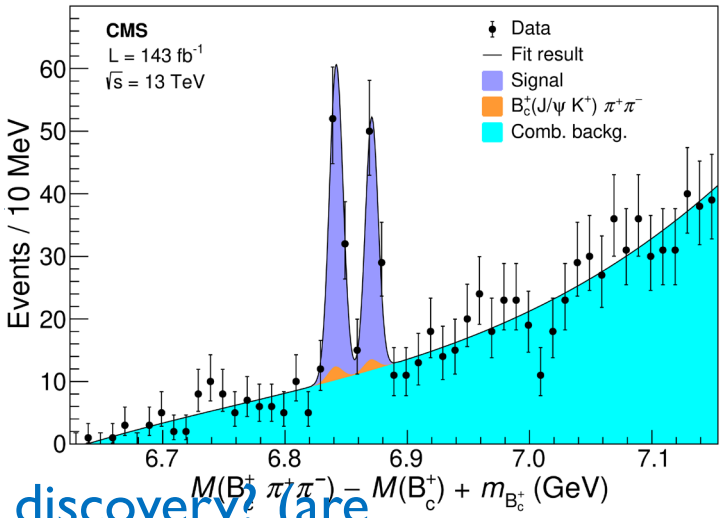
Evaluating extremely low p-values with importance sampling techniques in discovery-oriented HEP analyses.

Francisco Matorras
IFCA
Instituto de Física de Cantabria (Santander)

Introduction



- This study was triggered by this b-physics example
  - Is there a peak or two peaks?
  - How many "sigma" can we quote for the discovery? (are we better than other experiments?)
  - Can we trust Wilks in this particular case to the 10^{-7} level or less?
  - Obviously, generating 10^7 toys of a full analysis (including non-trivial fits) is unattainable
- A situation rather frequent in b-physics (also in other fields)
- Starting to investigate a toy-based method with 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 expectations asymptotically converge to the desired value
  - If you play your cards, it will converge faster (i.e., need less toys)
- Here, seek for a very particular application, since we are interested in the tails of a test statistic q (usually PLR, but not necessarily) to get the p-value

Posing the problem

- A (pseudo) experiment defined by a set of variables $\{x_i\}^j$ or $\vec{x}^j$
  - j runs over PsExps, j=0 real data
  - i runs over the events in a PsExp, I'm assuming fixed number of events N and 1D distributions (a single variable)
- The background is described by a pdf $\varrho(\vec{x})$ and if we can assume independence of the events $\varrho(\vec{x}) = \prod_i^N \rho(x_i)$, based on the pdf of the individual events (see more on iid later)
- We can define a statistic $\mathrm{q}(\vec{x})$, usually a LR, which takes the value $\mathrm{q0} = \mathrm{q}(\vec{x}^0)$ for the real data
- We can write the p-value as $Pval = E[\theta(\mathrm{q}(\vec{x}) - q0)] = \int \theta(\mathrm{q}(\vec{x}) - q0)\, \varrho(\vec{x}) d\vec{x}$
  - θ is the step function, 1 if argument positive 0 otherwise
- Or estimated from a sample as
  - $\mathrm{P}val = {}^1/_M \sum_1^M \theta\left(\mathrm{q}\left(\vec{x}^j\right) - q0\right)$

Importance sampling

- We can also write $Pval = \int \theta(\vec{x})\, \varrho(\vec{x}) d\vec{x} = \int \theta(\vec{x})\, \frac{\varrho(\vec{x})}{\tilde{\varrho}(\vec{x})} \tilde{\varrho}(\vec{x}) d\vec{x} = \int \theta(\vec{x})\, W(\vec{x}) \tilde{\varrho}(\vec{x}) d\vec{x}$
- Or the familiar $\mathrm{P}val = {}^{1}/_{M} \sum_{1}^{M} \theta\left(\mathrm{q}\left(\vec{x}^{\,j}\right) - q0\right) W(\vec{x}^{\,j})$, when x are sampled from $\tilde{\varrho}$
- Remarks:
  - Asymptotically it must work for any $\tilde{\varrho}$ provided some regularity conditions are fulfilled
  - W is the weight of the PsExp, **is a likelihood ratio**
  - Note that if the events are independent it is derived from the product of the **event** weights/LR $\prod_i^N \frac{\rho(x_i)}{\tilde{\rho}(x_i)}$
    - Rather easily grows to very large numbers or goes down to negligible values (prod of many events)

Importance sampling II

- However, not all $\tilde{\varrho}$ work better (converge faster) than unweighted samples, intuitively
  - we want to sample PsExp where θ is different from zero (q>q0), i.e., more "signal like" events
  - We do NOT want PsExp with a large weight (amplify fluctuations). Not always easy to know in advance, q is a complex function of the events.
- An optimal (in the sense of minimizing the variance of the estimation) $\tilde{\varrho}$ can be derived [1]:
  - $\varrho^*(\vec{x}) = \frac{\theta(\vec{x})\varrho(\vec{x})}{\int \theta(\vec{x})\varrho(\vec{x})d\vec{x}}$
  - But useless ☹, the integral in the denominator is the pvalue we want to get!

My conjecture

- Use as sampling pdf, your signal-included model which better fits your data provides a better way to estimate the p-value (of the data q0)
  - The weights become background/signal likelihood ratio
- Note that we don't need the best solution, a good solution is enough!
- Why that makes sense?
  - Sampling with this pdf will produce q in the neighborhood of q0, the region of more interest (we do not care much of the 99.99999% of the background-like events whose q is small)
  - Some mathematical considerations support this is a good choice (next slide)
  - ...And examples confirm it

Some maths

- A common approach [1] is to minimize the variance for a parametric family of pdf's and chose the optimal (set of) parameter(s)
- Let's use our signal+background model and try pdfs $\varrho(\vec{x}|\mu)$ where μ is the signal strength, or any other (set of) parameter(s)
  - Remember, ϱ is the model of the experiment, in the simple case product of the pdf's of the individual events.
- An optimal μ can be obtained minimizing the variance (now a parametric minimization), look for the μ which provides a smaller variance on the p-value estimation

Some maths

- It can be seen [1] that the minimum variance can be achieved for
  - $min_{\mu}\left({}^{1}/_{M}\sum_{1}^{M}\theta\left(\mathrm{q}\left(\vec{x}^{\,j}\right)-q0\right)\right)^{2}W\left(\vec{x}^{\,j}|\mu\right)\Big)$ if the sample is taken from the background only pdf
  - Or $min_{\mu}\left({}^{1}/_{M}\sum_{1}^{M}\theta\left(\mathrm{q}\left(\vec{x}^{\,j}\right)-q0\right)\right)^{2}W^{2}\left(\vec{x}^{\,j}|\mu\right)\Big)$ if the sample is taken from the background+signal pdf
- If we can replace the W_j by its average and can accept that in the neighborhood of μdata, q>q0, minimizing this function is equivalent to the MLR we perform on data
- Not exactly a proof, but...

Proposed procedure

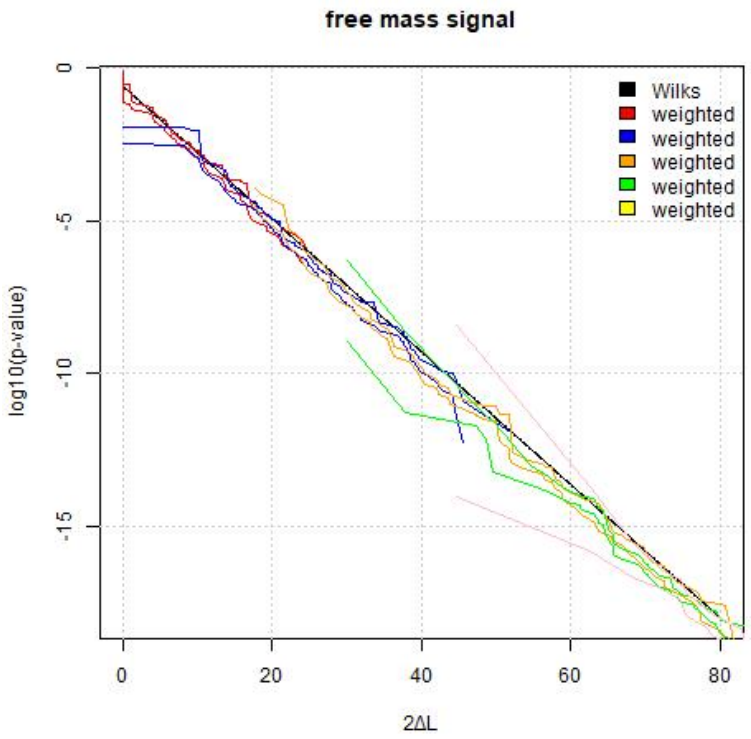
1. Start generating toys according to your S+B model best fit to real data
   - Note that you want to test the pvalue for q0 obtained from data, if you want to test another situation (i. e. find q corresponding to 5-sigma) it might not be an optimal choice
2. Perform your pseudoanalysis
3. Weight the PsExp according to the likelihood ratio of B model and your reference S+B model (a very small number)
4. Calculate your pvalue as the sum of weights divided by the number of PsExps

A few simple examples

- Fixed number of events in each "experiment" (10, 100, 1000)
- Only 100 pseudo-experiments to *force* the limits of the method
- Compare to Wilks prediction (in some cases to unweighted toys)
- Assume a known parametric pdf per event and independence (some considerations about that at the end)
- Different signal parameters tested exploring different regions of p-value
- Unbinned LR fit to get the best μ

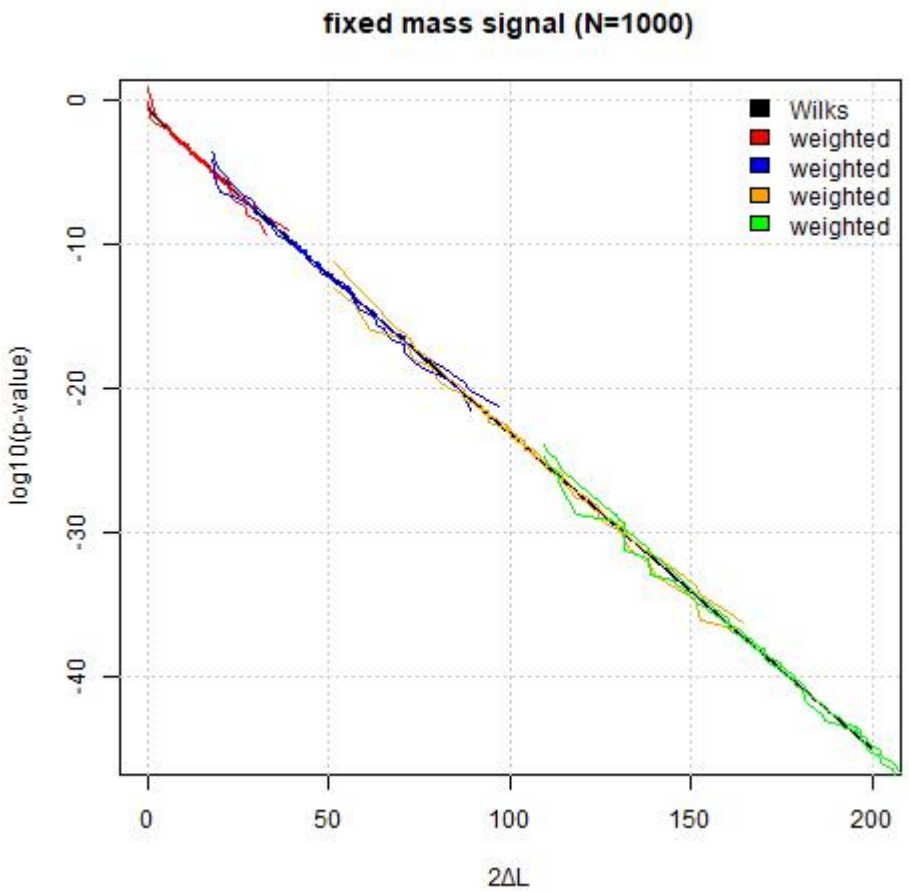
Summary plot of the tests

- For each test I'm showing a plot like this one
- It shows the p-value (upper tail prob) as a function of twice the difference in the Log Likelihood
  - as calculated by this method and different weighted samples optimized for a particular pvalue range
  - Compared with the Wilks prediction (in black)
- Weighted calculations shown as a $\pm 1\sigma$ band



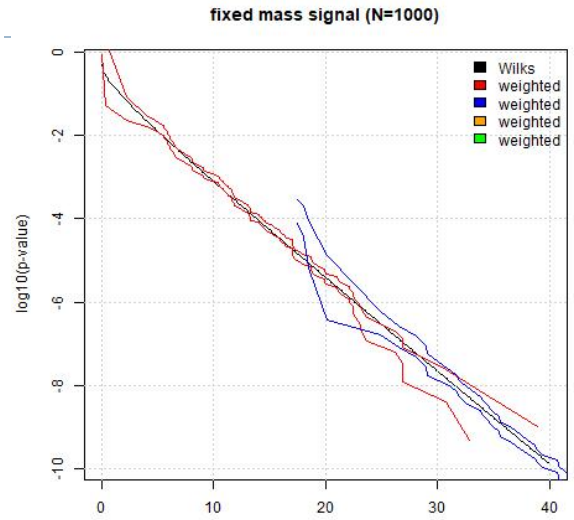
Exponential + fixed mass signal

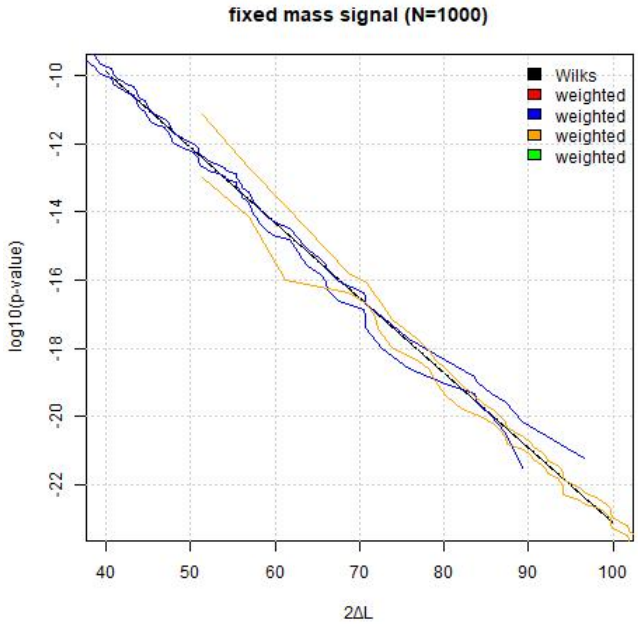
- Exponential background + fixed mass gaussian signal
- 1000 events/ps-exp
- Impressive agreement with Wilks down to $p<10^{-40}$
- Only 100 PsExp
- Zoom in next slides
- Cannot see on the plots but tested that indeed using the fitted μ is (at least close to) minimal variance. Not strongly dependent though

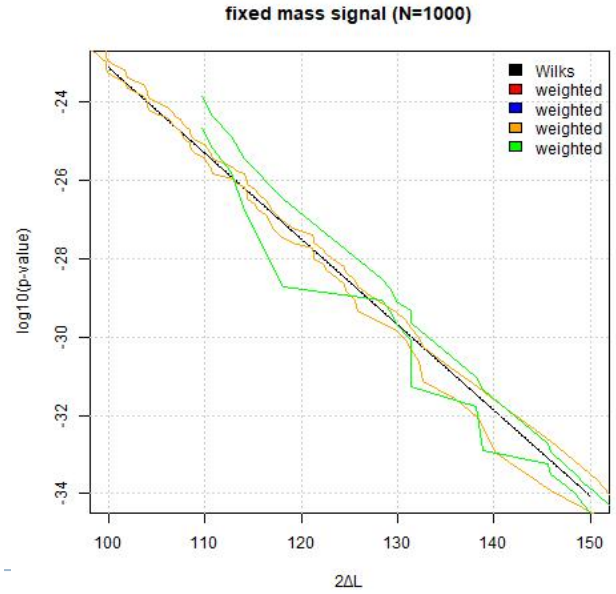


Exponential + fixed mass signal

- Can see that each sample has a range of appropriate prediction, corresponding to the expected signal

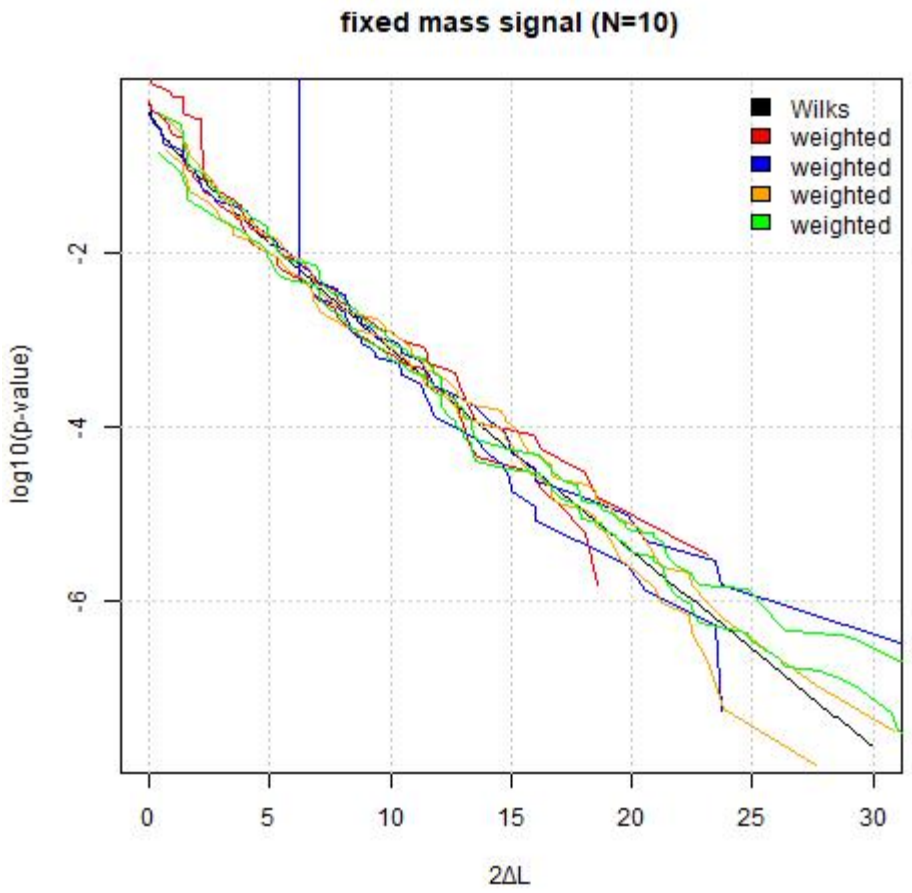






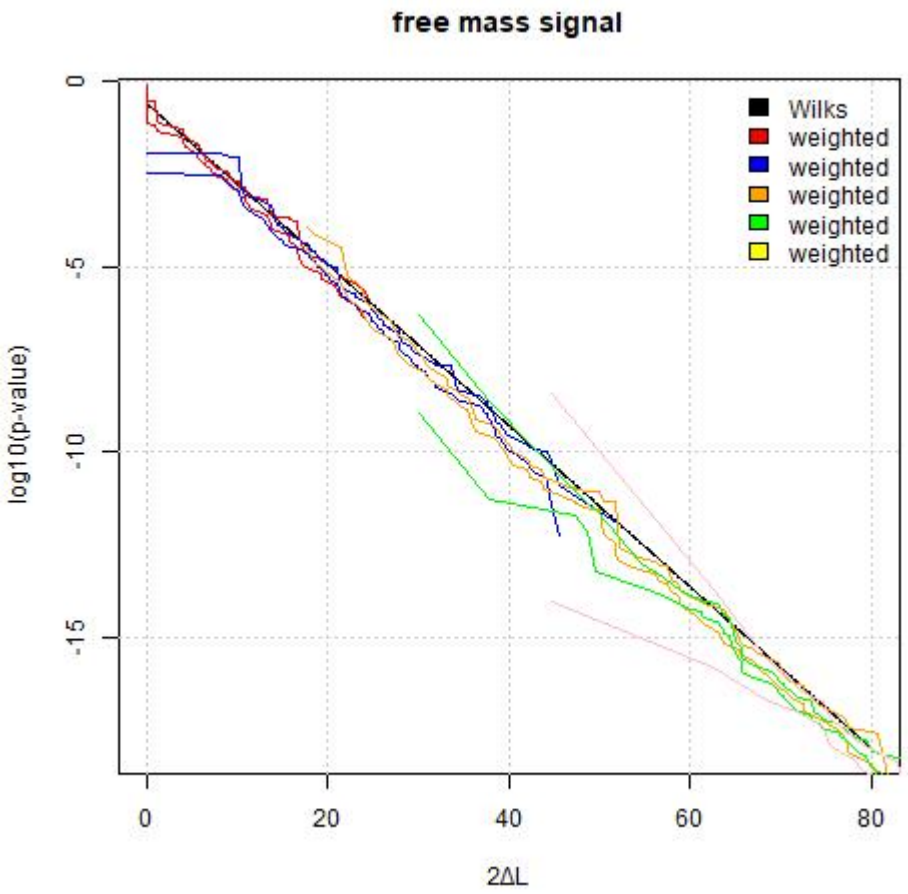
Exponential + fixed mass signal

- Same exponential background + fixed mass gaussian signal but only 10 events per PsExp
- still only 100 PsExp
- Still following Wilks but cannot go too far

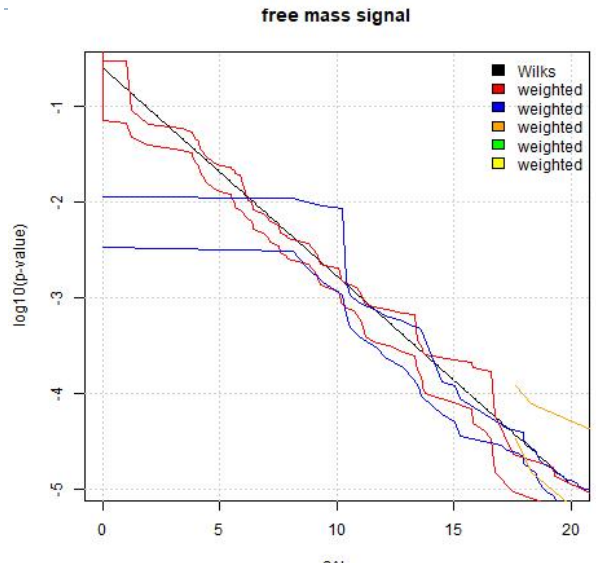


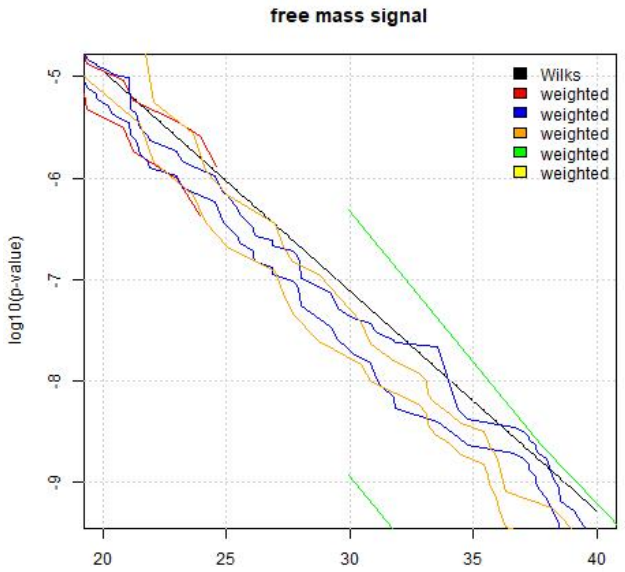
Exponential + free mass signal

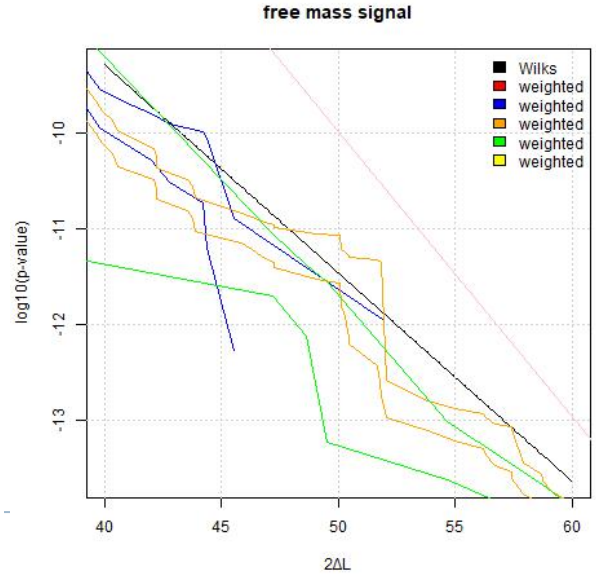
- Same model but the mass of the peak is allowed to vary (2 dof)
- 1000 events PsExp
- still 100 PsExp
- Still following Wilks although maybe some departure at 10^{-15}
  - is Wilks failing or the method does not work?

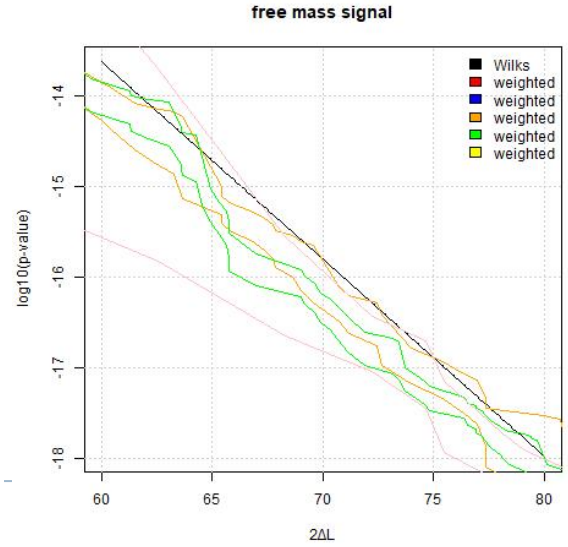


Exponential + free mass signal



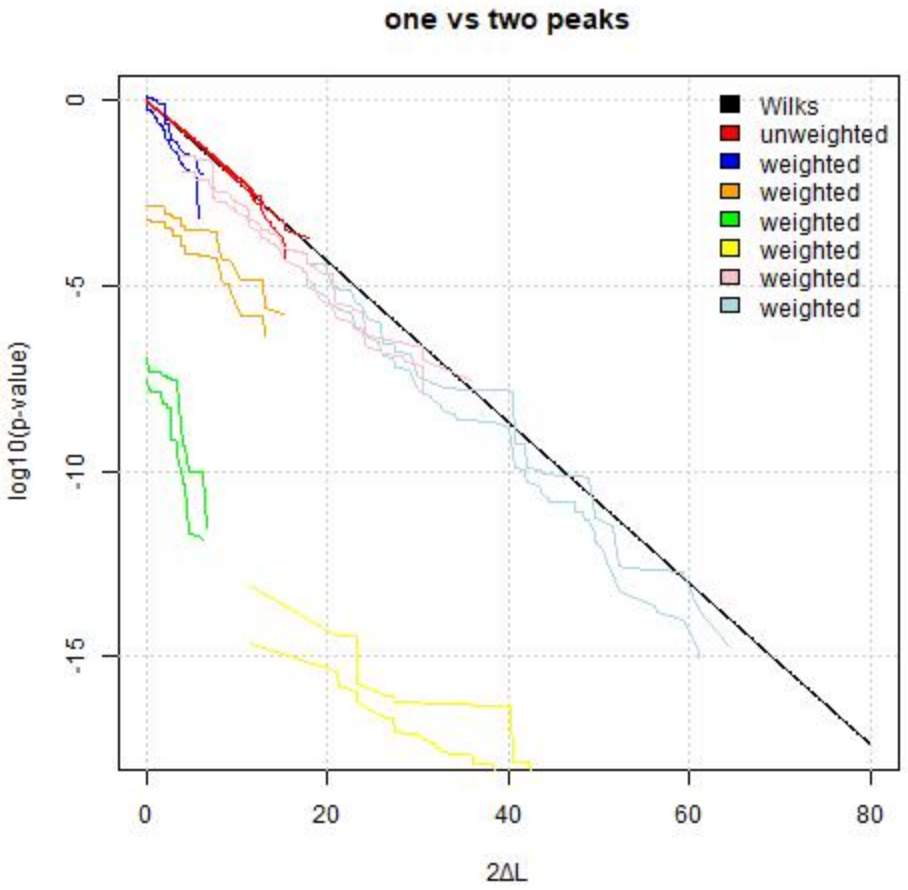






Two vs one peak

- Let's try a case where we have no guarantee that Wilks holds
- Compare the hypothesis of two gaussians vs one gaussian (no constrains on the parameters)
- Compared also to 10000 unweighted toys
- 100 toys
- First run (slightly) disappointing
  - "seems" to work but not always
  - Wilks prediction works down to 10^{-4}

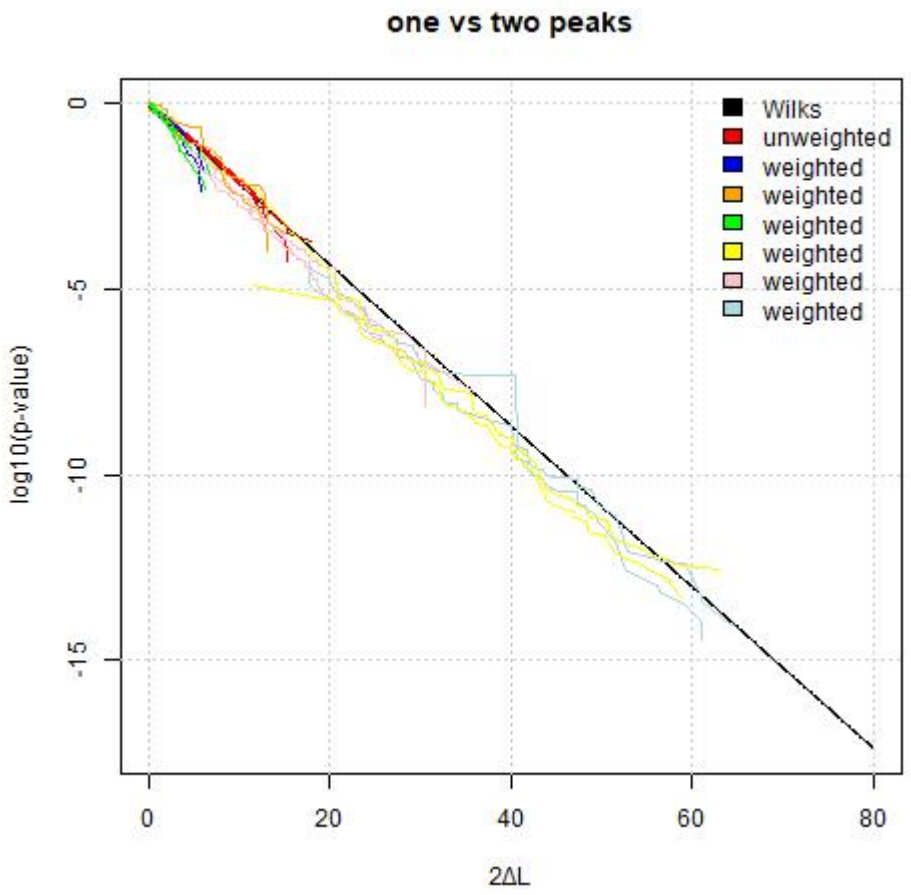


Two vs one peak

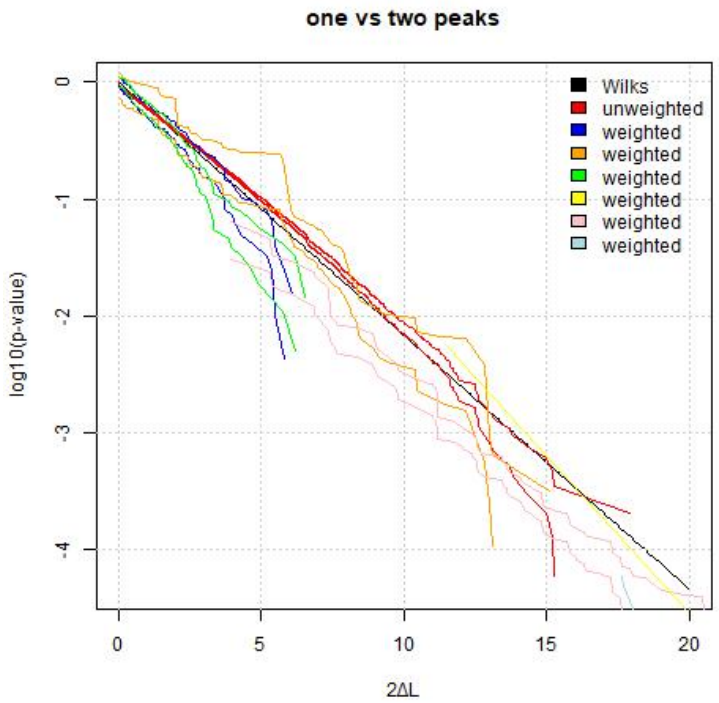
- The problem tracked down to the background model definition.
- What is the "background" let's say SM in this case?
  - Ill-posed problem also for unweighted toys?
- What is the B only pdf we have to draw from the toys?
- I had used a fixed mean and width gaussian
  - ANY gaussian? Which mean, spread range? Which law?
- On a second run, use the single gaussian that best fit to the BSM model

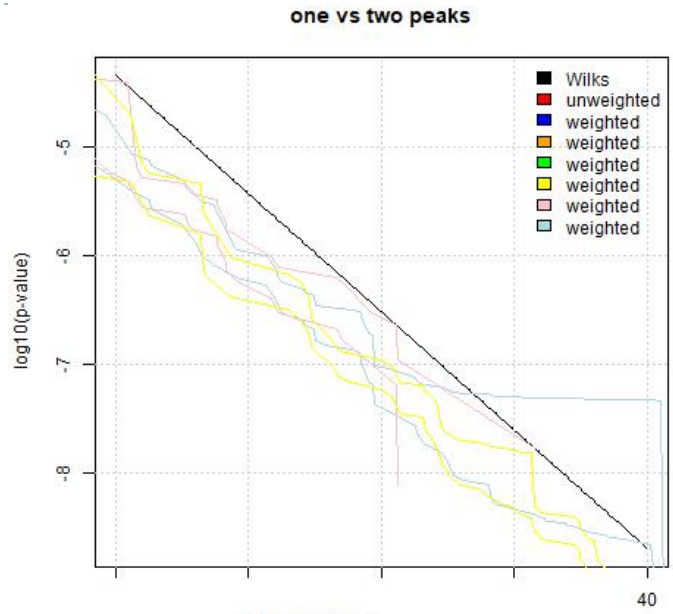
Two vs one peak

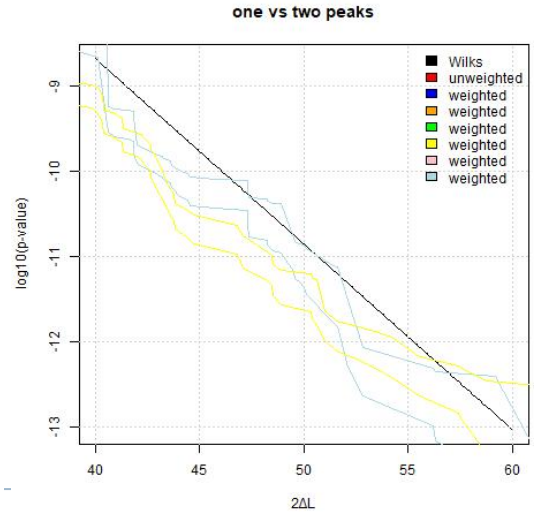
- Things turn back to normal
- Different samples give compatible results
- Wilks-like trend but some departure visible at low p-values
- Is really the problem ill-posed? How would you draw unweighted toys if you could?



Zoom for two peaks







Towards a realistic case

- So far, everything calculated assuming known analytical pdf and independent events
- What about a binned case?
  - Should not be a problem
  - $\varrho(\vec{x}) = \prod_i^N \wp\ (n_i|\lambda_i)$the product of the Poisson probabilities of each bin, given the bin expectation λ (with or without signal)
  - Can build your weights and sample from these pdf's, if any, even simpler
- Can we include nuisances?
  - In principle yes, similarly you should have your analytical pdf including nuisances for the likelihood
  - $\varrho(\vec{x}) = \prod_i^N \wp(n_i|\lambda_i(\vec{\nu}))\ \prod_j(f\ (\nu_i))$
  - You can sample the nuisances too or fix to the fit result

Conclusions

- A method based on importance sampling is proposed to estimate very small p-values with an acceptable number of pseudoexperiments
  - Generate weighted toys according to the signal model which better fits the data
- Promising results:
  - Can reproduce low-p tails of simple examples
  - Seem feasible to extrapolate to real cases
- Importance sampling can provide a handle to calculate p-values for discovery when asymptotic calculation are not trusted