Can we go beyond Wilks theorem for significance calculation?

Estimating p-values with importance sampling

Francisco Matorras
IFCA
Instituto de Física de Cantabria (Santander)

Introduction & Motivation

Motivation

- We are all familiar with the 5σ convention for discoveries and its "issues"
  - It is almost always taken as a sharp cut
  - Often struggling to reach that 5 or discussing if it is 4.9 or 5.0
- But almost always p-value calculation is based on Wilks theorem
  - Often without guarantee the conditions are fulfilled or that asymptotic regime can be trusted (to probs~10^{-7})
- The alterative is running toys
  - but O(10^8) needed, usually impractical
- Most interpretations limited to *local* p-values

Is there a better approach?

- Quote from "Data Analysis in High Energy Physics" O. Behnke, K. Kröninger, G. Schott, and T. Schörner-Sadenius (Ed. WILEY-VCH)
  - "*For computing very small p-values with reasonable precision, a large number of MC iterations is required. In that case the tail of the distribution of q is most important.* ***The procedure above may be improved by resorting to techniques such as importance sampling which concentrates on generating Monte Carlo datasets that lie in those tails.***"
- But how to concentrate events in those tails?
- Some ideas already presented at a previous Quark Confinement:
  - promising, but didn't always work (biased results in some cases)
  - Better understanding today

Importance sampling

Importance sampling

- Note: theory considerations based on
  - [1] "Simulation and the Monte Carlo Method" by Reuven Y. Rubinstein Dirk P. Kroese, Ed Wiley
- The basic idea behind IP,
  - Sample from a more convenient pdf
  - Assign weights so that the expectation values asymptotically converge to the desired value
  - If you play your cards, it will converge faster (i.e., need less toys)

Importance sampling

- In general, given a pdf $\rho(\vec{x})$, we want to estimate the expectation **h** of an observable $H(\vec{x})$
  - $h = E[H(\vec{x})] = \int H(\vec{x})\,\rho(\vec{x})d\vec{x}$
  - or with sampling $h = \sum H(\vec{x_i})$ with $\vec{x_i}$ drawn from $\rho(\vec{x})$
- The importance sample trick
  - Use a *better* pdf $\tilde{\rho}(\vec{x})$ and reweight: $h = \int H(\vec{x})\frac{\rho(\vec{x})}{\tilde{\rho}(\vec{x})}\tilde{\rho}(\vec{x})d\vec{x} = \int H(\vec{x})W(\vec{x})\tilde{\rho}(\vec{x})d\vec{x} = E_{\sim}[H(\vec{x})W(\vec{x})]$
  - We average or sample over $\tilde{\rho}(\vec{x})$, correcting with a weight $W(\vec{x}) = \frac{\rho(\vec{x})}{\tilde{\rho}(\vec{x})}$
- Given some conditions (basically avoid infinites) you should get correct results asymptotically independently of $\tilde{\rho}(\vec{x})$
- … but not all improve the sampling

Importance sampling II

- An optimal (in the sense of minimizing the variance of the estimation) can be derived [1]:
  - $\rho^*(\vec{x}) = \frac{H(\vec{x})\rho(\vec{x})}{\int H(\vec{x})\rho(\vec{x})d\vec{x}}$
  - But useless ☹, the integral in the denominator is the quantity we want to get!
- In practice, one can instead use a family of pdf, $\tilde{\rho}(\vec{x},\vec{\alpha})$
- An optimal $\tilde{\rho}$ can be obtained minimizing the variance, look for the $\vec{\alpha}$ which provides a smaller variance on the p-value estimation
  - Minimize the variance as a function of $\vec{\alpha}$, now a parametric minimization
  - Or maximize cross entropy w.r.t. $\rho^*(\vec{x})$
  - Note, we do not guarantee the global best, but if the pdf is chosen wisely, we can still gain a lot

Importance sampling in discoveries

Francisco Matorras, IFCA, Spain

How to relate IP to p-values?

- We have an H_0 (background) driven by $\rho(\vec{x})$
- We have a H_1 (signal) driven by $\rho'(\vec{x})$, usually in a parametric way and such H_0 is contained $\rho'(\vec{x}) = \rho(\vec{x}, \vec{\alpha})$ and very often just depending on a signal strength $\rho'(\vec{x}) = \rho(\vec{x}, \mu)$ such that μ=0 means no signal
- We define a test statistic based on the likelihood ratio

$$q(\vec{x}) = -2\log\left(\frac{\rho(\vec{x}|\mu = 0)}{\rho(\vec{x}|\mu = \mu best)}\right)$$

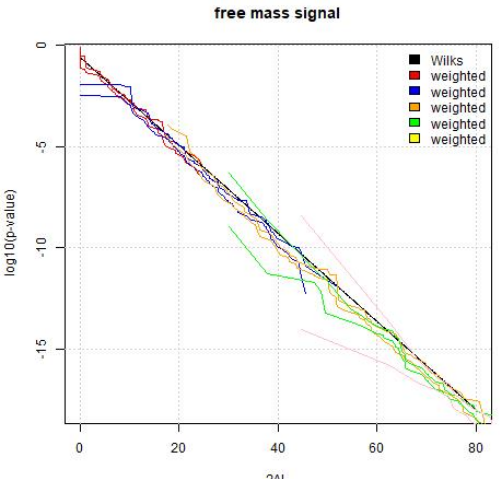
- And have an observed data $\vec{x}_0$ with $q_0 = q(\vec{x}_0)$
- P-value is defined as $p = \int_{q(\vec{x})>q_0} \rho(\vec{x}) d\vec{x}$
  - *Is it sufficient to claim discovery?*

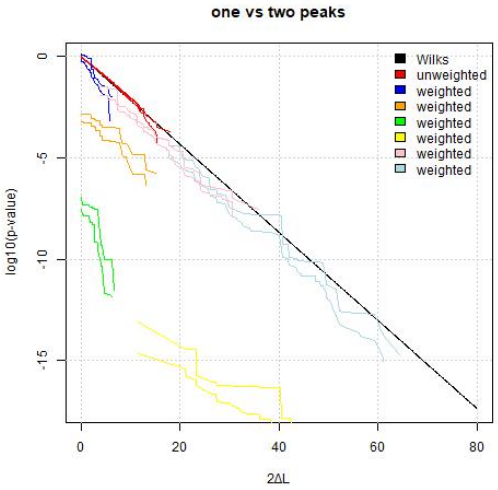
My proposal

- Turn this calculation into an IP problem
  - Calculate the expectation of $\mathrm{H}(\vec{x}) = \theta(\mathrm{q}(\vec{x}) - q_0)$
    - (θ is the step function 1 if argument positive 0 otherwise)
    - $p = \int_{q(\vec{x})>q_0} \rho(\vec{x}) d\vec{x} = \int_{\vec{x}} \theta(\mathrm{q}(\vec{x}) - q_0)\rho(\vec{x}) d\vec{x}$
    - $\mathrm{p} = {}^1/_M \sum_1^M \theta\left(\mathrm{q}\left(\vec{x}^{\,j}\right) - q_0\right)$
- Use as pdf family those from H_1, our S+B model (function of μ and possibly other params)
  - These pdfs are known and available for some points if derived from full MC
  - Take advantage from the fact that our LR resembles the weights
    - Suggest that minimizing the variance could be related to the MLE

My conjecture

- At the time of '21 (virtual) quark confinement I wondered:
- Why not using as sampling pdf, your signal-included model which better fits your data?
  - **Take as μ for importance sampling the one obtained from the MLE fit to the data μ_0**
  - Easy and convenient, you already have the model, either analytical of with simulation
  - examples showed impressive performance
  - Some mathematical arguments supported the idea (populating the tails, similar LR)
- But it didn't always work ☹
  - Struggling with some cases… now I understand when and why it does not work





One-sided, one POI

1 POI

Monotonic dependence of q

One POI with monotonic dependence

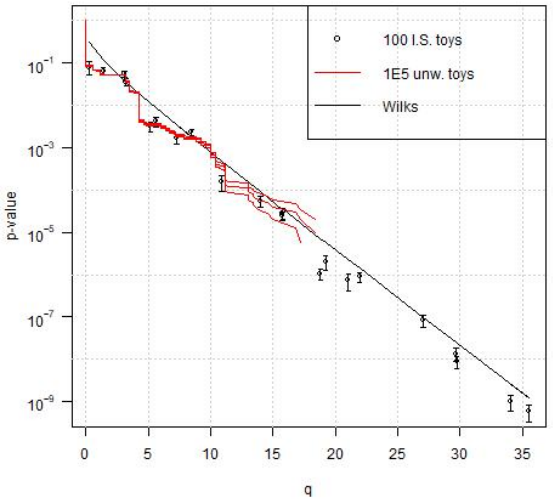
- Quite general case, for example one sided signal strength, $\mu>0$
- Monotonic in the sense that larger values of the parameter, imply larger q and lower p-values
- Conjecture can be proved to be exact, $\mu=\mu_0$ optimal, in this case provided that
  - it is *not far* from Wilks conditions or if $q(x|\mu)$ not *too wide*
- Improves variance several orders of magnitude
- Improvement still valid over a wide range of μ
  - Don't need the exact solution, for example can use the nearest MC sample

Algorithm

1. Fit the **data** to H_1, B+S, model and get μ_0
2. Generate a handful (M) of pseudoexperiments
   - $\{\vec{x}\}_j, \sim\rho(\vec{x}, \mu_0)$
   - Or sample from the available full MC sample closer to μ_0
3. Fit each set $\{\vec{x}\}_j$ get μ_j and q_j (repeat the full analysis on this pseudodata)
4. Calculate the weights of **each psexp** $w_j = \frac{\rho(\vec{x}^j, \mu=0)}{\rho(\vec{x}^j, \mu=\mu_0)}$
   1. if independent, ρ is factorized and become products of N **event** weights
5. Calculate p as $\frac{1}{M}\sum_{q_j>q_0} w_j$

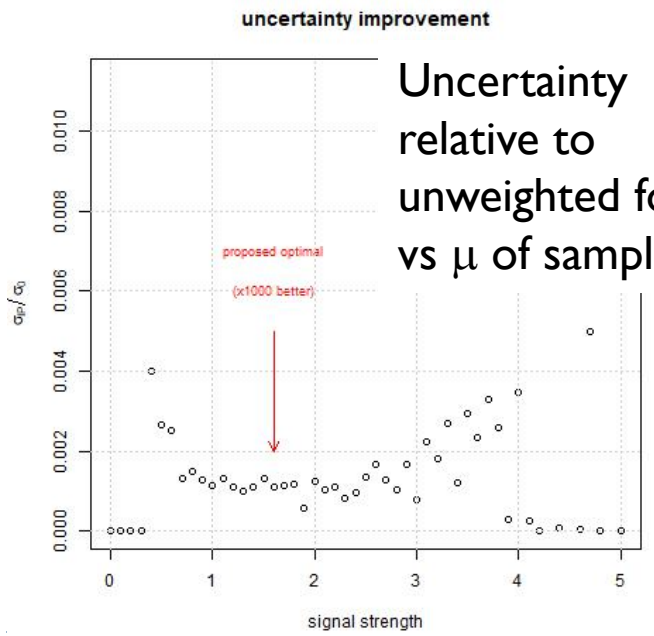
A few simple examples

- Next and following simplistic examples to illustrate the result
- Few, O(100), pseudo-experiments to *highlight* the power of the method
- P-value calculated with weighted events and compared to Wilks prediction and large-size unweighted toys, O(1e5)
- Uncertainty on weights calculated
- Shown as a function of q

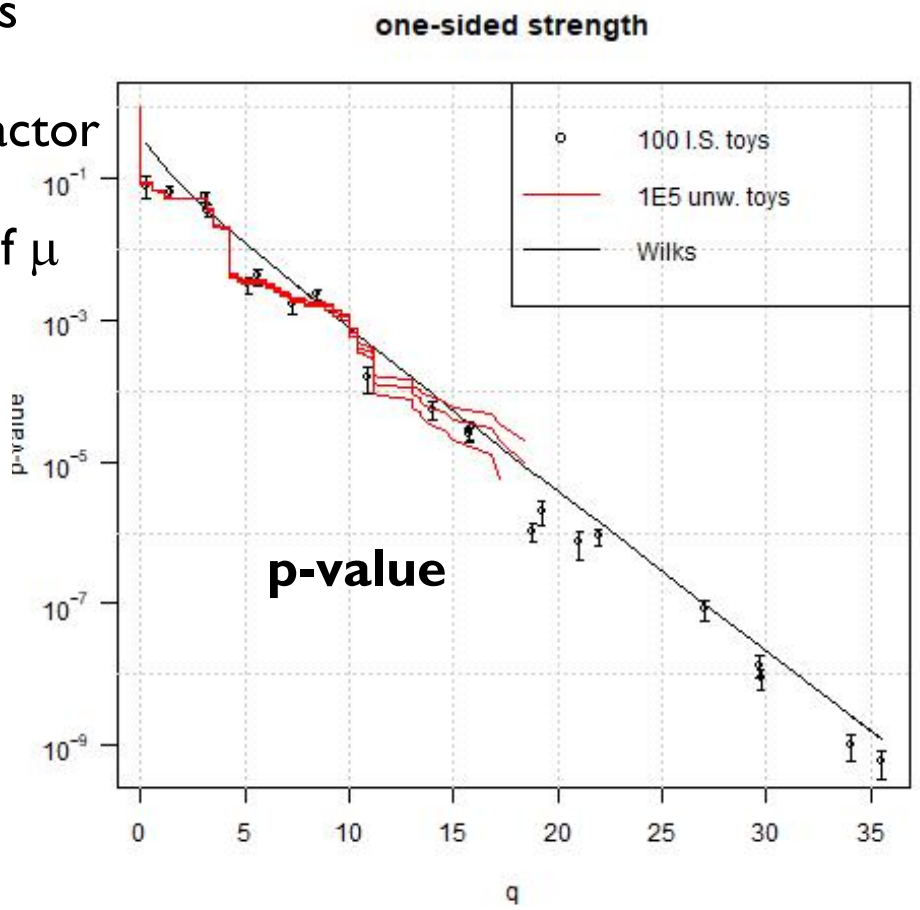


Binned low stat one sided

- Histograms; Poisson stat with $\lambda_i = b_i + \mu s_i$ μ positive
- With just 100 toys non-wilks wiggles perfectly reproduced
- For 5σ, uncertainty improved by a factor 1000 (ie, need factor 10^6 less toys)
- Big improvement for a wide range of μ



Uncertainty relative to unweighted for 5σ vs μ of sampling



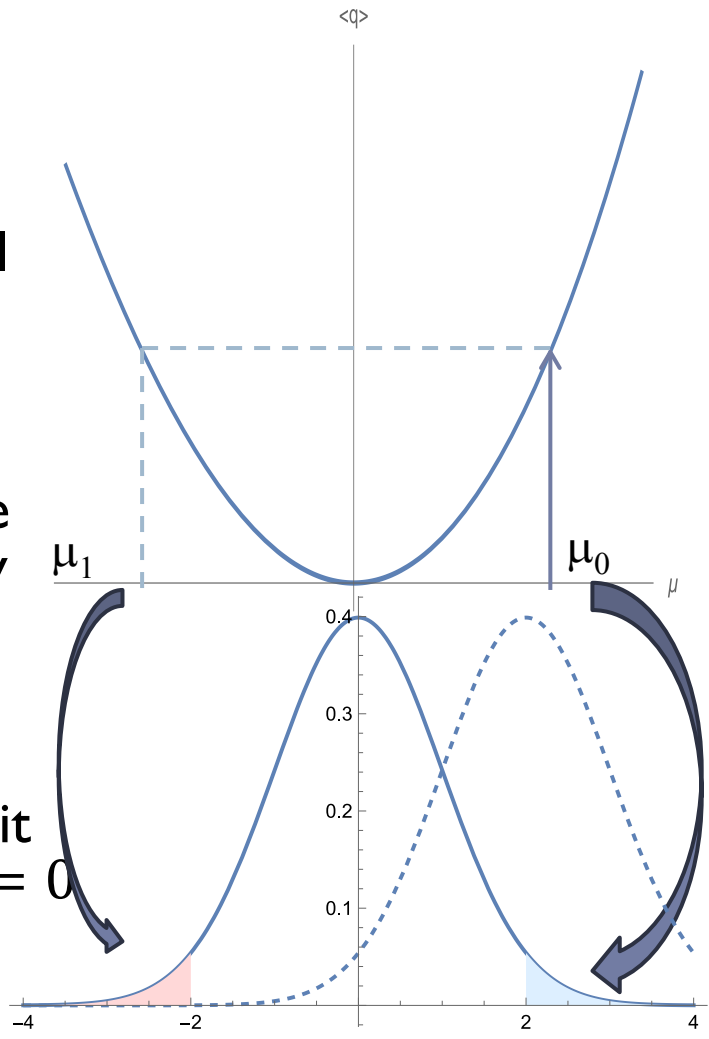
More general one POI

Two sided

Two minima

One POI two sided

- Will the same approach work? **NO**
- Imagine a case where q has a minimum
- Our data shows some excess and would prefer a relatively large (and positive) signal strength μ_0
- But a similarly unlikely situation exists for another signal strength μ_1 (negative)
- If we sample using μ_0 we will only populate the upper tail, lower tail events will be very unlikely and with a huge weight
  - Often a biased result, only upper tail
  - or a huge variance
- Previous proof fails in this case, and in fact it can be see that the optimal is found for $\mu = 0$ hence *unweighted*



Any way out?

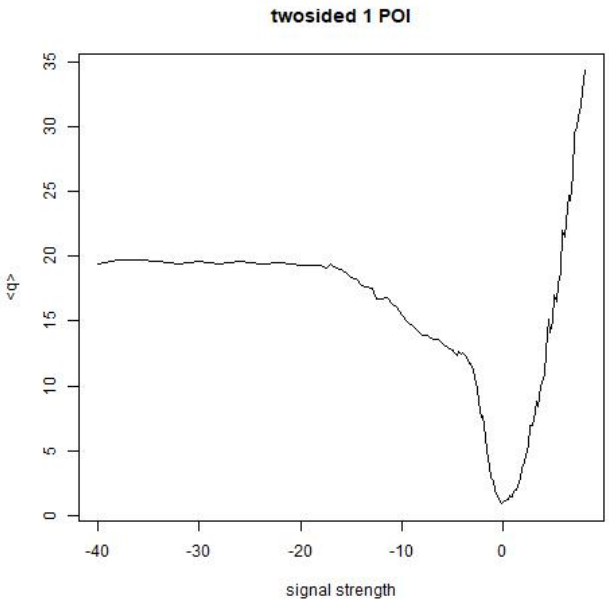
- Some ad-hoc cases can be solved (symmetric or well separated) but with some care of avoiding double counting
- I propose instead a more general approach:
  - use a mixture of both μ_0 and μ_1, the one preferred by the data and the other solution giving the same q
- Sample from $\tilde{\rho} = \frac{1}{2}\rho\left(\vec{x}^{\,j}, \mu = \mu_0\right) + \frac{1}{2}\rho\left(\vec{x}^{\,j}, \mu = \mu_1\right)$
- Drawbacks
  - Need to scan to get the second point
  - *Waste ½ of toys*
- **But it works!**

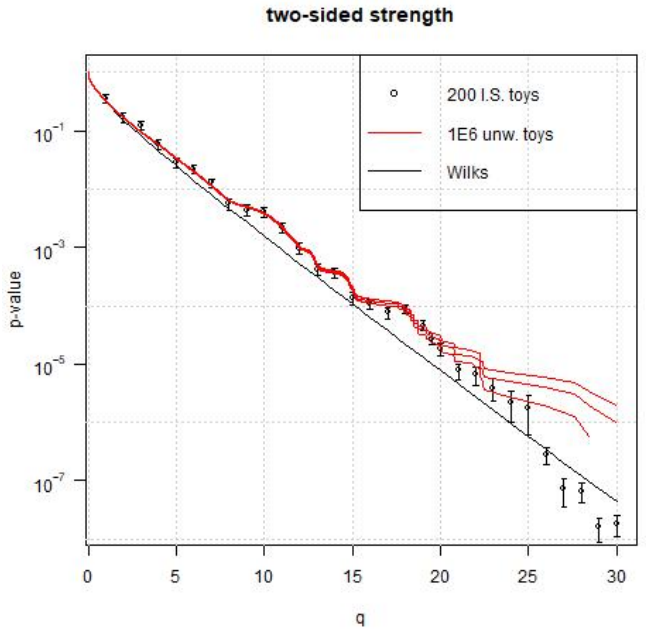
Algorithm (two-sided)

1. Fit the data to H_1, S+B, model and get q_0 and μ_0
2. Scan μ values, for each generate psexps $\{\vec{x}\}_j, \sim\rho(\vec{x},\mu)$ run the analysis, get q_j and calculate the average (note there might be a spread). Can be the available MC points
3. From <q> as a function of μ get the two values μ_0 and μ_1 in your scan closer to q_0
4. Get M psexp $\{\vec{x}\}_j, \sim \frac{1}{2}\rho(\vec{x}^j, \mu=\mu_0) + \frac{1}{2}\rho(\vec{x}^j, \mu=\mu_1)$
   - or a combination of the **two** closer full MC samples
5. Fit $\{\vec{x}\}_j$ get μ_j and q_j (repeat the full analysis on this pseudodata)
6. Calculate weights $w_j = \frac{\rho(\vec{x}^j,\mu=0)}{\frac{1}{2}\rho(\vec{x}^j,\mu=\mu_0)+\frac{1}{2}\rho(\vec{x}^j,\mu=\mu_1)}$
7. Calculate p as $\frac{1}{M}\sum_{q_j>q_0} w_j$

An (extreme) example

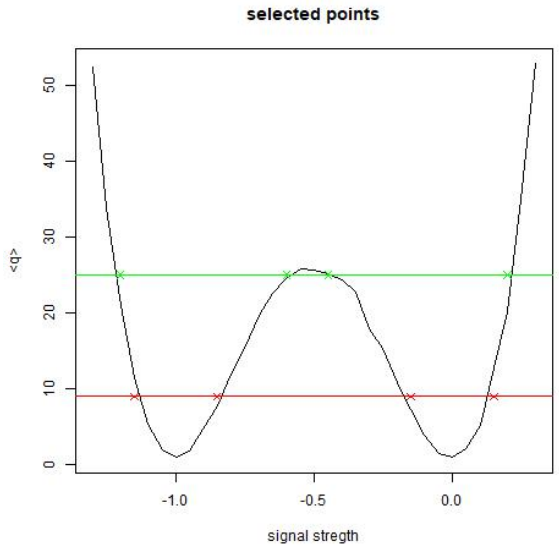
- Same low stat histogram, but allowing μ to be negative (forcing λ to be nonnegative)
  - Poisson stat with $\lambda_i = \max(b_i + \mu s_i, 0)$ μ positive or negative
- Note that Wilks does not work that well, but 200 weighted toys provide good results

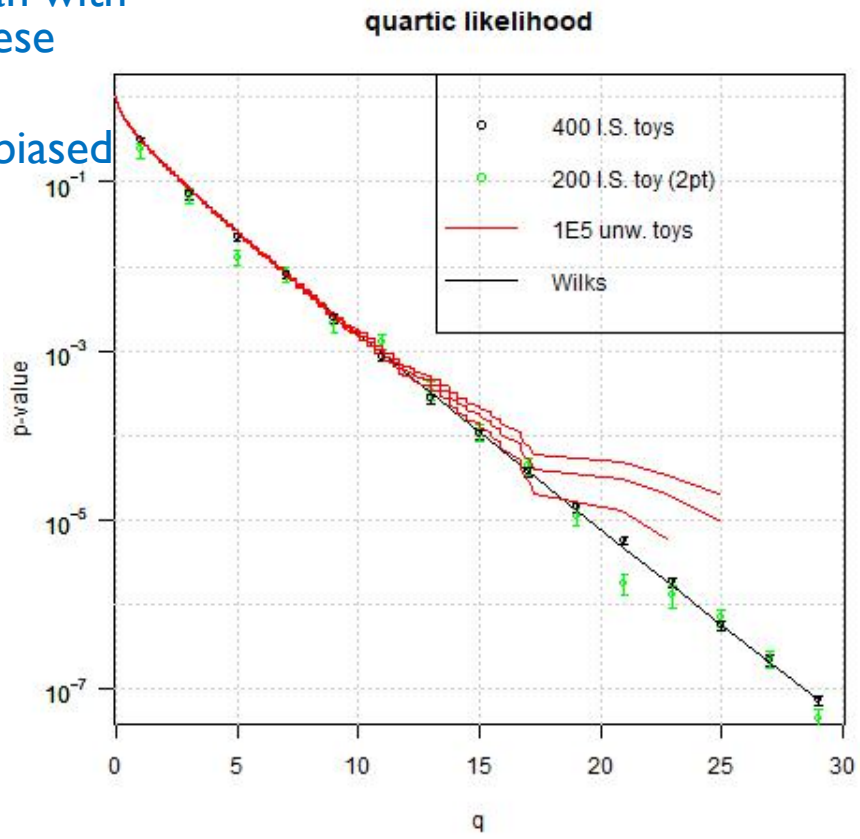




Quartic and more

- This idea can be extended to more complex LR shapes, forcing it to be quartic
  - $\lambda_i = b_i + (\mu + \mu^2)s_i,$
  - For a given q_0 look for all μ_i in your scan with the same q_0 and use an admixture of these $\sum \rho(\vec{x}^j, \mu = \mu_i)$
  - Admixture of only two pt can produce biased results





A glance to 2 or more POI

Francisco Matorras, IFCA, Spain

2 or more POI

- Can we extrapolate to several POIs?
  - Not trivially…
  - Need to populate an n-D region defined by <q>=q_0
- In principle can use the same trick of an admixture
  - But a continuous set of values to mix from
  - The sum on the weight denominator becomes an integral
  - No closed form except from trivial examples (and nontrivial solution, modified Bessel functions for n-D gaussian)
  - Efficiency degraded as one goes to higher dimensions
- My proposal:
  - do a grid scan of parameters $\vec{\alpha}$
  - find *a few* points $\vec{\alpha}_i$ in the POI space compatible with q_0
  - As before, use an admixture of these $\frac{1}{n}\sum_i \rho\left(\vec{x}^j, \vec{\alpha} = \vec{\alpha}_i\right)$
- Drawbacks: unclear how many points needed; n-D scanning can be time consuming…
- But seems to work

An example

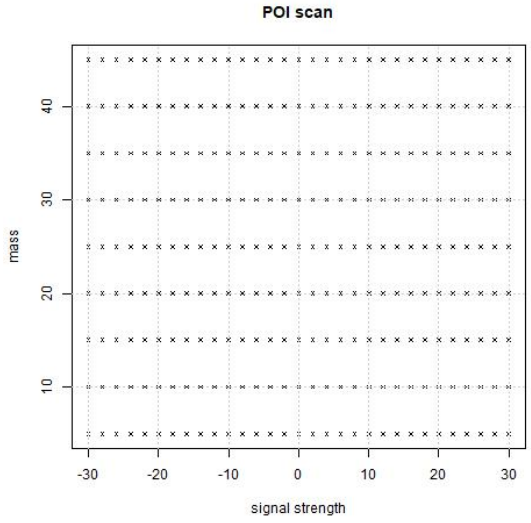
A Gaussian signal over an exponential background

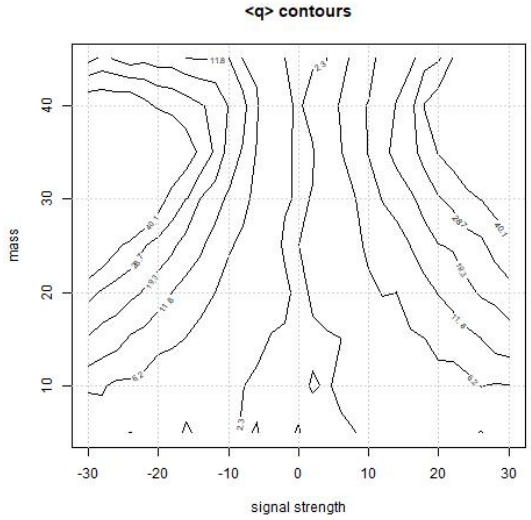
Free signal strength and mass

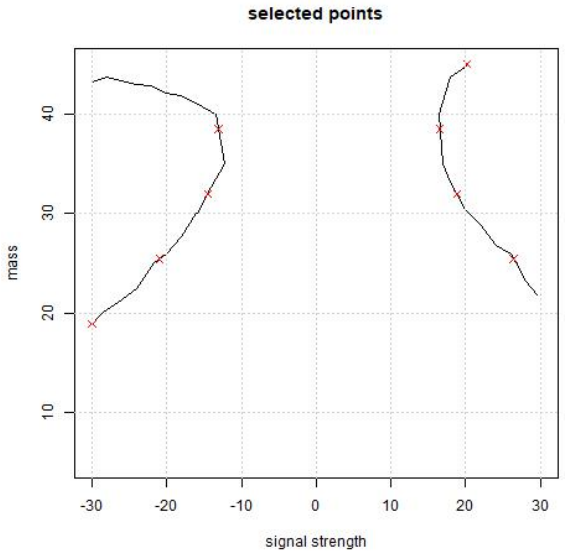
2 POI

- Exponential background $A\, e^{-\alpha x}$ (A and α, fixed)
- Gaussian signal $\frac{\mu}{\sqrt{2\pi}\sigma}\, e^{-\frac{(x-m)^2}{2\sigma^2}}$ (μ and m free, σ fixed to 5 times the bin width)
- 50 bins

- Define a grid of parameters
- Run psexp for each, calculate <q>
- Select the contour corresponding to q_0
- Select n points along that contour
- Admixture of their pdfs
- Generate and reweight as before

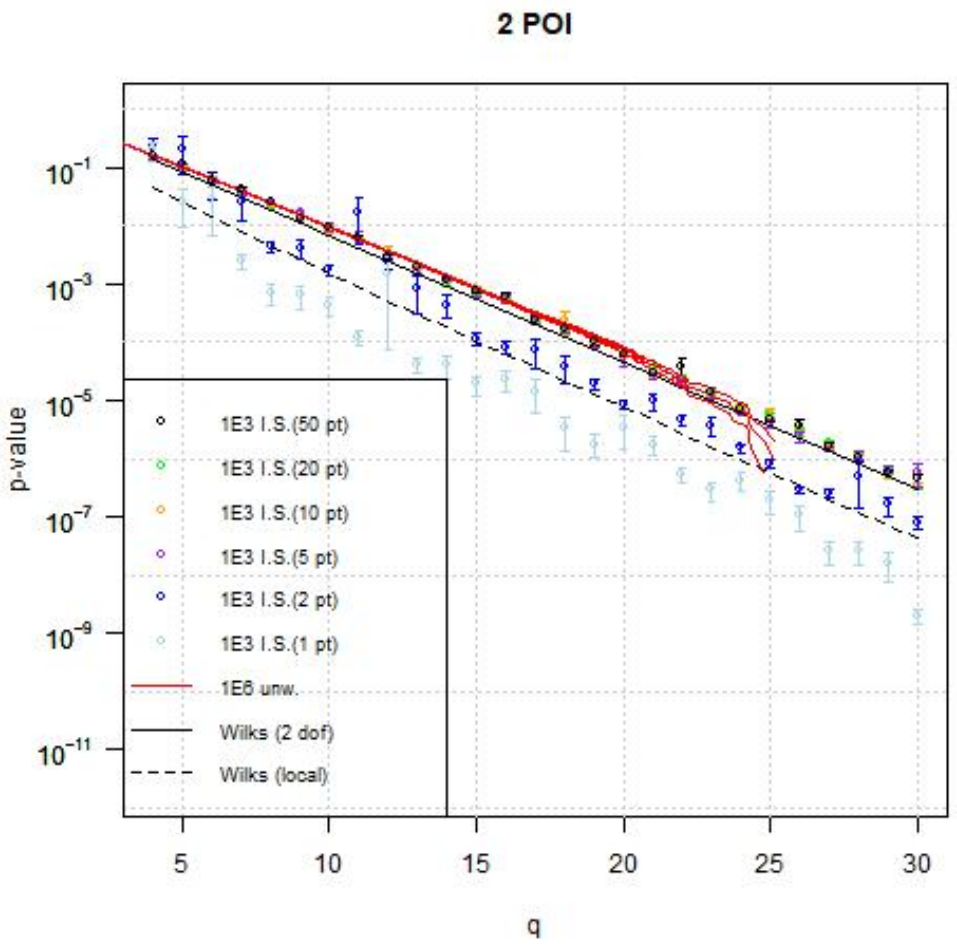






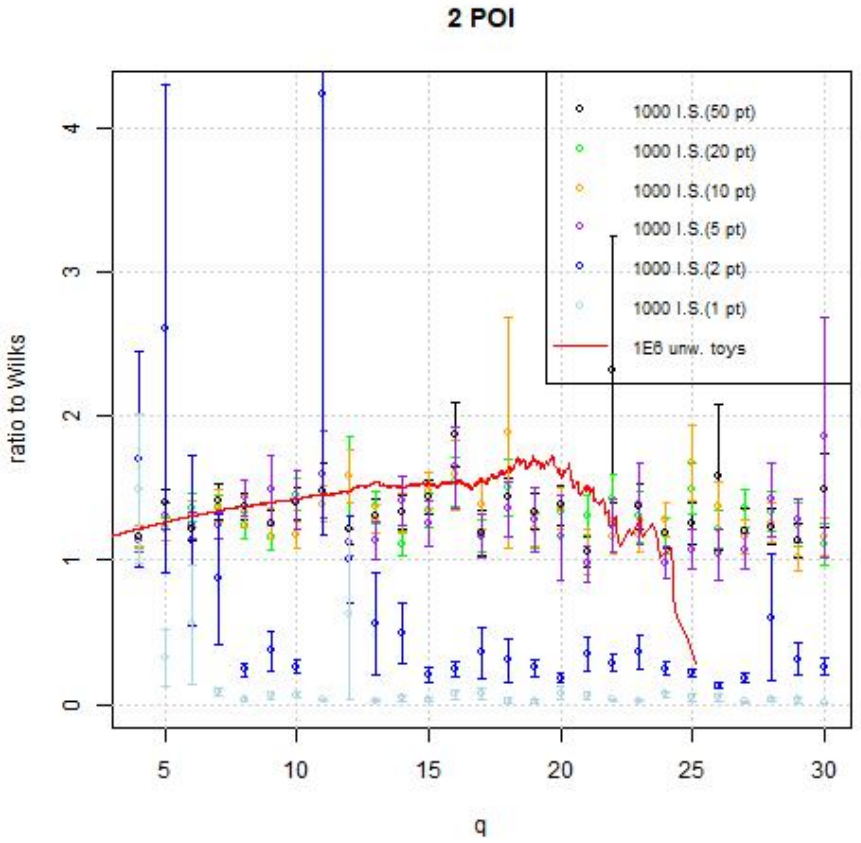
Results

- Less efficient but still can obtain good results (in the many sigma level) with O(1000) toys
- too few points along the contour provide biased results, but stabilize with pt~ 5



Results

- Same plot but normalized to Wilks prediction (dof=2)
- True p-values show small departure from 2 dof Wilks (~1.5 factor)
- additional 6-7 factor from "local p-value" (wilks and dof=1)
- **Can easily calculate true significance, in this case usual calculation off by a factor** 10



Summary & Conclusion

Summary

- The use of **importance sampling** to estimate very small p-values has been explored
  - **based on sampling from pdf taken from S+B models**
  - Avoid relying on Wilks
  - Can give a handle to step ahead of *local p-values*
- For models with only one POI and monotonic behavior, **sample from the S+B pdf that best fits to the data**
  - Rather general proof of validity
  - Reduce the toy sample size several order of magnitude for 5σ discoveries
  - Can use existing full MC for estimation
- A proposal to extend to more complex cases and to >1 POI is presented
  - Based on building admixtures for similarly significant μ
  - Require a likelihood scan and less efficient
  - Encouraging results
- Accounting for nuisances underway, but looks straightforward

Conclusion

Importance sampling can provide a handle to calculate p-values for discovery when asymptotic calculations cannot be trusted and to calculate global p-values

Thank you for your attention

Additional material

How many points?

- Rule of thumb: need to populate sufficiently all the regions along the contour $q(\vec{x}) = q_0$ (note this is a different contour)
- Each dashed circle represent the region sampled from the importance sampling
- Following sketch illustrates for a 2D normal
- For 5σ pt $\approx \frac{5\pi}{k}$ with k~1-3 depending on the size of the sample