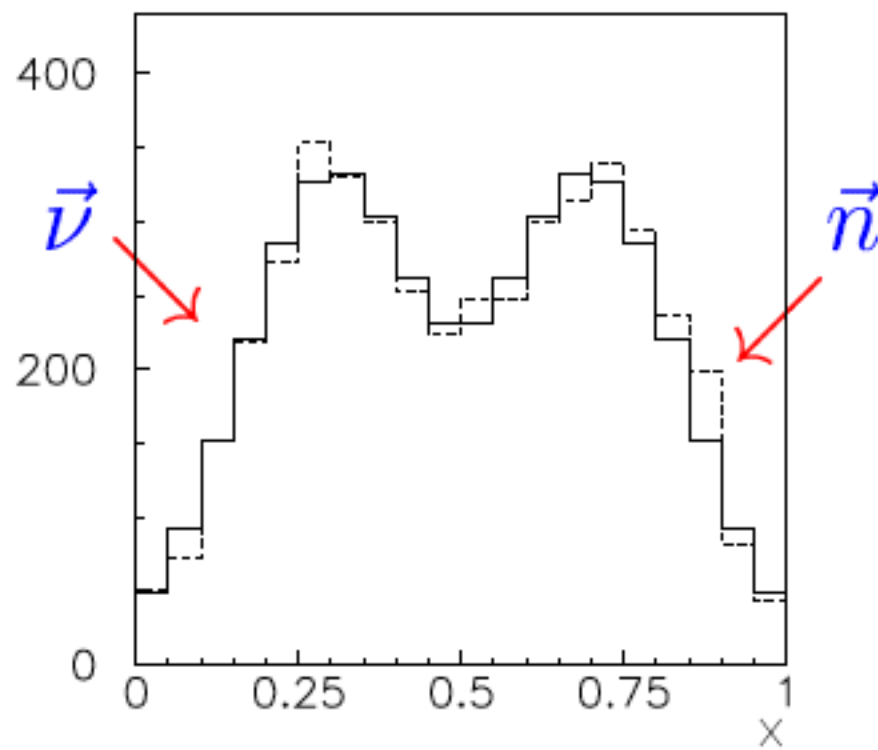
$\beta_i$  = expected number of background events in *observed* histogram.

For now, assume the  $\beta_i$  are known.

# The basic ingredients



“true”



“observed”

# Summary of ingredients

‘true’ histogram:  $\mu = (\mu_1, \dots, \mu_M), \quad \mu_{\text{tot}} = \sum_{j=1}^M \mu_j$

probabilities:  $\mathbf{p} = (p_1, \dots, p_M) = \mu / \mu_{\text{tot}}$

expectation values for observed histogram:  $\nu = (\nu_1, \dots, \nu_N)$

observed histogram:  $\mathbf{n} = (n_1, \dots, n_N)$

response matrix:  $R_{ij} = P(\text{observed in bin } i \mid \text{true in bin } j)$

efficiencies:  $\varepsilon_j = \sum_{i=1}^N R_{ij}$

expected background:  $\beta = (\beta_1, \dots, \beta_N)$

These are related by:

$$E[\mathbf{n}] = \nu = R\mu + \beta$$



# Maximum likelihood (ML) estimator from inverting the response matrix

Assume  $\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$  can be inverted:  $\boldsymbol{\mu} = R^{-1}(\boldsymbol{\nu} - \boldsymbol{\beta})$

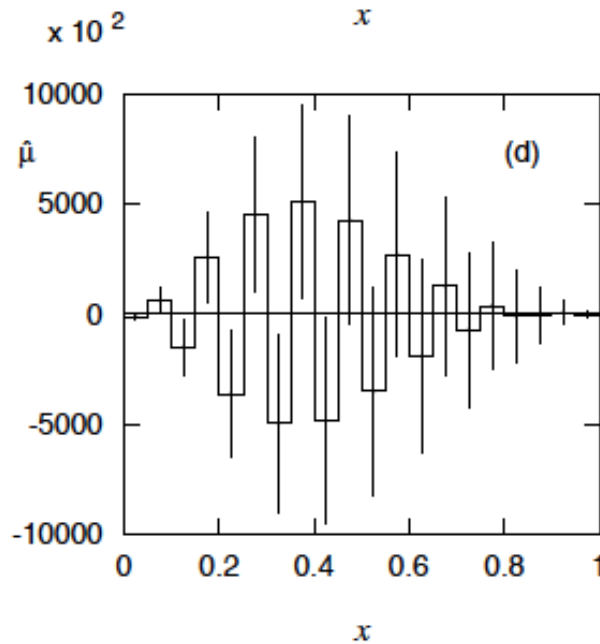
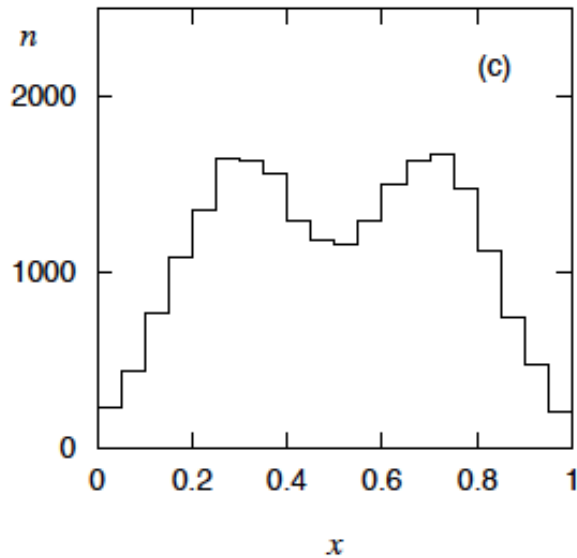
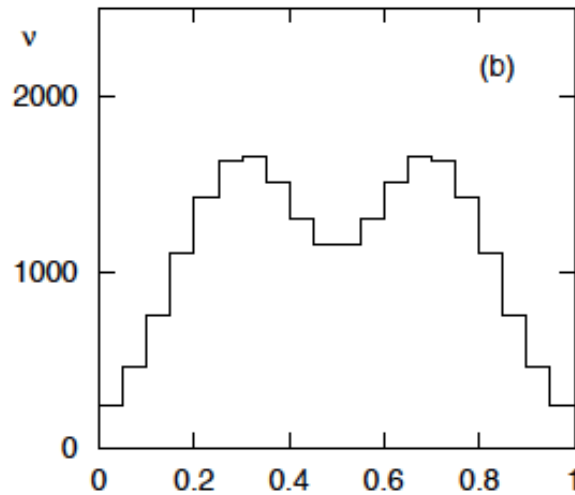
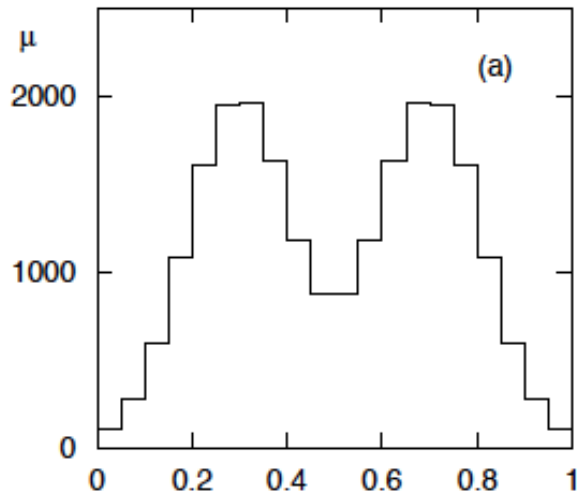
Suppose data are independent Poisson:  $P(n_i; \nu_i) = \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$

So the log-likelihood is  $\ln L(\boldsymbol{\mu}) = \sum_{i=1}^N (n_i \ln \nu_i - \nu_i)$

ML estimator is  $\hat{\boldsymbol{\nu}} = \mathbf{n}$

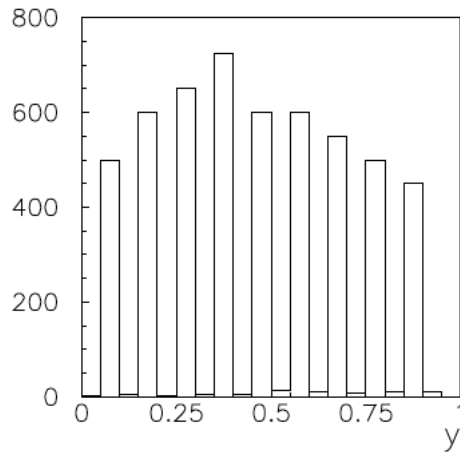
$$\longrightarrow \hat{\boldsymbol{\mu}} = R^{-1}(\mathbf{n} - \boldsymbol{\beta})$$

# Example with ML solution



Catastrophic failure???

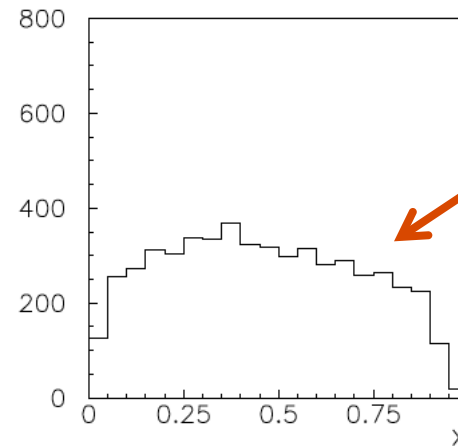
# What went wrong?



Suppose  $\mu$  really had a lot of fine structure.

←  $\vec{\mu}$

Applying  $R$  washes this out, but leaves a residual structure:



←  $\vec{\nu} = R\vec{\mu}$

Applying  $R^{-1}$  to  $\vec{\nu}$  puts the fine structure back:  $\vec{\mu} = R^{-1}\vec{\nu}$ .

But we don't have  $\nu$ , only  $n$ .  $R^{-1}$  “thinks” fluctuations in  $n$  are the residual of original fine structure, puts this back into  $\hat{\mu}$ .

# ML solution revisited

For Poisson data the ML estimators are unbiased:

$$E[\hat{\mu}] = R^{-1}(E[\mathbf{n}] - \beta) = \mu$$

Their covariance is:

$$\begin{aligned} U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j] &= \sum_{k,l=1}^N (R^{-1})_{ik} (R^{-1})_{jl} \text{cov}[n_k, n_l] \\ &= \sum_{k=1}^N (R^{-1})_{ik} (R^{-1})_{jk} \nu_k \end{aligned}$$

(Recall these statistical errors were huge for the example shown.)

## ML solution revisited (2)

The information inequality gives for unbiased estimators the minimum (co)variance bound:

$$(U^{-1})_{kl} = -E \left[ \frac{\partial^2 \log L}{\partial \mu_k \partial \mu_l} \right] = \sum_{i=1}^N \frac{R_{ik} R_{il}}{\nu_i}$$

invert  $\rightarrow U_{ij} = \sum_{k=1}^N (R^{-1})_{ik} (R^{-1})_{jk} \nu_k$

This is the same as the actual variance! I.e. ML solution gives smallest variance among all unbiased estimators, even though this variance was huge.

In unfolding one must accept some bias in exchange for a (hopefully large) reduction in variance.

# Correction factor method

Use equal binning for  $\vec{\mu}$ ,  $\vec{\nu}$  and take  $\hat{\mu}_i = C_i(n_i - \beta_i)$ , where

$$C_i = \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} \quad \nu_i^{\text{MC}} \text{ and } \mu_i^{\text{MC}} \text{ from Monte Carlo simulation (no background).}$$


$$U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j] = C_i^2 \text{cov}[n_i, n_j]$$

Often  $C_i \sim O(1)$  so statistical errors far smaller than for ML.

But the bias  $b_i = E[\hat{\mu}_i] - \mu_i$  is

$$b_i = \left( \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} - \frac{\mu_i}{\nu_i^{\text{sig}}} \right)$$

Nonzero bias unless MC = Nature.


$$\nu_i^{\text{sig}} = \nu_i - \beta_i$$

## Reality check on the statistical errors

Suppose for some bin  $i$  we have:

$$C_i = 0.1 \qquad \beta_i = 0 \qquad n_i = 100$$

$$\longrightarrow \hat{\mu}_i = C_i n_i = 10 \qquad \sigma_{\hat{\mu}_i} = C_i \sqrt{n_i} = 1.0 \qquad (10\% \text{ stat. error})$$

But according to the estimate, only 10 of the 100 events found in the bin belong there; the rest spilled in from outside.

How can we have a 10% measurement if it is based on only 10 events that really carry information about the desired parameter?

# Discussion of correction factor method

As with all unfolding methods, we get a reduction in statistical error in exchange for a bias; here the bias is difficult to quantify (difficult also for many other unfolding methods).

The bias should be small if the bin width is substantially larger than the resolution, so that there is not much bin migration.

So if other uncertainties dominate in an analysis, correction factors may provide a quick and simple solution (a “first-look”).

Still the method has important flaws and it would be best to avoid it.



# Regularized unfolding

Consider ‘reasonable’ estimators such that for some  $\Delta \log L$ ,

$$\log L(\vec{\mu}) \geq \log L_{\max} - \Delta \log L$$

Out of these estimators, choose the ‘smoothest’, by maximizing

$$\Phi(\vec{\mu}) = \alpha \log L(\vec{\mu}) + S(\vec{\mu}),$$

$S(\vec{\mu})$  = regularization function (measure of smoothness),

$\alpha$  = regularization parameter (choose to give desired  $\Delta \log L$ )

## Regularized unfolding (2)

In addition require  $\sum_{i=1}^N \nu_i = \sum_{i,j} R_{ij} \mu_j = n_{\text{tot}}$ , i.e. maximize

$$\varphi(\vec{\mu}, \lambda) = \alpha \log L(\vec{\mu}) + S(\vec{\mu}) + \lambda \left[ n_{\text{tot}} - \sum_{i=1}^N \nu_i \right]$$

where  $\lambda$  is a Lagrange multiplier,  $\partial\varphi/\partial\lambda = 0 \rightarrow \sum_{i=1}^N \nu_i = n_{\text{tot}}$ .

$\alpha = 0$  gives smoothest solution (ignores data!),

$\alpha \rightarrow \infty$  gives ML solution (variance too large).

We need: regularization function  $S(\vec{\mu})$ ,

a prescription for setting  $\alpha$ .

# Tikhonov regularization

Take measure of smoothness = mean square of  $k$ th derivative,

$$S[f_{\text{true}}(y)] = - \int \left( \frac{d^k f_{\text{true}}(y)}{dy^k} \right)^2 dy, \text{ where } k = 1, 2, \dots$$

If we use Tikhonov ( $k = 2$ ) with  $\log L = -\frac{1}{2}\chi^2$ ,

$$S(\boldsymbol{\mu}) = - \sum_{i=1}^{M-2} (-\mu_i + 2\mu_{i+1} - \mu_{i+2})^2$$

$$\varphi(\vec{\mu}, \lambda) = -\frac{\alpha}{2}\chi^2(\vec{\mu}) + S(\vec{\mu}) \quad \text{quadratic in } \mu_i,$$

→ setting derivatives of  $\varphi$  equal to zero gives linear equations.

Programs: RUN (Blobel); SVD (Hoecker & Kartvelishvili)

# SVD implementation of Tikhonov unfolding

A. Höcker, V. Kartvelishvili, NIM A**372** (1996) 469.

→ program **GURU**, see [www.hep.man.ac.uk/~vato](http://www.hep.man.ac.uk/~vato)

Minimizes  $\chi^2 + \tau \sum_i [(\mu_{i+1} - \mu_i) - (\mu_i - \mu_{i-1})]^2$

Numerical implementation → Singular Value Decomposition

Recommendations for setting regularization parameter  $\tau$ :

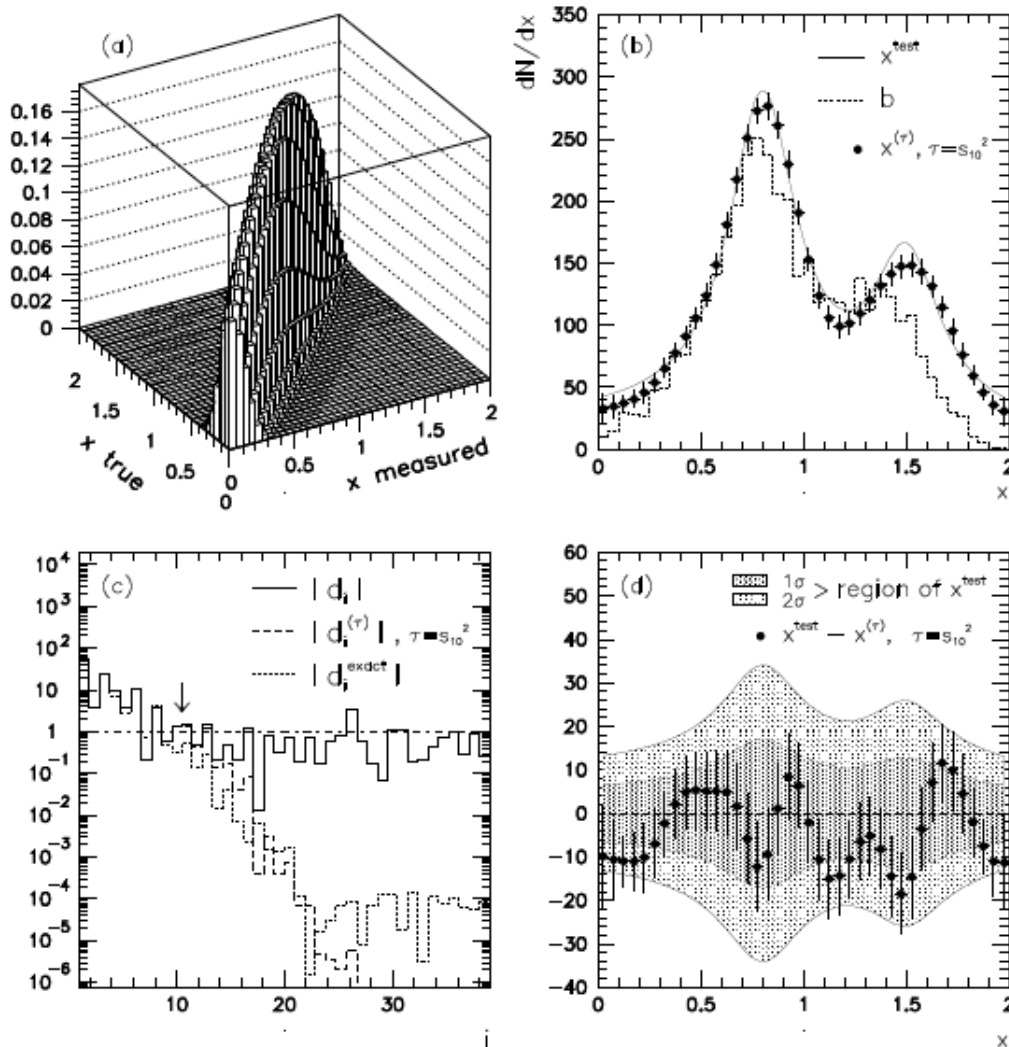
Transform variables so estimators are  $N(0, 1)$ ,

number significantly different from zero → prescription for  $\tau$ ;

or base choice on unfolding of test distributions.

# SVD example

A. Höcker, V. Kartvelishvili, NIM A**372** (1996) 469.



# Regularization function based on entropy

Shannon entropy of a set of probabilities is

$$H = - \sum_{i=1}^M p_i \log p_i$$

All  $p_i$  equal  $\rightarrow$  maximum entropy (maximum smoothness)

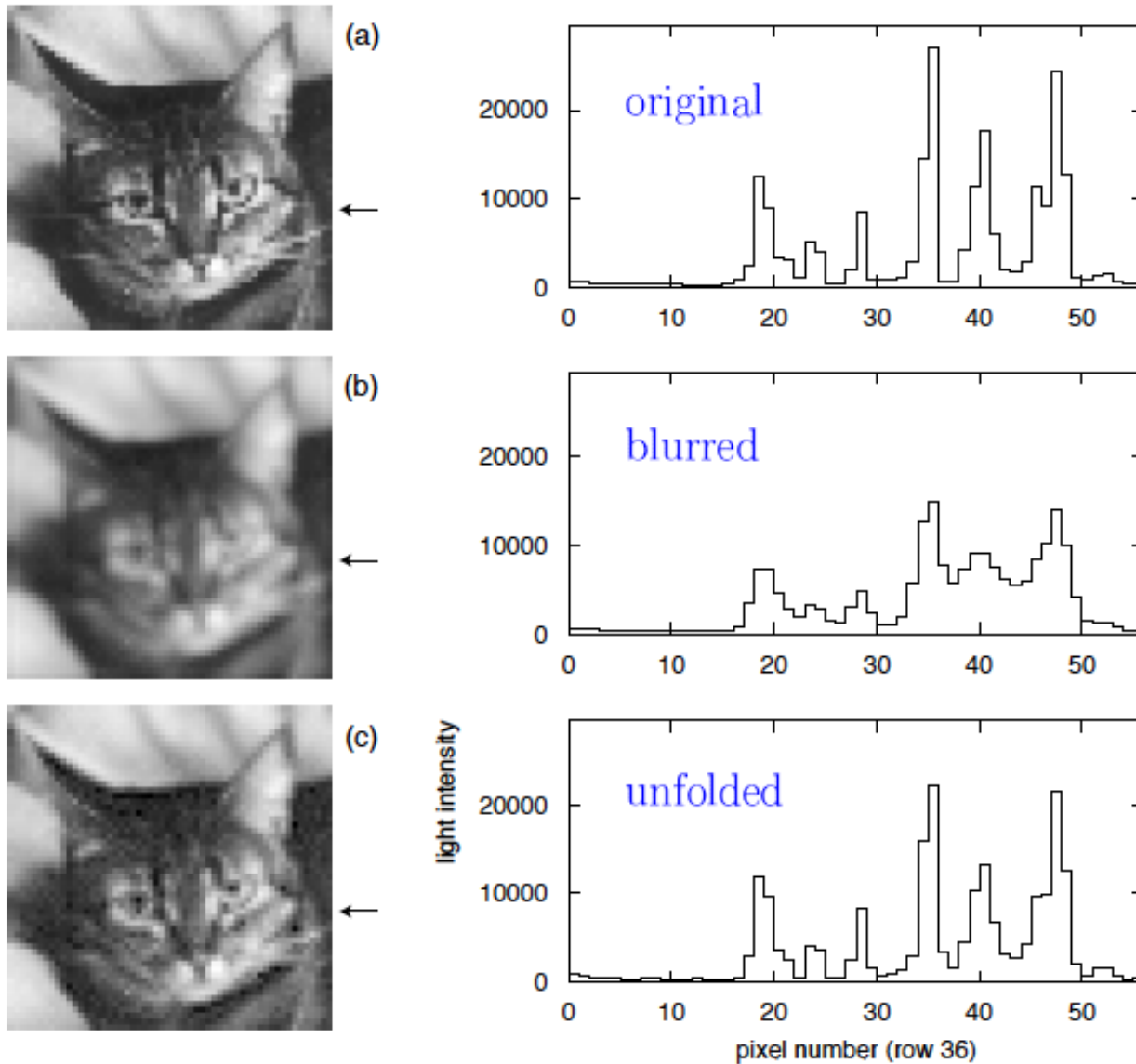
One  $p_i = 1$ , all others = 0  $\rightarrow$  minimum entropy

Use entropy as regularization function,

$$S(\vec{\mu}) = H(\vec{\mu}) = - \sum_{i=1}^M \frac{\mu_i}{\mu_{\text{tot}}} \log \frac{\mu_i}{\mu_{\text{tot}}}$$

$$\propto \log(\text{number of ways to arrange } \mu_{\text{tot}} \text{ entries in } M \text{ bins})$$

# Example of entropy-based unfolding



# Iterative unfolding (“Bayesian”)

G. D’Agostini, NIM A **362** (1995) 487

Goal is to estimate probabilities:  $\mathbf{p} = (p_1, \dots, p_M)$

For initial guess take e.g.  $p_i = 1/M$

Initial estimators for  $\mu$  are  $\hat{\mu}_0 = n_{\text{tot}} \mathbf{p}_0$ ,

Update according to the rule

$$\hat{\mu}_i = \frac{1}{\varepsilon_i} \sum_{j=1}^N P(\text{true value in bin } i | \text{ found in bin } j) n_j$$

uses Bayes’ theorem here


$$= \frac{1}{\varepsilon_i} \sum_{j=1}^N \left( \frac{R_{ij} p_i}{\sum_k R_{jk} p_k} \right) n_j$$

Continue until solution stable  
using e.g.  $\chi^2$  test with previous  
iteration.



# Estimating systematic uncertainty

We know that unfolding introduces a bias, but quantifying this (including correlations) can be difficult.

Suppose a model predicts a spectrum

$$f(y; \theta) \sim 1/y^\theta \rightarrow \mu(\theta)$$

A priori suppose e.g.  $\theta \approx 8 \pm 2$ . More precisely, assign prior  $\pi(\theta)$ . Propagate this into a systematic covariance for the unfolded spectrum:

$$U_{ij}^{(\theta)} = \int (\hat{\mu}_i - \mu_i(\theta))(\hat{\mu}_j - \mu_j(\theta)) \pi(\theta) d\theta$$

(Typically large positive correlations.)

Often in practice, one doesn't have  $\pi(\theta)$  but rather a few models that differ in spectrum. Not obvious how to convert this into a meaningful covariance for the unfolded distribution.

# Stat. and sys. errors of unfolded solution

In general the statistical covariance matrix of the unfolded estimators is not diagonal; need to report full

$$U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j]$$

But unfolding necessarily introduces biases as well, corresponding to a systematic uncertainty (also correlated between bins).

This is more difficult to estimate. Suppose, nevertheless, we manage to report both  $U_{\text{stat}}$  and  $U_{\text{sys}}$ .

To test a new theory depending on parameters  $\theta$ , use e.g.

$$\chi^2(\theta) = (\mu(\theta) - \hat{\mu})^T (U_{\text{stat}} + U_{\text{sys}})^{-1} (\mu(\theta) - \hat{\mu})$$

Mixes frequentist and Bayesian elements; interpretation of result can be problematic, especially if  $U_{\text{sys}}$  itself has large uncertainty.

# Folding

Suppose a theory predicts  $f(y) \rightarrow \mu$  (may depend on parameters  $\theta$ ).

Given the response matrix  $R$  and expected background  $\beta$ , this predicts the expected numbers of observed events:

$$\nu = R\mu + \beta$$

From this we can get the likelihood, e.g., for Poisson data,

$$L(\mathbf{n}|\nu) = \prod_{i=1}^N \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

And using this we can fit parameters and/or test, e.g., using the likelihood ratio statistic

$$q = -2 \ln \frac{L(\mathbf{n}|\nu)}{L(\mathbf{n}|\hat{\nu})} \sim \chi_N^2$$

## Versus unfolding

If we have an unfolded spectrum and full statistical and systematic covariance matrices, to compare this to a model  $\mu$  compute likelihood

$$L(\hat{\mu}|\mu) \sim e^{-\chi^2/2}$$

where

$$\chi^2 = (\mu - \hat{\mu})^T (U_{\text{stat}} + U_{\text{sys}})^{-1} (\mu - \hat{\mu})$$

Complications because one needs estimate of systematic bias  $U_{\text{sys}}$ ; also assumes data  $\sim$  Gaussian (and difficult to avoid need for this approximation).

Quadratic sum of stat. and sys. errors makes interpretation difficult.

Even beyond these issues, a test based on the unfolded distribution will not in general be as optimal as one obtained through folding the theory and comparing it to the uncorrected data.

# ML solution again

From the standpoint of testing a theory or estimating its parameters, the ML solution, despite catastrophically large errors, is equivalent to using the uncorrected data (same information content).

There is no bias (at least from unfolding), so use

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}}_{\text{ML}})^T U_{\text{stat}}^{-1} (\boldsymbol{\mu}(\boldsymbol{\theta}) - \hat{\boldsymbol{\mu}}_{\text{ML}})$$

The estimators of  $\boldsymbol{\theta}$  should have close to optimal properties: zero bias, minimum variance.

The corresponding estimators from any unfolded solution cannot in general match this.

Ditto for power of statistical tests; ML is optimal.

Crucial point is to use full covariance, not just diagonal errors.

# Summary/discussion

Unfolding can be a minefield and not necessary if goal is to compare model with theory.

Even comparison of uncorrected distribution with *future* theories not a problem, as long as it is reported together with the expected background and response matrix.

In practice complications because these ingredients have uncertainties, and they must be reported as well.

Unfolding useful for getting an actual estimate of the distribution we think we've measured; can e.g. compare with CMS.

Model test using unfolded distribution should take account of the (correlated) bias introduced by the unfolding procedure.

Unfolded distributions easier to work with in cases where bias introduced is small compared to other uncertainties (and thus can be neglected).