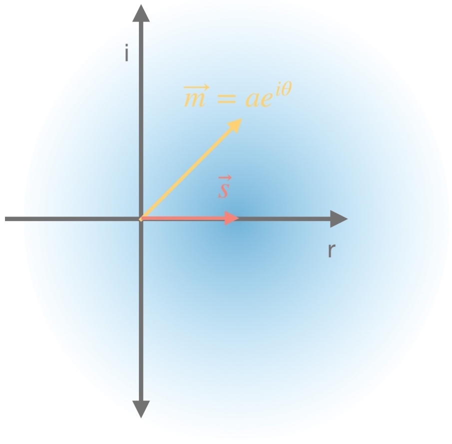
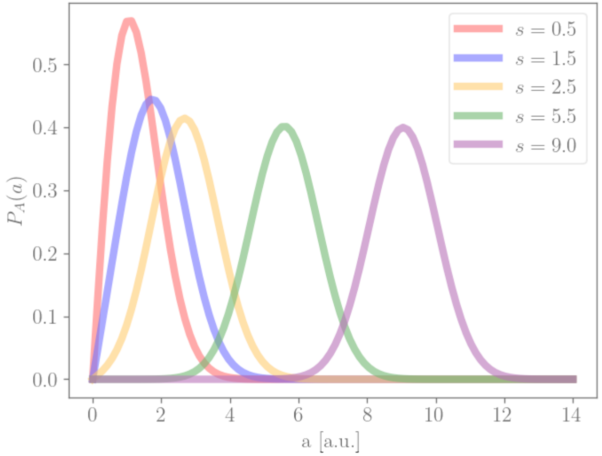
Annual Meeting of DDAp/DDEIT and HIRSAP 2023

Determining radio signals in the presence of noise

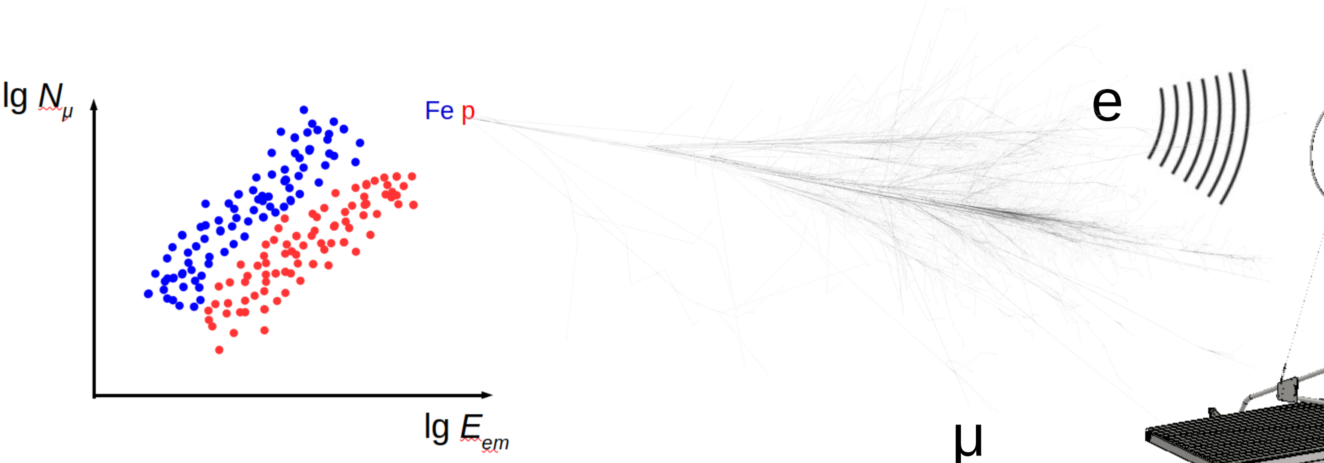
PhD student: Sara Martinelli
Supervisors: Prof. Dr. R. Engel, Dr. T. Huege, Dr. D. Ravignani

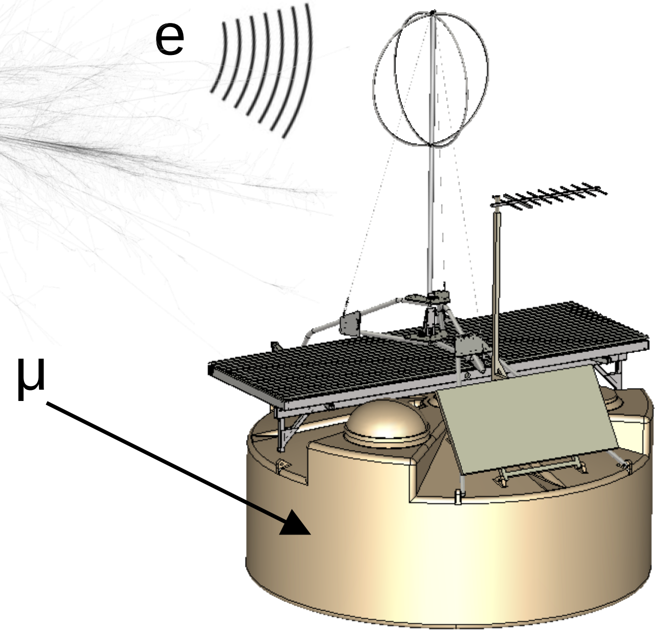




sara.martinelli@kit.edu

Mass composition study using the radio signals of RD

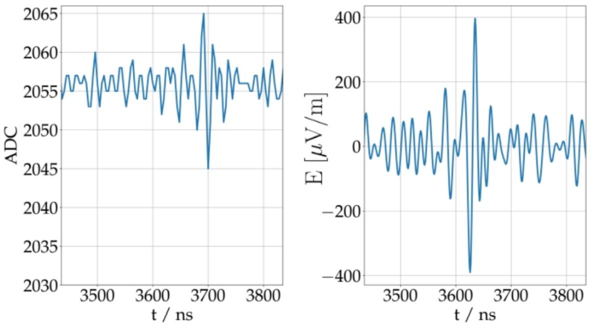




RD (Radio Detector) +
WCD (Water Cherenkov Detector)

Works well with <u>inclined air showers</u> ($65° \lesssim \theta \lesssim 85°$)

Reconstructing the electromagnetic energy of the shower with RD



1. E-field reconstruction

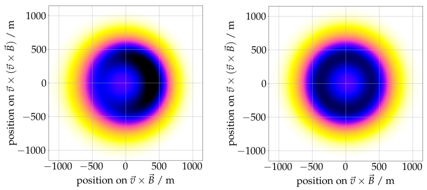
- Digital to analog conversion, upsampling, Hann window etc.
- Unfolding of the response of the signal-processing chain (LNA, impedance matching, filter amplifiers...)
- **Unfolding of the antenna response** (NEC-2) to get the E-field (EW, NS, N)

2. Calibrated signals

- Decomposition of the E-field in the **shower plane** coordinate system
- Estimation of signal-to-noise ratio (SNR)
- Estimation of the **energy fluence** f [eV m^{-2}], the energy deposit per unit area

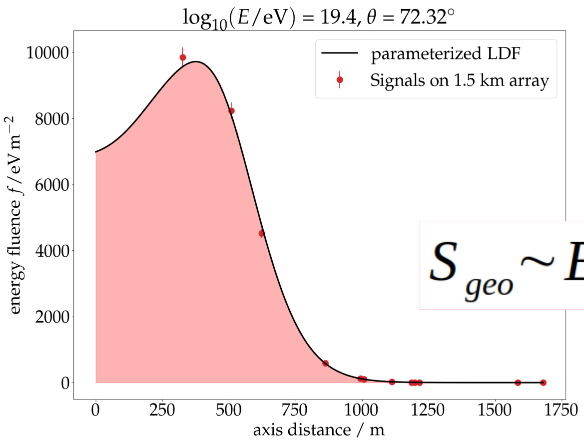
3. Geomagnetic energy fluence

- Analytic correction of early-late asymmetry
- Parameterized **subtraction of charge-excess emission** → 1-dim LDF



4. Radiation energy

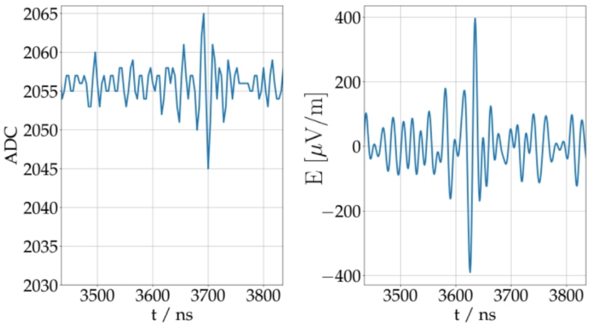
- **LDF fit** to estimate the geo radiation energy E_{geo} (energy emitted in form of waves)
- Correction on E_{geo} to compensate for the second-order scaling with the geomagnetic angle and air density at Xmax



$$S_{geo} \sim E_{em}^2$$

5. Elm energy

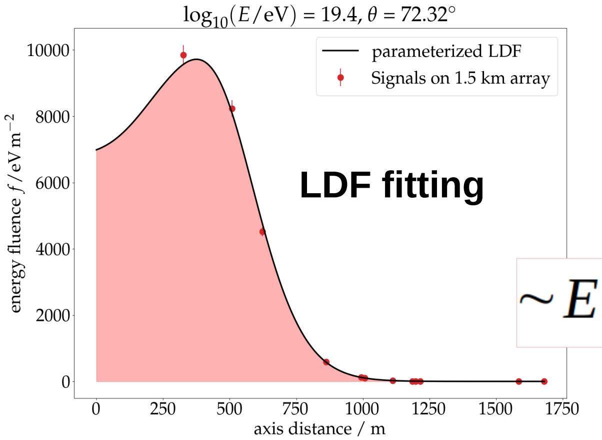
Reconstructing the electromagnetic energy of the shower with RD (in a nutshell)



energy deposit per unit area

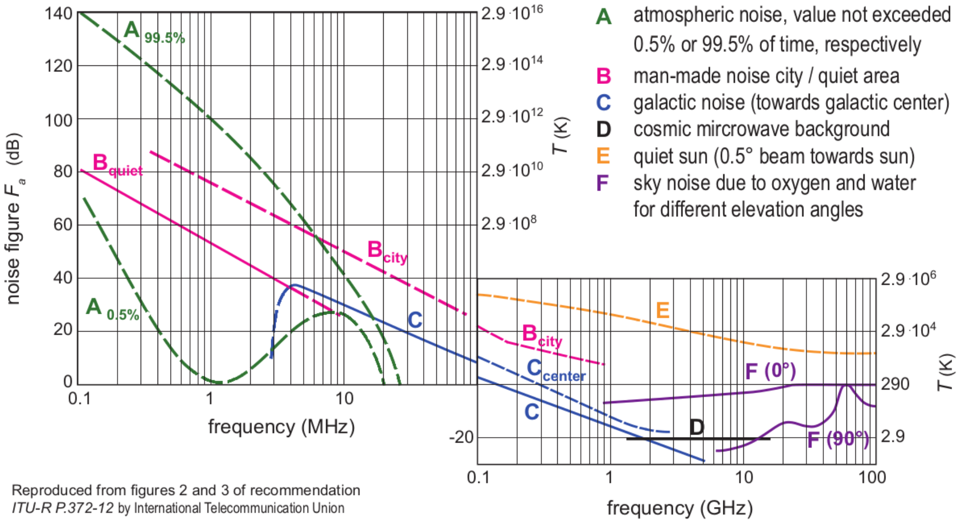
Energy fluence

f [eV m^{-2}]

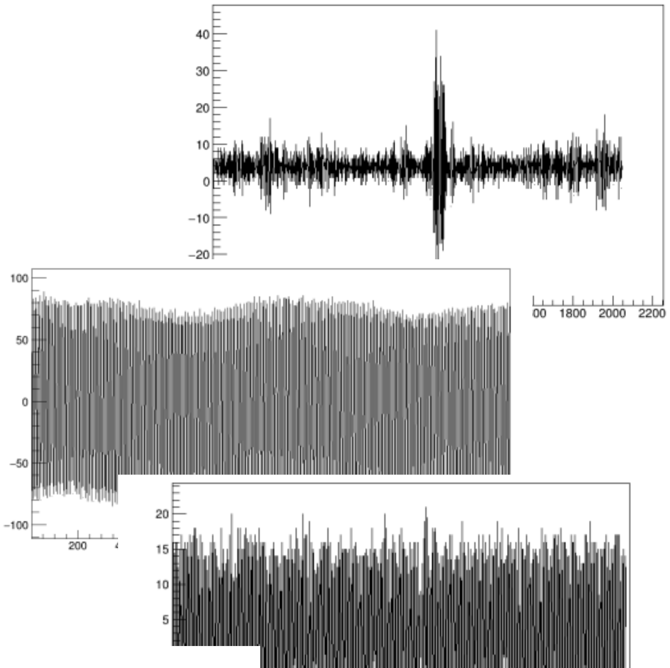


$\sim E_{em}^2$

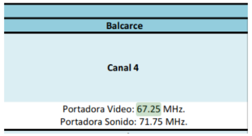
Radio background at Auger



Reproduced from figures 2 and 3 of recommendation *ITU-R P.372-12* by International Telecommunication Union

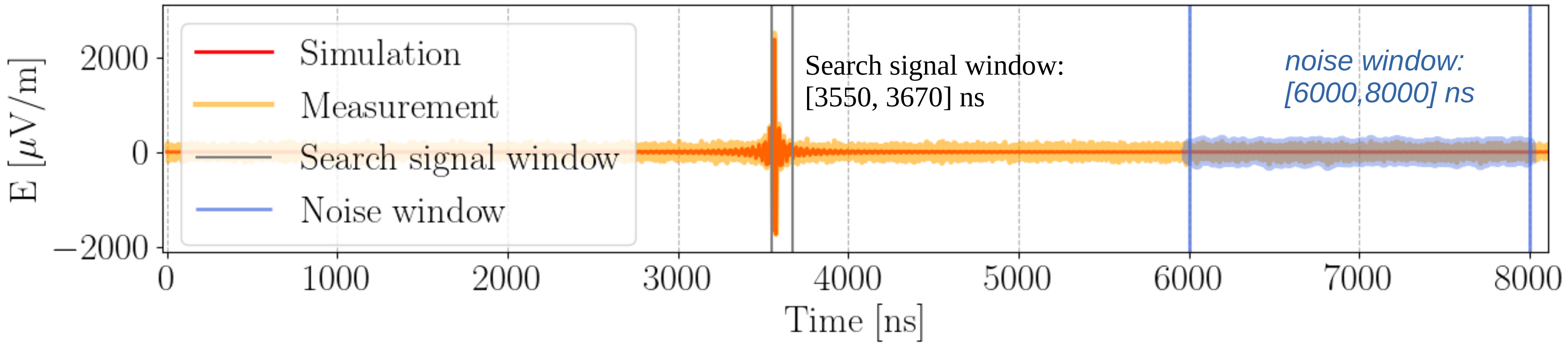


Industria TV Argentina



How do we determine the signal (fluence) from the noisy radio measurements?

Current signal and signal uncertainty estimation (Offline)

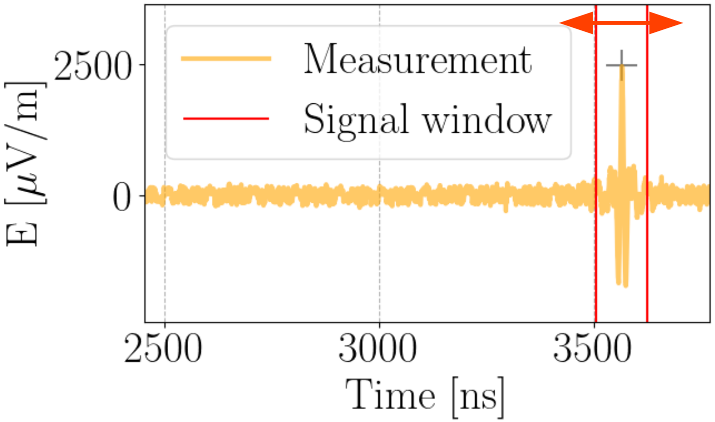


Estimation of the energy fluence [eV m^{-2}]

$$f = \epsilon_0 c \Delta t \left(\sum_{t_i=t_1}^{t_2} A(t_i)^2 - \frac{t_2 - t_1}{t_4 - t_3} \sum_{t_i=t_3}^{t_4} A(t_i)^2 \right)$$

signal window *noise window*

100 ns signal window around the peak



The method breaks down at low signal-to-noise ratio

Current signal and signal uncertainty estimation (Offline)

Signal-to-noise ratio cut at station level
SNR < 7 (10)

$$\mathrm{SNR} = \left(\frac{|A^{\mathrm{hilb}}_{\mathrm{tot}}|_{\mathrm{max}}}{V^{\mathrm{noise}}_{\mathrm{RMS}}}\right)^2$$

The uncertainties on the reconstructed electromagnetic energy (LDF fitting by χ^2 min.) are underestimated

$$\sigma_{\mathrm{f}}^2 = \sigma_{\mathrm{Gauss}}^2 + \sigma_{\mathrm{det}}^2 + f_{\mathrm{noise}}^2$$

(backup)

We want to achieve a better estimation of the fluence and its uncertainty exploiting a robust mathematical and statistical background.

Signal estimation in presence of noise with the Rice distribution

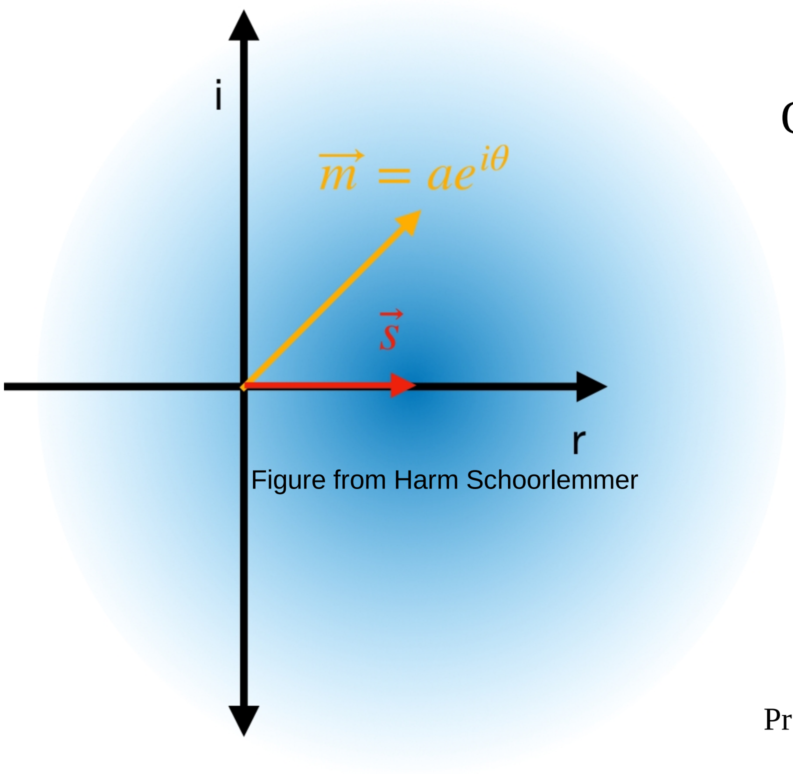


Figure from Harm Schoorlemmer

CAVEAT: Radio measurements have amplitude and phase!

Our **measurement** can be expressed as:
sum of **constant known phasor s** and a **random phasor sum**

Marginal density function for **amplitude** only:

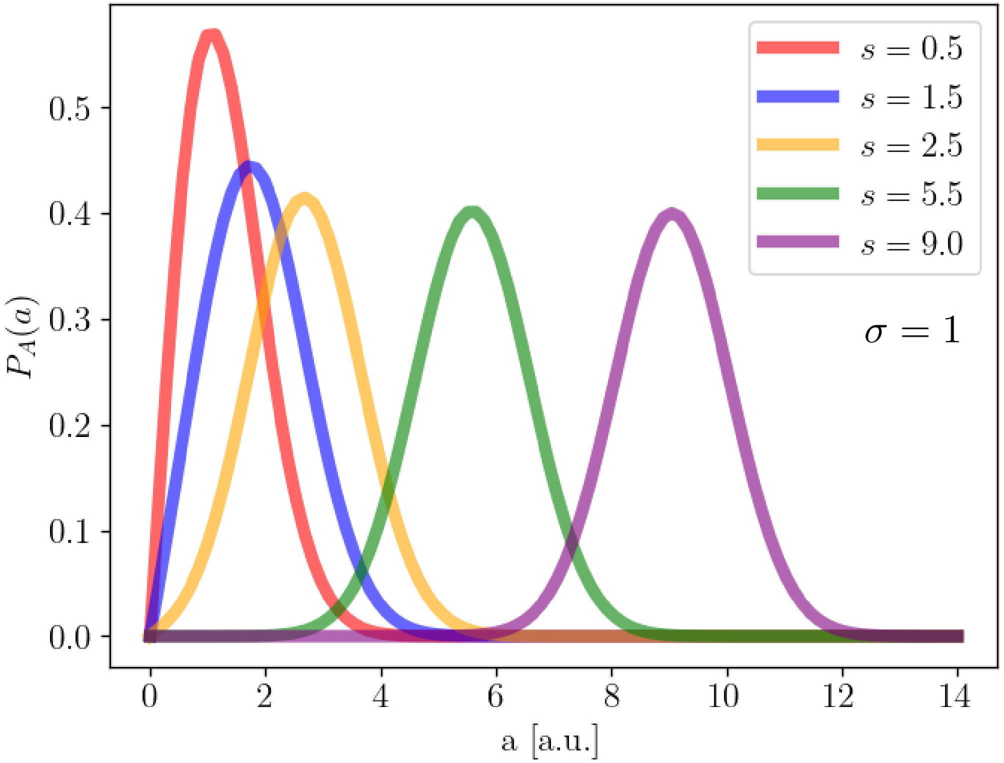
$$p_A(a) = \begin{cases} \frac{a}{\sigma^2}\, exp\left(-\frac{a^2+s^2}{2\sigma^2}\right)\, I_0\left(\frac{as}{\sigma^2}\right) & \text{a} > 0 \\ 0 & \text{otherwise} \end{cases}$$

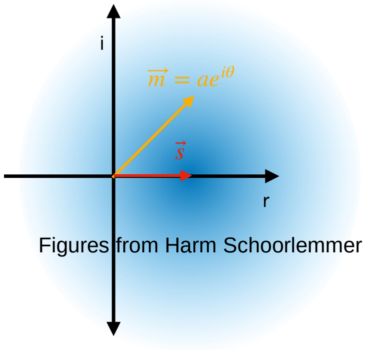
with I_0 modified Bessel function 1st kind of 0-order

Proof and details: Chapter 2.9 from J. W. Goodman, Statistical Optics (2015)

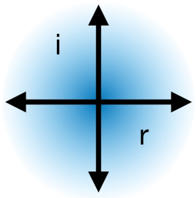
Signal estimation in presence of noise with the Rice distribution

$$p_A(a) = \begin{cases} \frac{a}{\sigma^2}\, exp\left(-\frac{a^2+s^2}{2\sigma^2}\right)\, I_0\left(\frac{as}{\sigma^2}\right) & \text{a} > 0 \\ 0 & \text{otherwise} \end{cases}$$

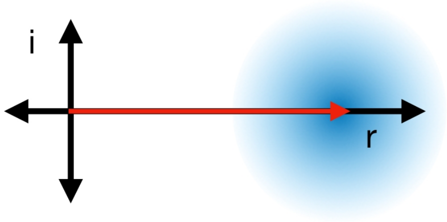




Figures from Harm Schoorlemmer

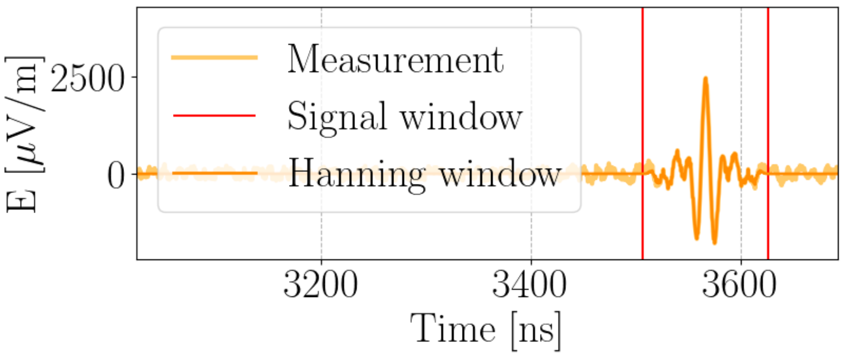


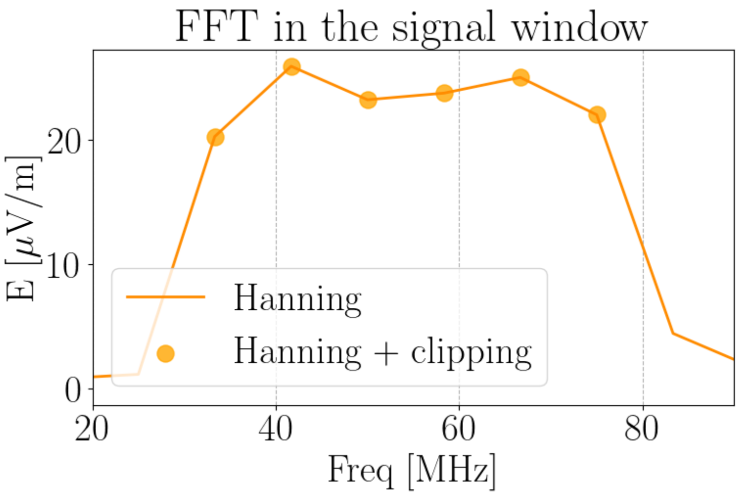
No signal
Rayleigh distribution $\mathcal{R}(a, \sigma)$



Large signal
Gaussian distribution
$\mathcal{N}(\sqrt{s^2 + \sigma^2}, \sigma)$

Signal estimation in presence of noise with the Rice distribution



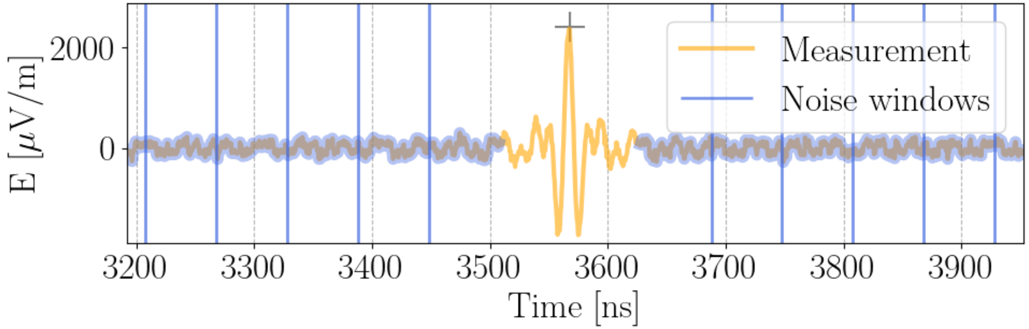


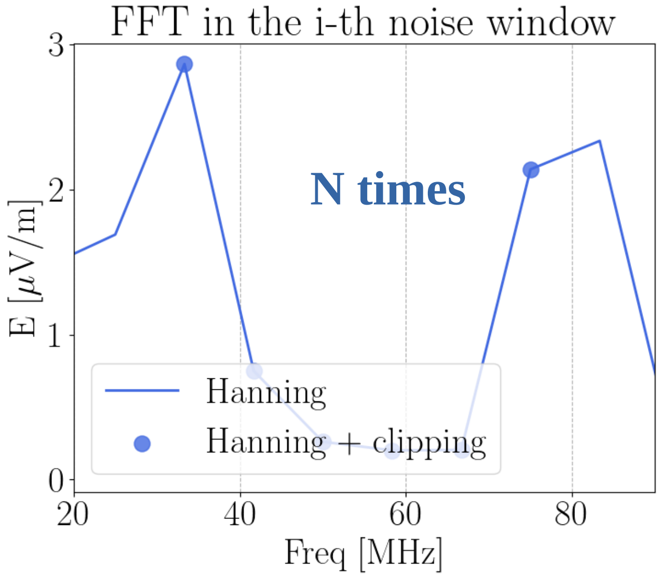
$$p_A(a) = \begin{cases} \frac{a}{\sigma^2}\, exp\left(-\frac{a^2+s^2}{2\sigma^2}\right)\, I_0\left(\frac{as}{\sigma^2}\right) & \text{a} > 0 \\ 0 & \text{otherwise} \end{cases}$$

a(f) amplitudes of FFT in the signal window

1. Find the Hilbert peak
2. Hanning window around the peak
3. Clipping in the signal window
4. FFT of the clipped trace
5. Select frequencies in [30-80] MHz

Signal estimation in presence of noise with the Rice distribution





$$p_A(a) = \begin{cases} \frac{a}{\sigma^2} exp\left(-\frac{a^2+s^2}{2\sigma^2}\right) I_0\left(\frac{as}{\sigma^2}\right) & \text{a} > 0 \\ 0 & \text{otherwise} \end{cases}$$

σ(f) noise level of each frequency

1. N noise windows all over the trace

2. In each window:

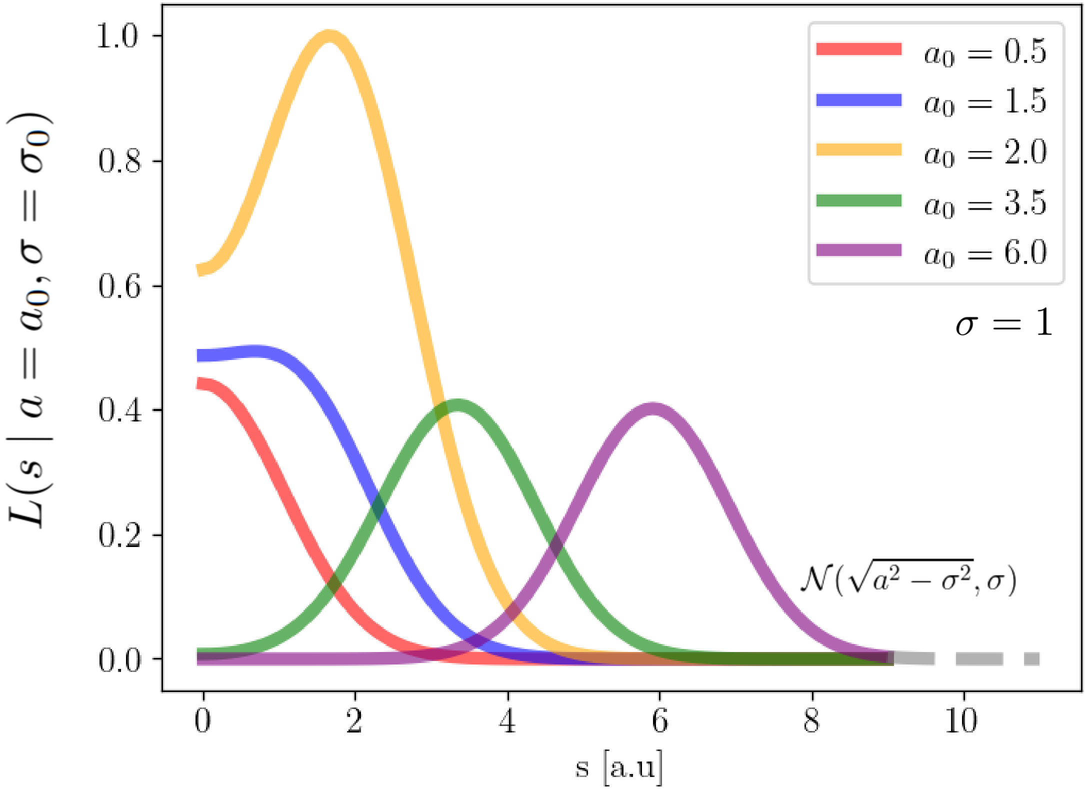
   - Hanning, Clipping
   - FFT, select frequencies in 30-80 MHz

3. Mean value of each frequency

$$\sigma = \mu / \sqrt{(\pi/2)}$$ Rayleigh distribution

Signal estimation in presence of noise with the Rice distribution

$$L(s \mid a = a_0, \sigma = \sigma_0) = \frac{a}{\sigma^2} \cdot exp\left(-\frac{a^2 + s^2}{2\sigma^2}\right) \cdot I_0\left(\frac{as}{\sigma^2}\right), \quad \mathrm{s} \geq 0$$



By storing in Offline the parameters **a(f), σ(f)** we have access to the likelihood L(s).

Get the estimators of **s(f)** by maximizing L(s)
(it can be zero!)

$$\hat{f} = \epsilon_0\, \mathrm{c} \cdot 2\,\Delta f \sum_{i=0}^{N_{\mathrm{bins}}} \left(\hat{s_{\mathrm{ML}}}(f_i) \cdot \Delta t\right)^2$$

Once we get the fluence likelihood, it can be used in the **LDF fitting procedure** instead of χ^2minimization → uncertainties estimation

Testing the Rice distribution method for RD

Simulations

2000 proton/iron qgsjet sims. (RdIdealGrid)

E-field traces - EW polarisation

Excluding stations strongly affected by thinning

Simulated measurements

Noise library

RD traces from EA stations (2021-09 until 2022-08) cleaned from the showers signals, traces from the stations with broken LNA, duplicated traces and partially from corrupted traces

RECONSTRUCTION

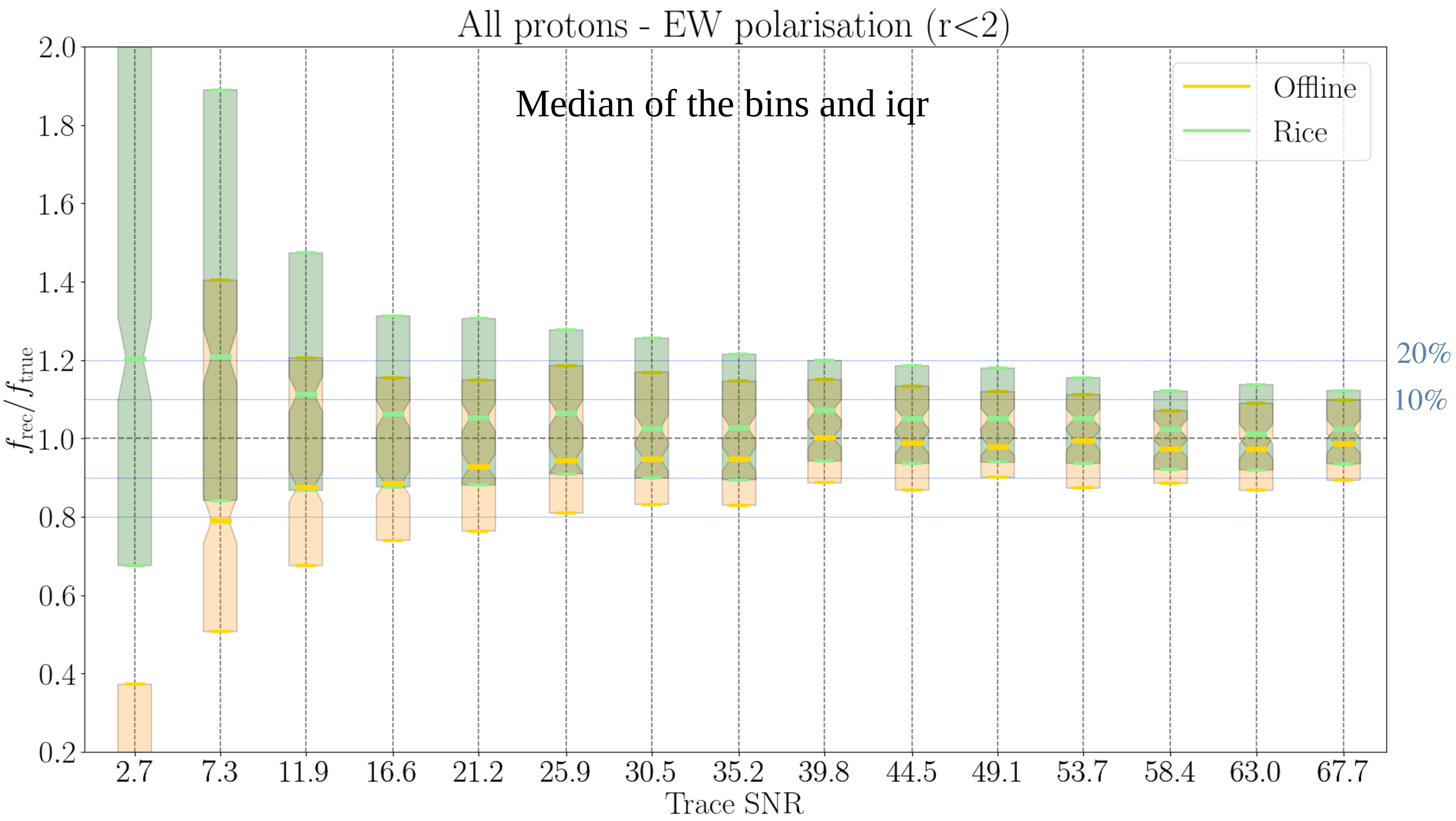
Rice method

$$\hat{f} = \epsilon_0 \, c \cdot 2 \, \Delta f \sum_{i=0}^{N_{\mathrm{bins}}} \left(\hat{s_{\mathrm{ML}}}(f_i) \cdot \Delta t \right)^2$$

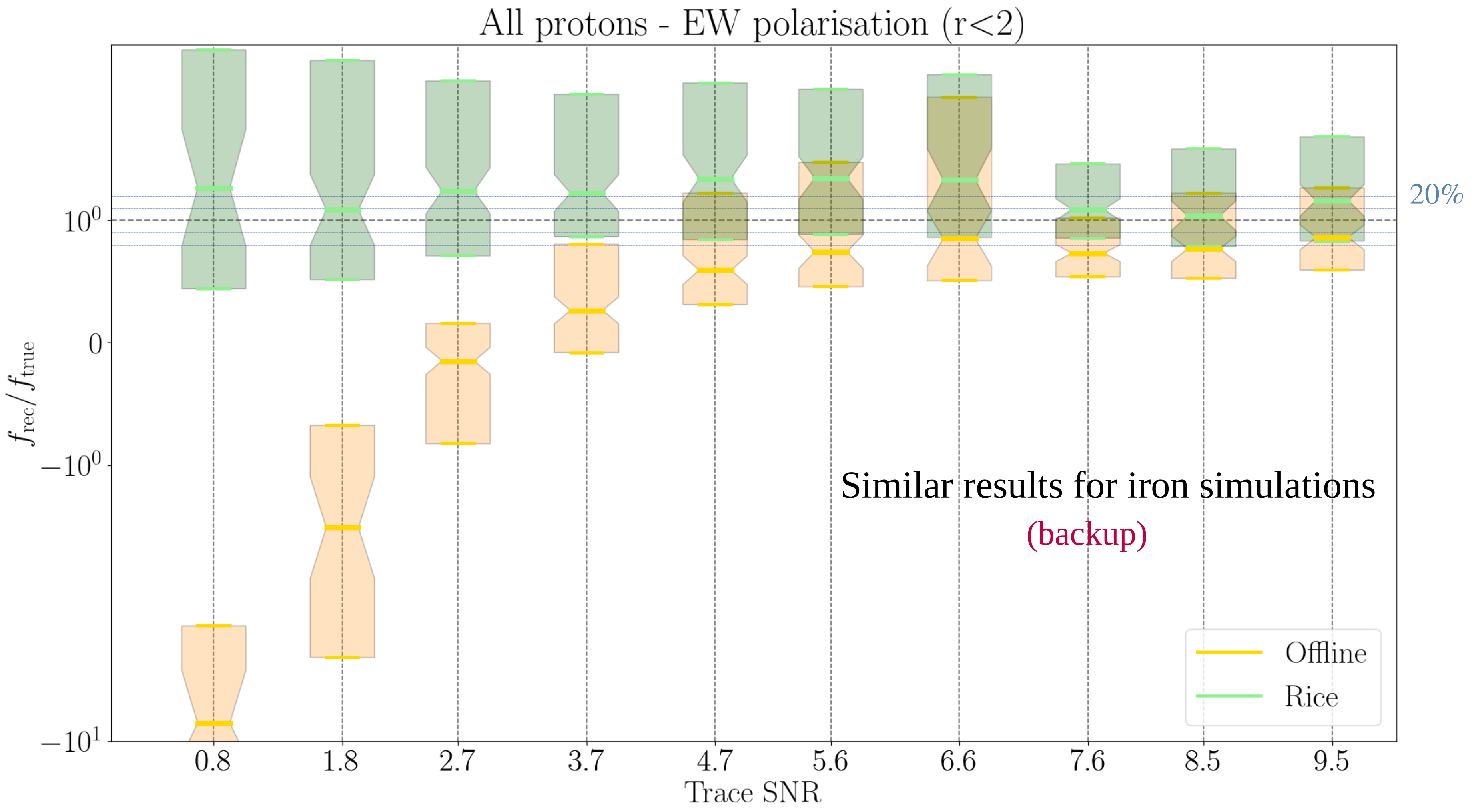
Offline current method

$$f = \epsilon_0 c \Delta t \left(\sum_{t_i=t_1}^{t_2} A(t_i)^2 - \frac{t_2 - t_1}{t_4 - t_3} \sum_{t_i=t_3}^{t_4} A(t_i)^2 \right)$$

Testing the Rice distribution method for RD



Testing the Rice distribution method for RD



Similar results for iron simulations

(backup)

Conclusions & Outlook

- The fluence estimation using the Rice distribution method shows a smaller bias than the actual Offline method for low SNR values, both for proton and iron simulations

- At higher SNRs, the bias is comparable

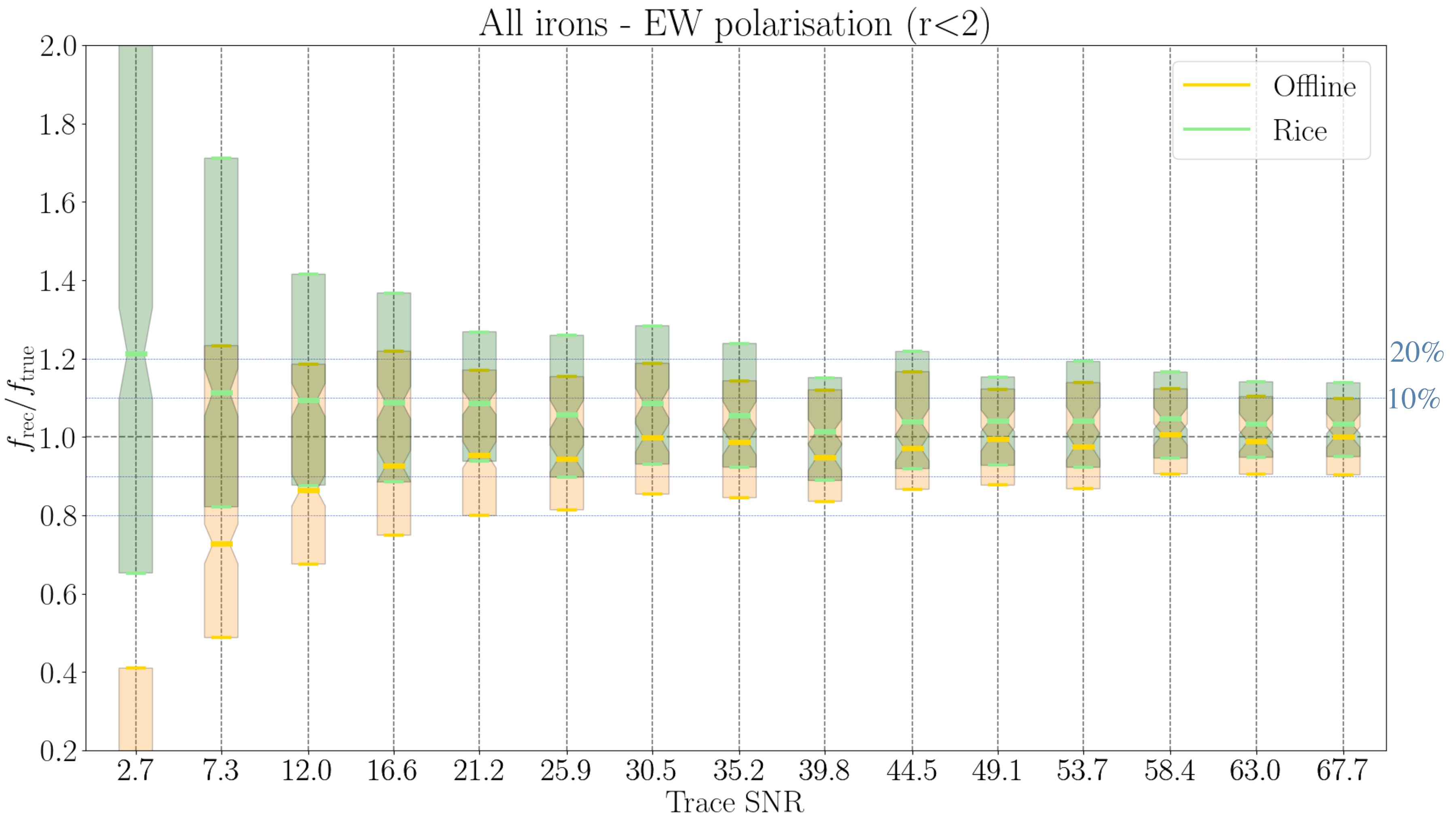
- By storing the Rice parameters we have access to the Likelihood

- LDF fitting using the Rice fluence Likelihood will provide uncertainties

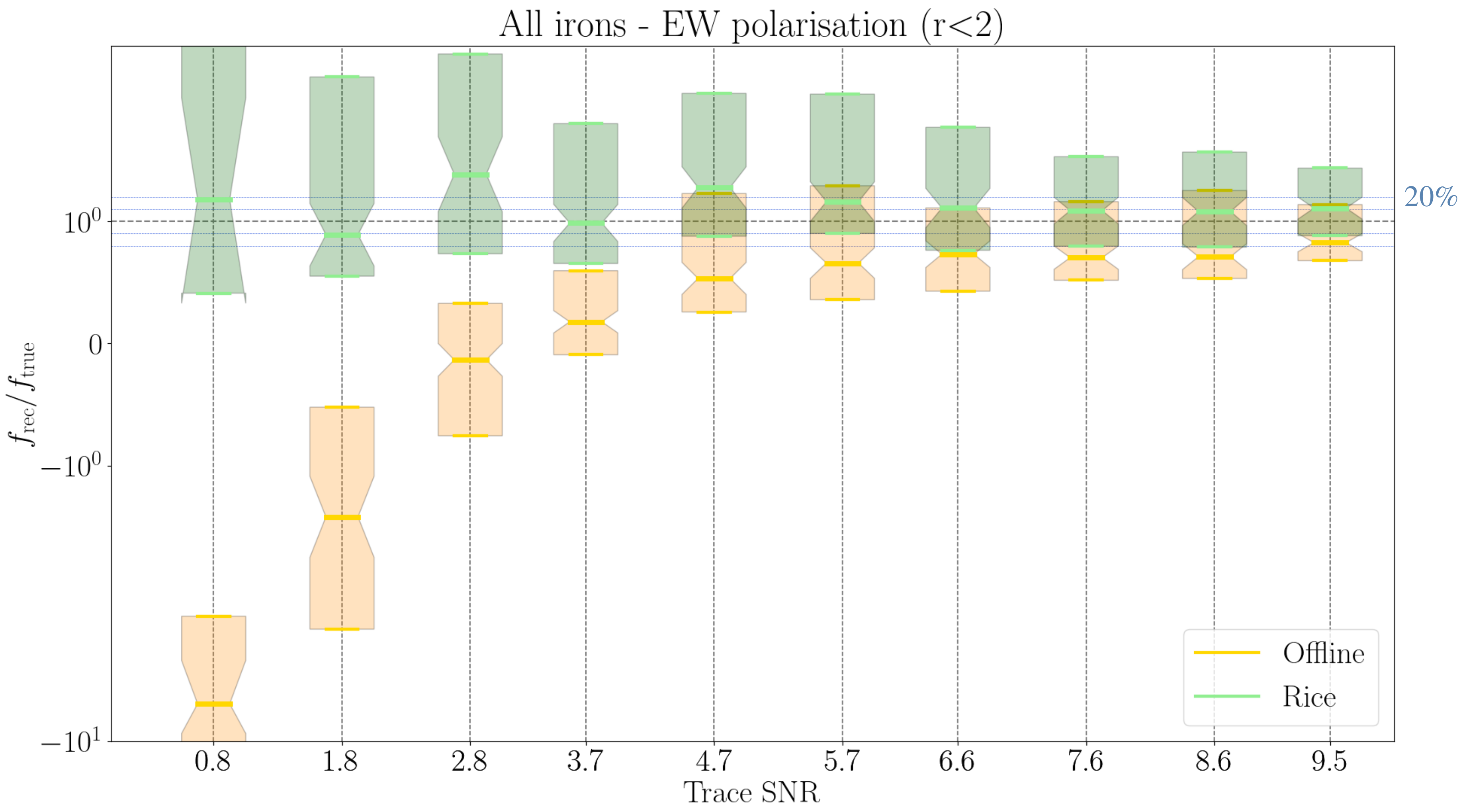
- For backward compatibility, we will store also the uncertainty on the fluence obtained by propagating the ones on the s estimators (work in progress, see backup)

Backup

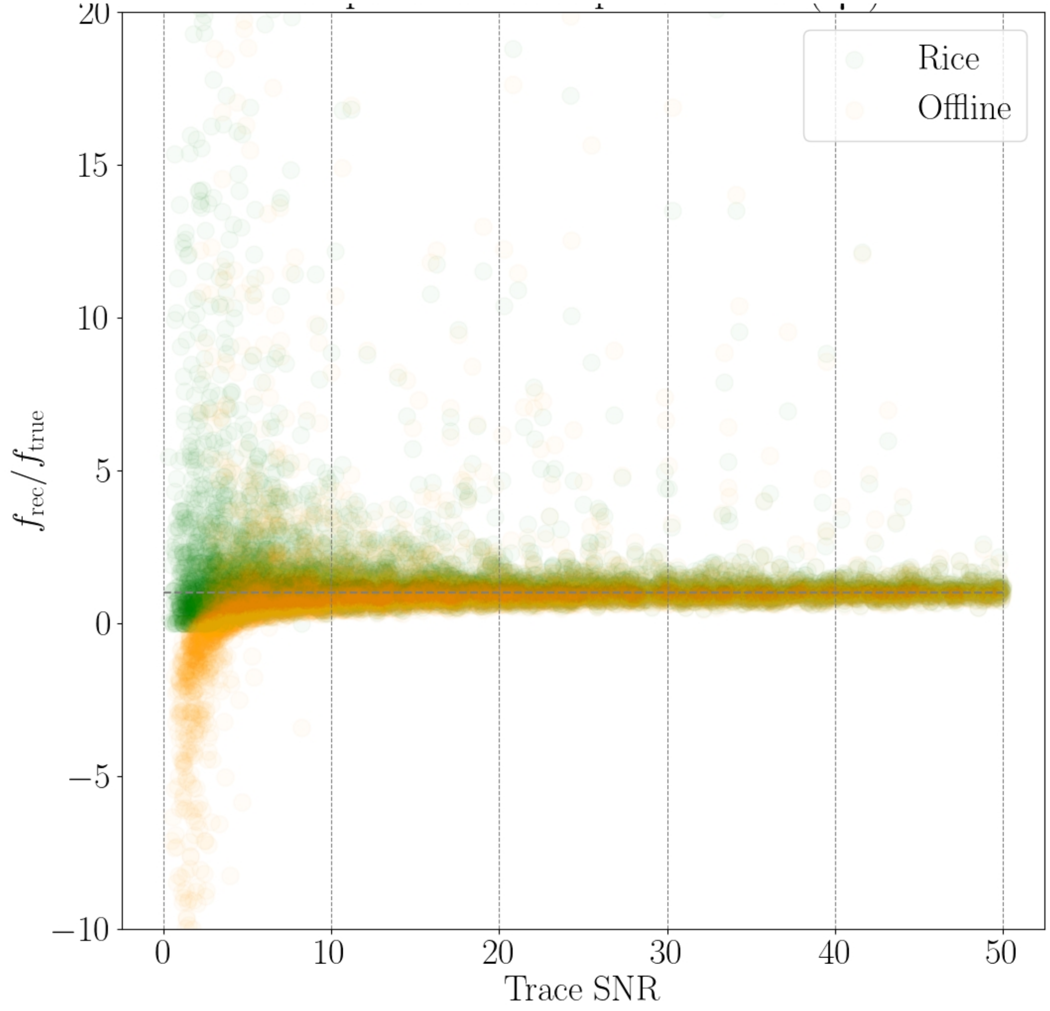
Testing the Rice distribution method for RD

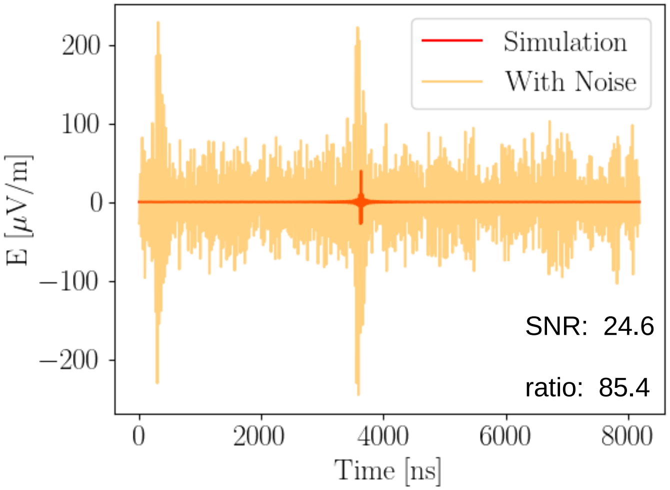


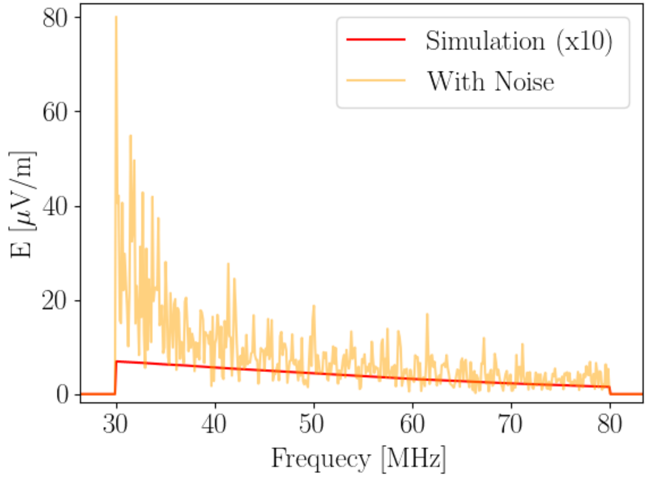
Testing the Rice distribution method for RD



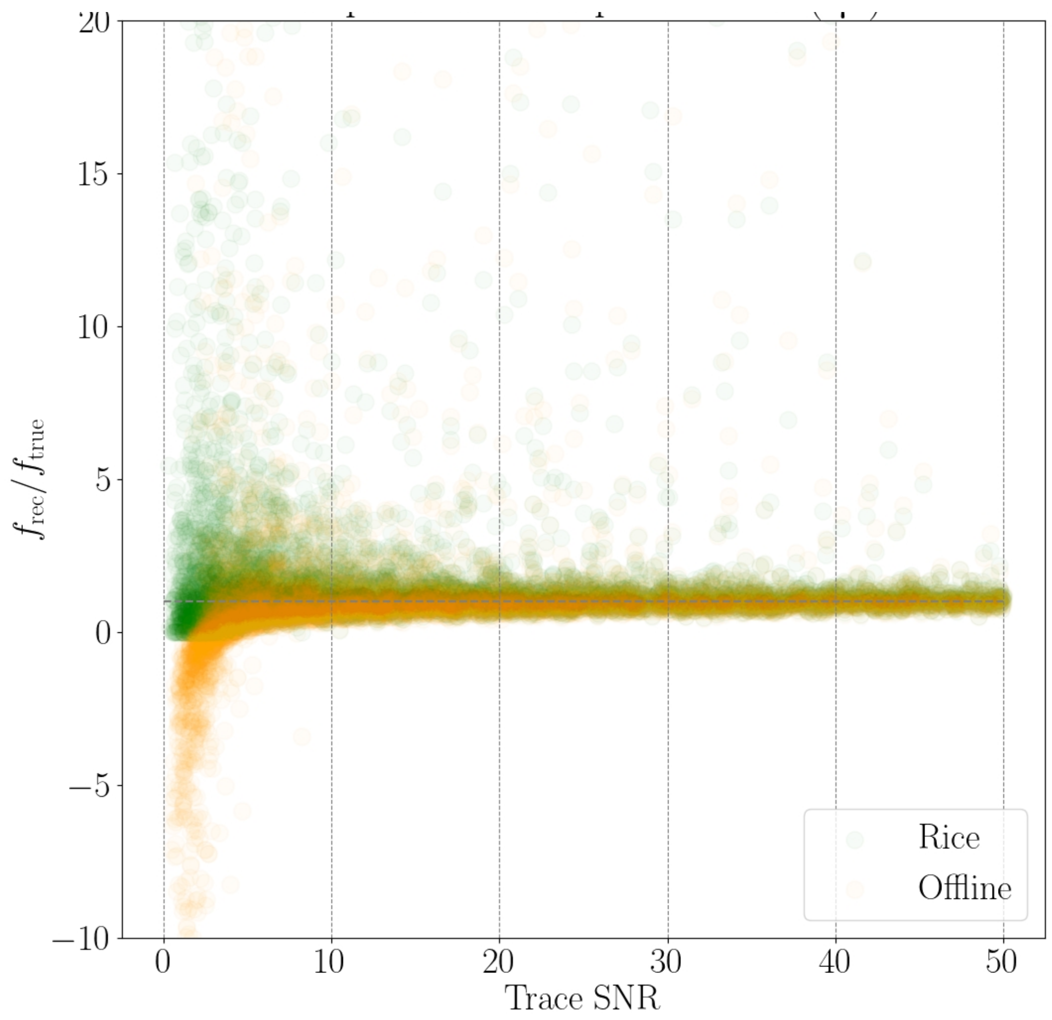
Testing the Rice distribution method for RD







Testing the Rice distribution method for RD

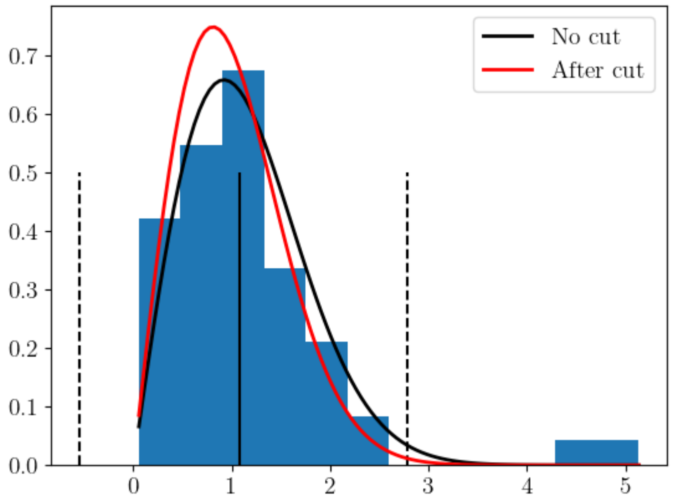


Cut on timing information

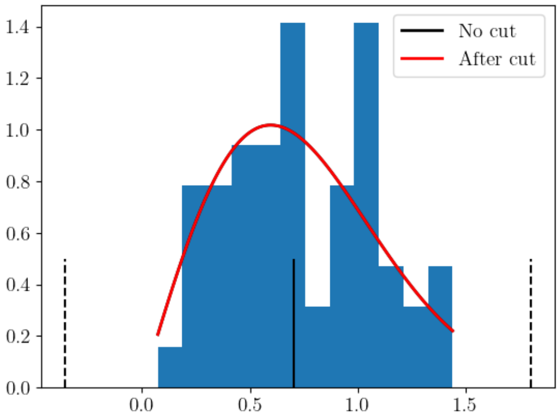
$|(T_{peak})^{REC} - (T_{peak})^{TRUE}| < 2ns$

in order to compare the goodness of the methods independently of the PulseFinder

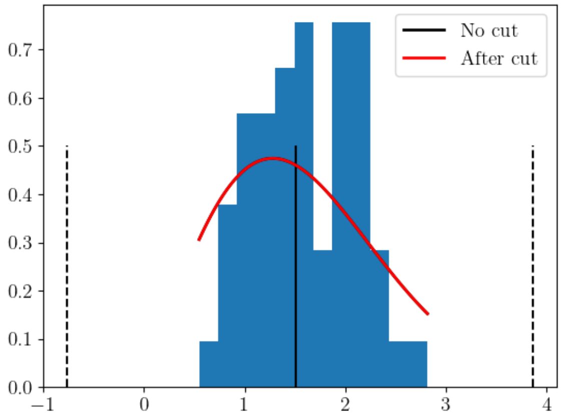
Noise windows and Rayleigh distributed noise assumption



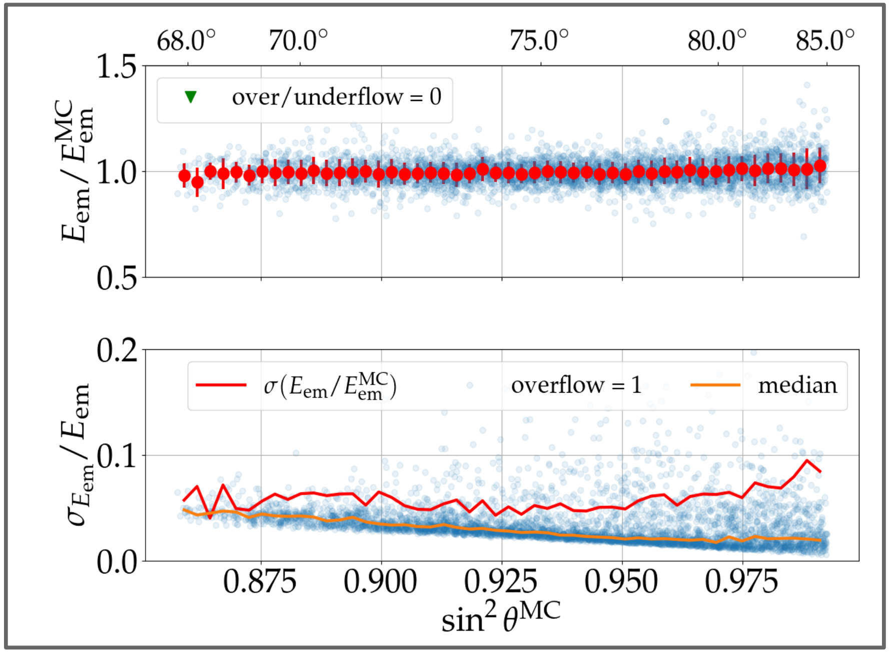
Cut of outliers in the frequency bin



Rayleigh distributed bins?

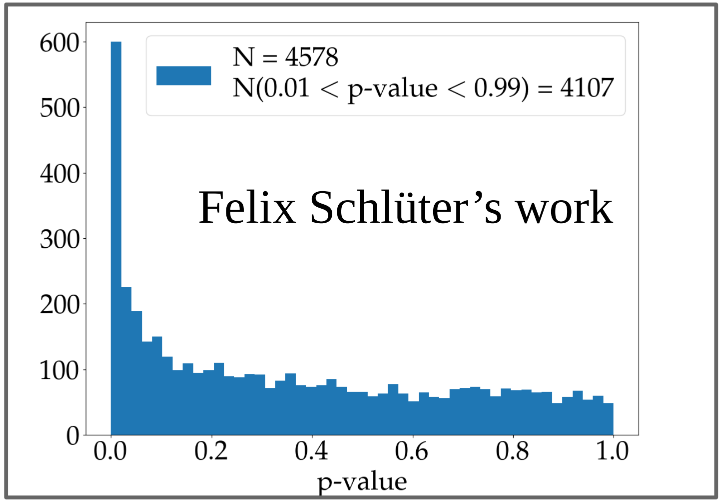


Reconstructed E_{em} and uncertainty



Resolution: how good we reconstruct the energy, std of Eem /EemMC in zenith bins

The relative uncertainties do not match the resolution → uncertainties underestimated



Goodness of the LDF fits:
too small p-values → uncertainties probably underestimated

Signal uncertainty estimation (Offline)

Uncertainty model derived in C. Glaser's PhD thesis

$$\sigma_{\mathrm{f}}^2 = \sigma_{\mathrm{Gauss}}^2 + \sigma_{\mathrm{det}}^2 + f_{\mathrm{noise}}^2$$

$$\sigma_{f_{\mathrm{Gauss}}}^2 = 4|f|\epsilon_0 c \Delta t V_{\mathrm{RMS}}^{\mathrm{noise}^2} + 2(t_2 - t_1)(\epsilon_0 c)^2 \Delta t V_{\mathrm{RMS}}^{\mathrm{noise}^4}$$

uncertainty due to noise after subtracting it, assumes amplitudes are superposition of signal and white noise Gaussian distributed

$$\sigma_{\mathrm{det}} = \sigma_{\mathrm{A}}^{\mathrm{detector}} \cdot 2 \cdot f$$

f~A² *5% uncertainty on the amplitudes*

$$\sigma_{\mathrm{A}}^{\mathrm{detector}} = 0.05$$ *(antenna variation)*

5.5.2 Uncertainty of the energy fluence

We estimate the uncertainty of the energy fluence by assuming that the measured electric-field amplitude $A(t)$ is the sum of the cosmic-ray radio pulse $S(t)$ and noise $e(t)$. Furthermore, we assume that the noise $e(t)$ is Gaussian distributed with mean $\mu = 0$ and standard deviation $\sigma = \sigma_e$. The energy fluence of A is then given by the equation

$$f(A) = \epsilon_0 c \Delta t \sum_{t_1}^{t_2} A(t_i)^2 = \epsilon_0 c \Delta t \sum_{t_1}^{t_2} [S(t_i) + e(t_i)]^2 = \epsilon_0 c \Delta t \sum_{t_1}^{t_2} \left[S(t_i)^2 + 2S(t_i)e(t_i) + e(t_i)^2\right] \quad (5.16)$$

and the expectation value of $f(A)$ is

$$\begin{aligned}\langle f(A)\rangle =& \epsilon_0 c \Delta t \sum_{t_1}^{t_2} A(t_i)^2 = \epsilon_0 c \Delta t \sum_{t_1}^{t_2} \left[\langle S(t_i)^2\rangle + 2\langle S(t_i)e(t_i)\rangle + \langle e(t_i)^2\rangle\right] \\ =& \epsilon_0 c \Delta t \sum_{t_1}^{t_2} \Big[\langle S(t_i)\rangle^2 + \underbrace{Var(S(t_i))}_{=0} + 2\langle S(t_i)\rangle \underbrace{\langle e(t_i)\rangle}_{=0} \\ &+ 2\underbrace{Cov(S(t_i), e(t_i))}_{=0} + \underbrace{\langle e(t_i)\rangle^2}_{=0} + \underbrace{Var(e(t_i))}_{\sigma_e^2}\Big] \\ =& \epsilon_0 c \Delta t \sum_{t_1}^{t_2} \left[\langle S(t_i)\rangle^2 + \sigma_e^2\right].\end{aligned} \quad (5.17)$$

Hence, the best estimate of the energy fluence of the radio signal S is indeed

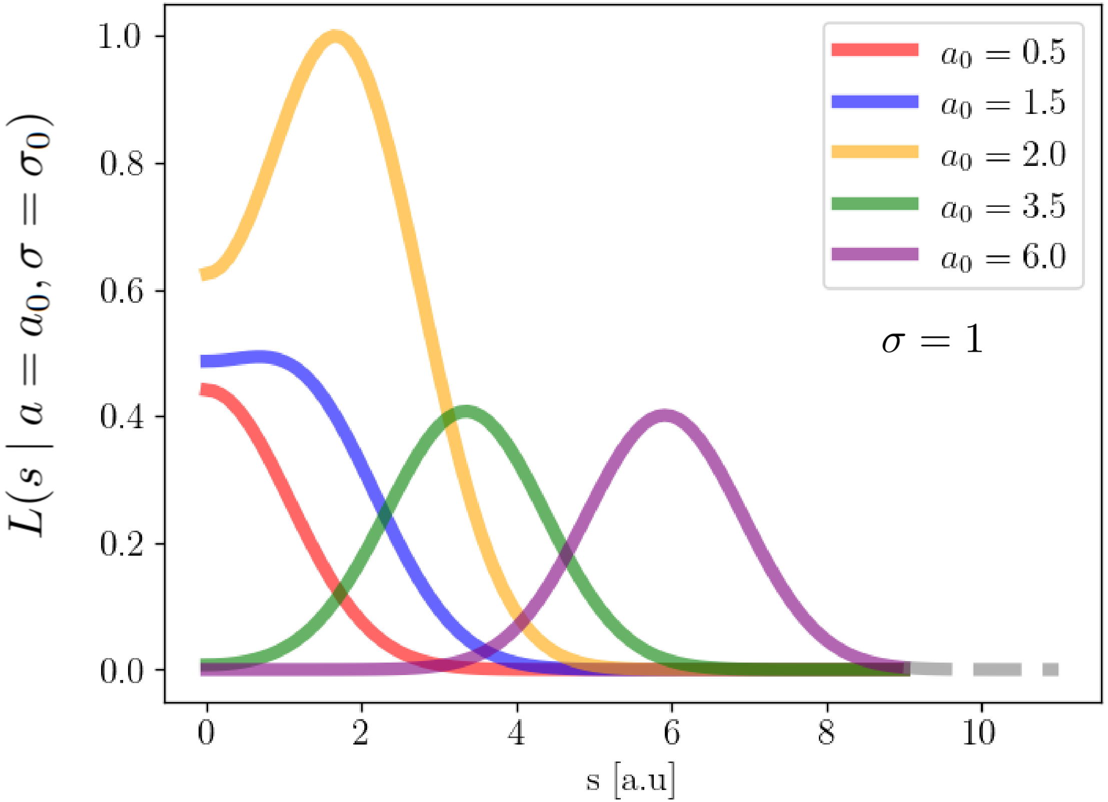
$$f(S) = \epsilon_0 c \Delta t \sum_{t_1}^{t_2} \left[A(t_i)^2 - \sigma_e^2\right] \quad (5.18)$$

as defined in Eq. (5.8) where σ_e^2 is also calculated from the electric-field trace in a part where no signal is present. Following a similar calculation we can estimate the uncertainty of $f(S)$ by computing $\sigma_f^2 = Var(f) = \langle f^2\rangle - \langle f\rangle^2$. After several lines of calculation it follows that

$$\sigma_f^2 = 4f\,\epsilon_0 c\,\Delta t\,\sigma_e^2 + 2\,(\epsilon_0 c)^2\,(t_2 - t_1)\,\Delta t\,\sigma_e^4. \quad (5.19)$$

Bias of the s estimators

$$L(s \mid a = a_0, \sigma = \sigma_0) = \frac{a}{\sigma^2} \cdot exp\left(-\frac{a^2+s^2}{2\sigma^2}\right) \cdot I_0\left(\frac{as}{\sigma^2}\right), \quad \text{s} \geq 0$$



What we actually need to know to understand the Likelihood estimator and its bias:

- $\text{a}/\sigma \gtrsim 3 \rightarrow \hat{s_{\text{ML}}} \approx \sqrt{a^2 - \sigma^2}$

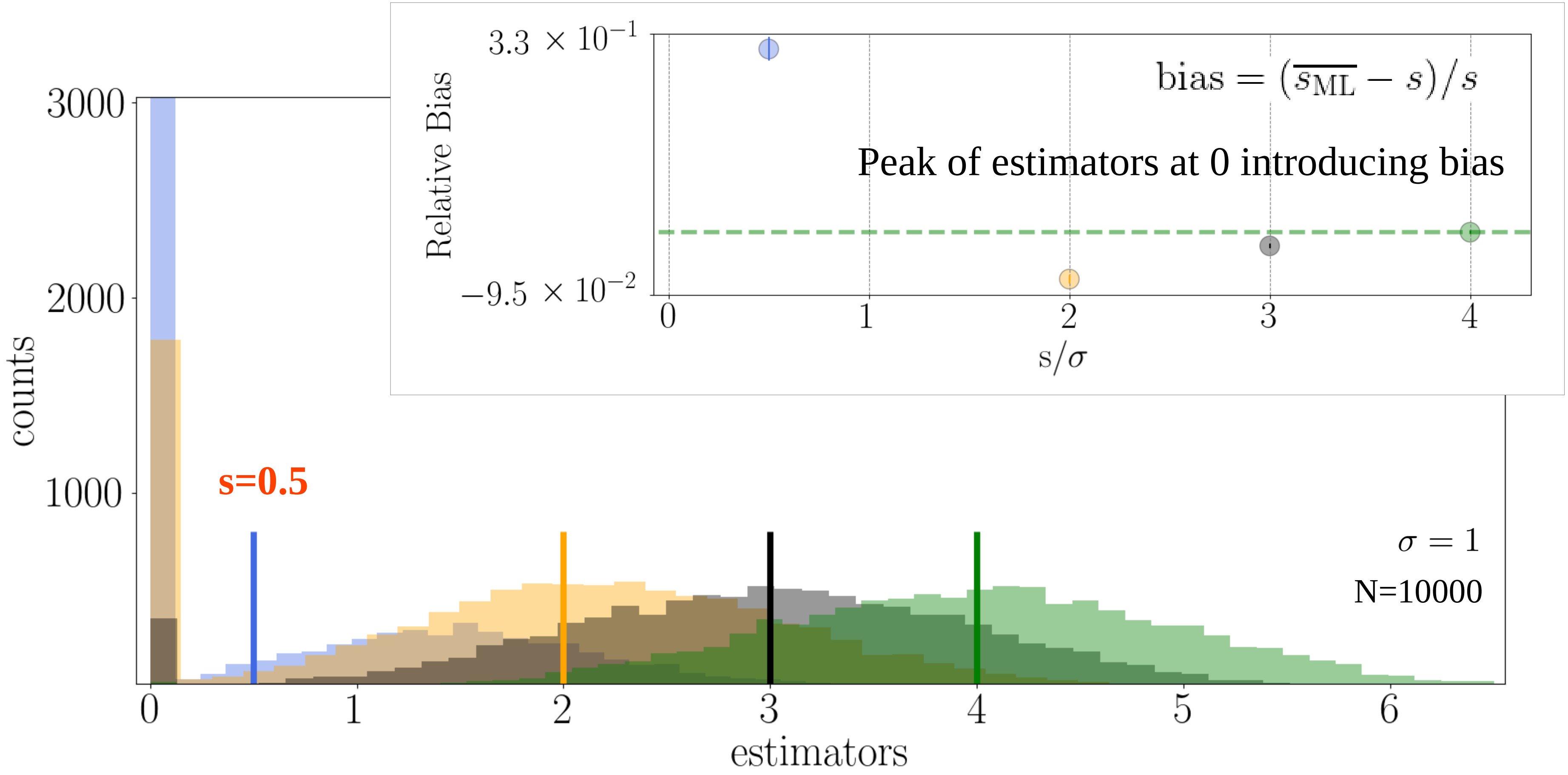
  $\mathcal{N}(\sqrt{a^2 - \sigma^2}, \sigma)$

- $a/\sigma \leqslant 1.4 \rightarrow \hat{s_{\text{ML}}} = 0$

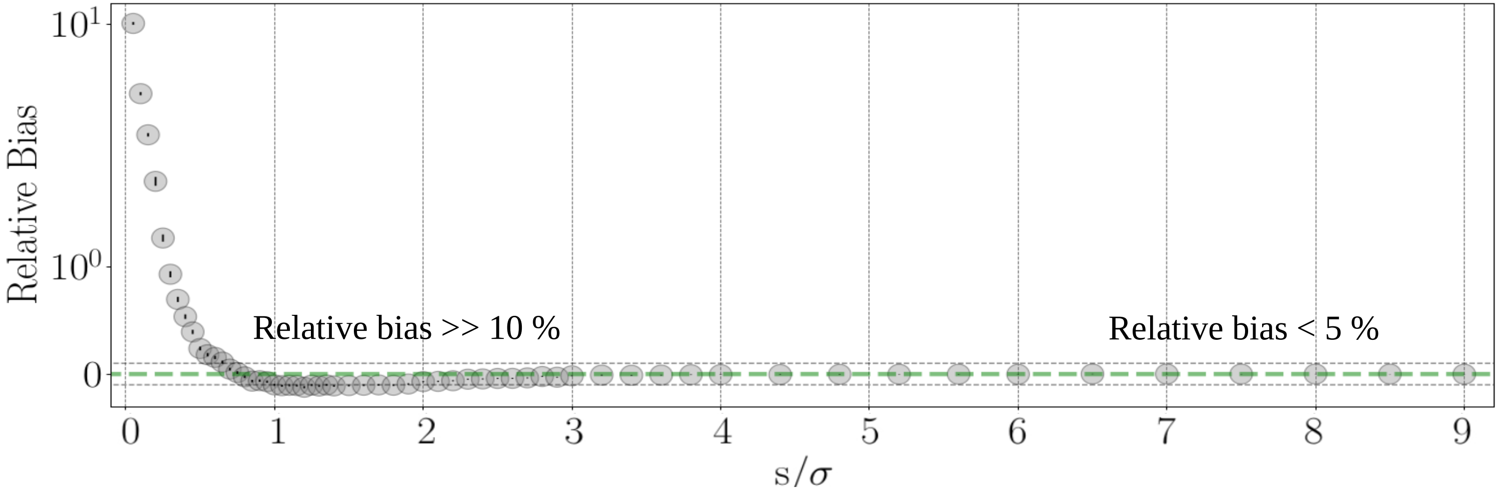
- Else:
  $a - 1.5\,\sigma < \hat{s_{\text{ML}}} < a + 1.5\,\sigma$

Note that the fluence will be always positive ($f = 0\,\text{eV}/\text{m}^2$ in principle possible!)

Bias of the s estimators

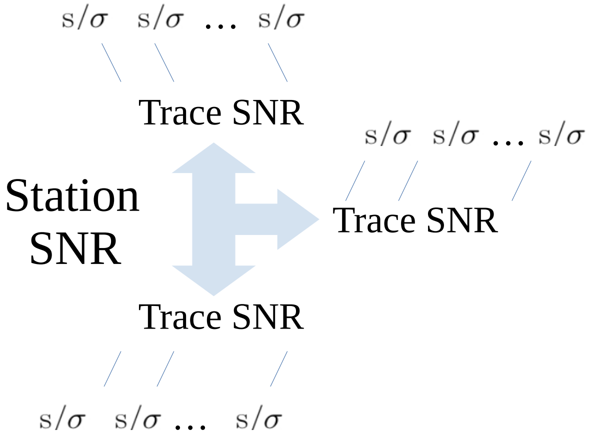


Bias of the s estimators



Verify how relevant the bias is for the fluence

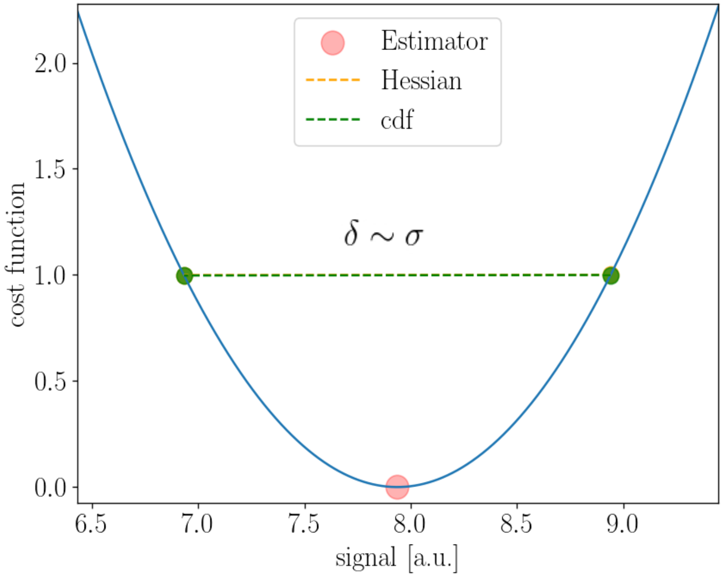
→ Large set of simulations + noise
(see previous slides)



Signal and signal uncertainty estimation with the Rice distribution

Once we get the fluence likelihood, it can be used in the **LDF fitting procedure** instead of χ^2 minimization → uncertainties estimation

For **backward compatibility**, we want to store in Offline the fluence and its error (e.g. get the uncertainties of the estimators of **s(f)** and propagate them to the fluence).



$$J(s) = -2\ln\left(\frac{L(s)}{L(\hat{s}_{\mathrm{ML}})}\right)$$

(backup)

$$J(s) = k^2$$

$$\delta = k\sigma$$

k=1 →

$$\delta = 1/\sqrt{\frac{1}{2}\frac{\partial^2 J}{\partial s^2}\Big|_{s=\hat{s_{\mathrm{ML}}}}}$$

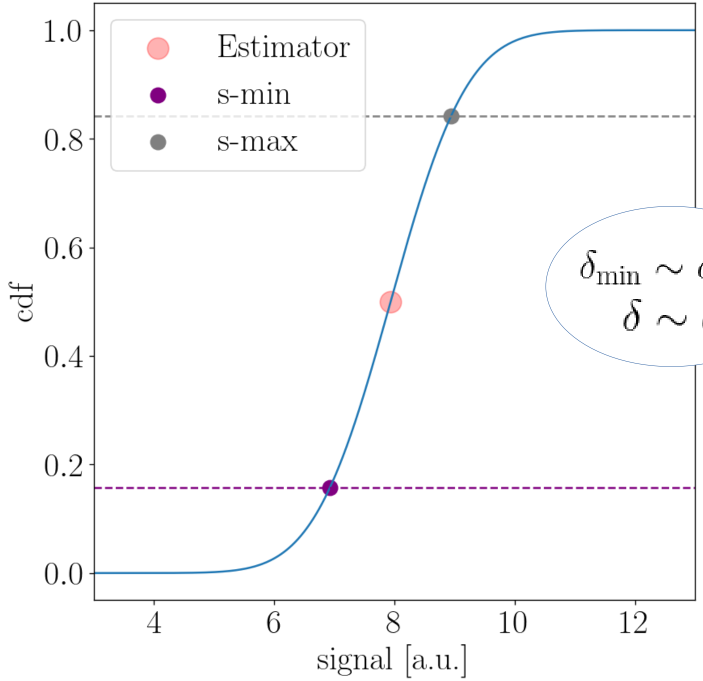
Gaussian approximation: $\delta \sim \sigma$

In non Gaussian approximation we would need to define asymmetrical errors to have a 68% interval

Estimators uncertainty

- Ideally we want to define a **68% interval** around the estimator.
  Looking at the cdf of the normalized Likelihood function we can distinguish two main cases:

  - Gaussian approximation (cdf of the estimator ~ 50%): symmetrical errors



$$\delta_{\text{min}} \sim \delta_{\text{max}}$$
$$\delta \sim \sigma$$

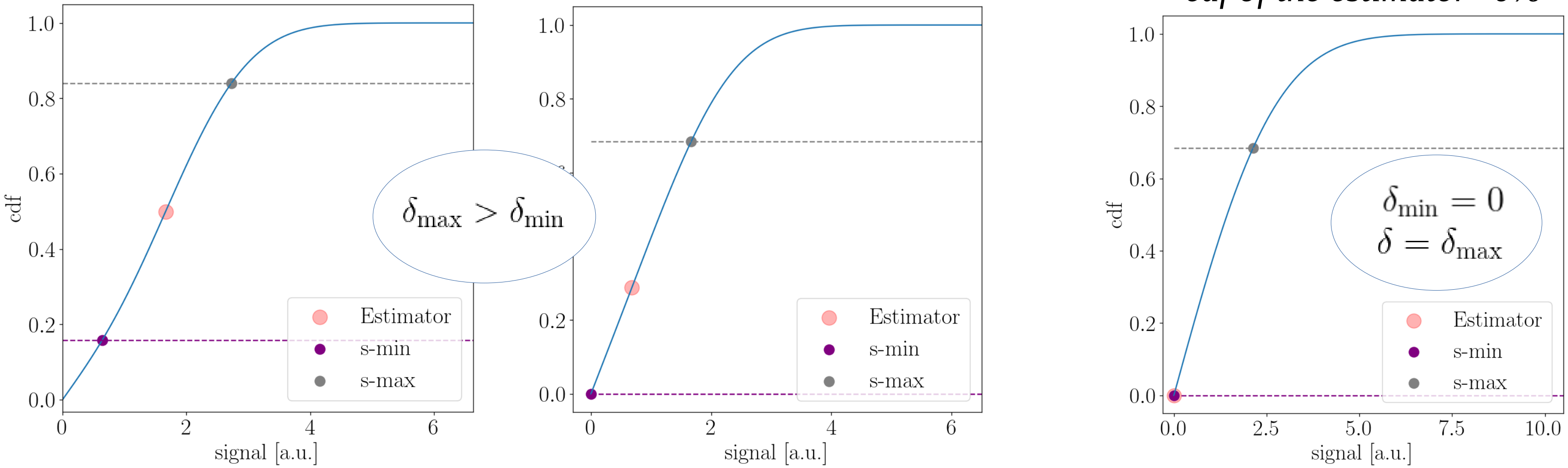
Estimators uncertainty

- Ideally we want to define a **68% interval** around the estimator.
  Looking at the cdf of the normalized Likelihood function we can distinguish two main cases:

  - Gaussian approximation (cdf of the estimator ~ 50%): symmetrical errors $\delta \sim \sigma$

  - Non-Gaussian: asymmetrical errors

0%<cdf of the estimator <50% *cdf of the estimator=0%*



Estimators uncertainty

- Ideally we want to define a **68% interval** around the estimator.
  Looking at the cdf of the normalized Likelihood function we can distinguish two main cases:

  - Gaussian approximation (cdf of the estimator ~ 50%): symmetrical errors $\delta \sim \sigma$

  - Non-Gaussian: asymmetrical errors $\delta_{\text{max}} > \delta_{\text{min}}$, $\delta = \delta_{\text{max}}$

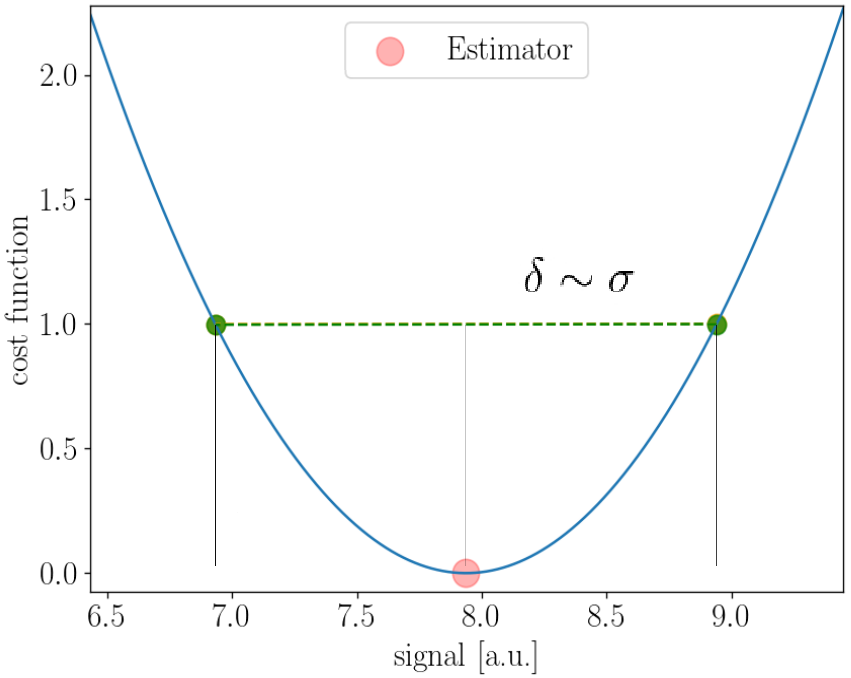
  But, we would like to store 1 single value (symm.) in Offline → $\delta = \delta_{\text{max}}$

- There's no analytical form of the cdf available → numerical integration
  - discrete n values
  - integrals in [i, i+1] , with i=0,..,n

Estimators uncertainty

- Hessian of the cost function $J(s) = -2\ln\left(\frac{L(s)}{L(\hat{s}_{\mathrm{ML}})}\right)$

$$J(s) \cong \cancel{J(\hat{s}_{\mathrm{ML}})} + \cancel{\frac{\partial J}{\partial s}\Big|_{s=\hat{s}_{\mathrm{ML}}}}(s - \hat{s}_{\mathrm{ML}}) + \frac{1}{2}\frac{\partial^2 J}{\partial s^2}\Big|_{s=\hat{s}_{\mathrm{ML}}}(\overbrace{s - \hat{s}_{\mathrm{ML}}}^{\delta})^2$$



Gaussian approximation:

$J(s) = k^2$

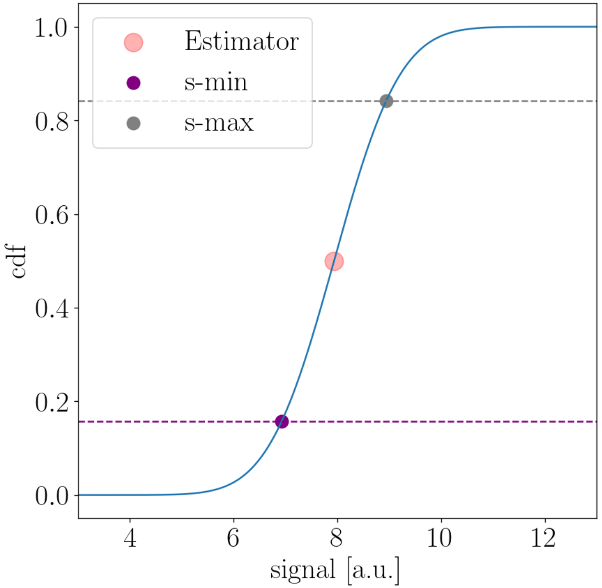
$\delta = k\sigma$

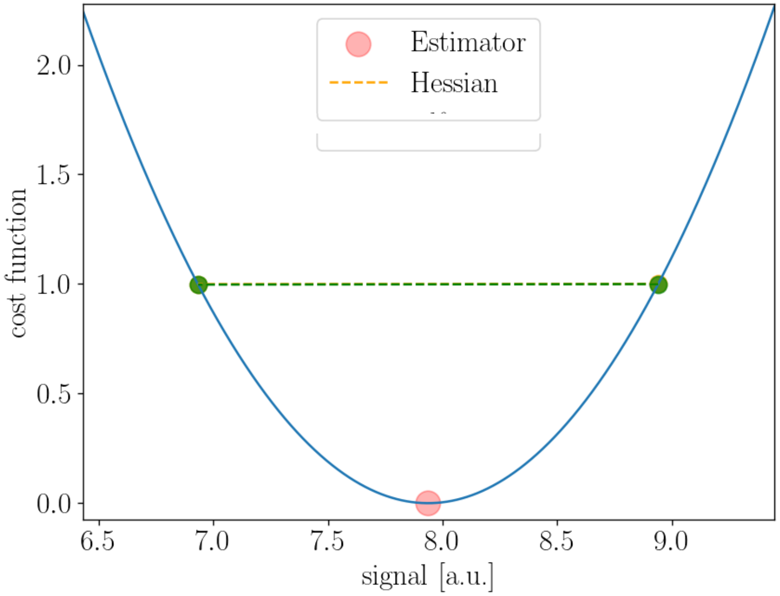
k=1 →

$$\delta = 1\Big/\sqrt{\frac{1}{2}\frac{\partial^2 J}{\partial s^2}\Big|_{s=\hat{s}_{\mathrm{ML}}}}$$

Estimators uncertainty: Hessian vs cdf

Gaussian approximation

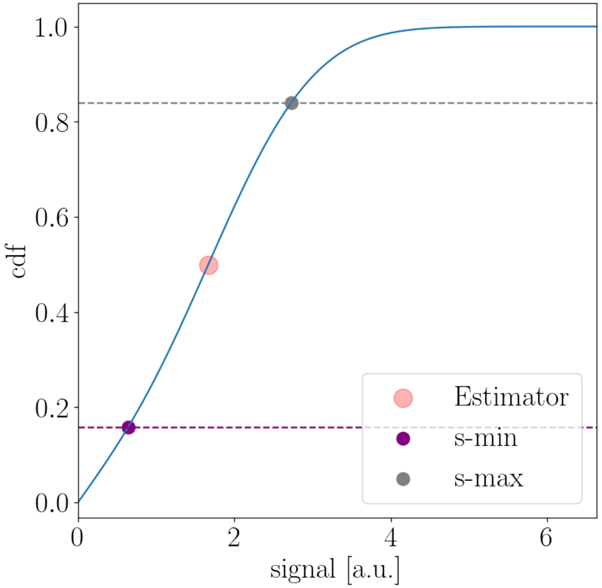


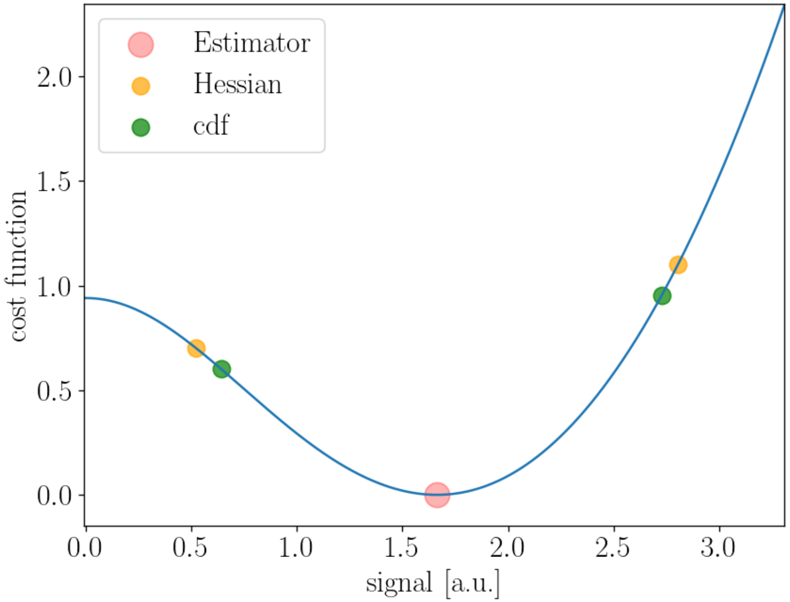


The two methods are equivalent (as it should)

Estimators uncertainty: Hessian vs cdf

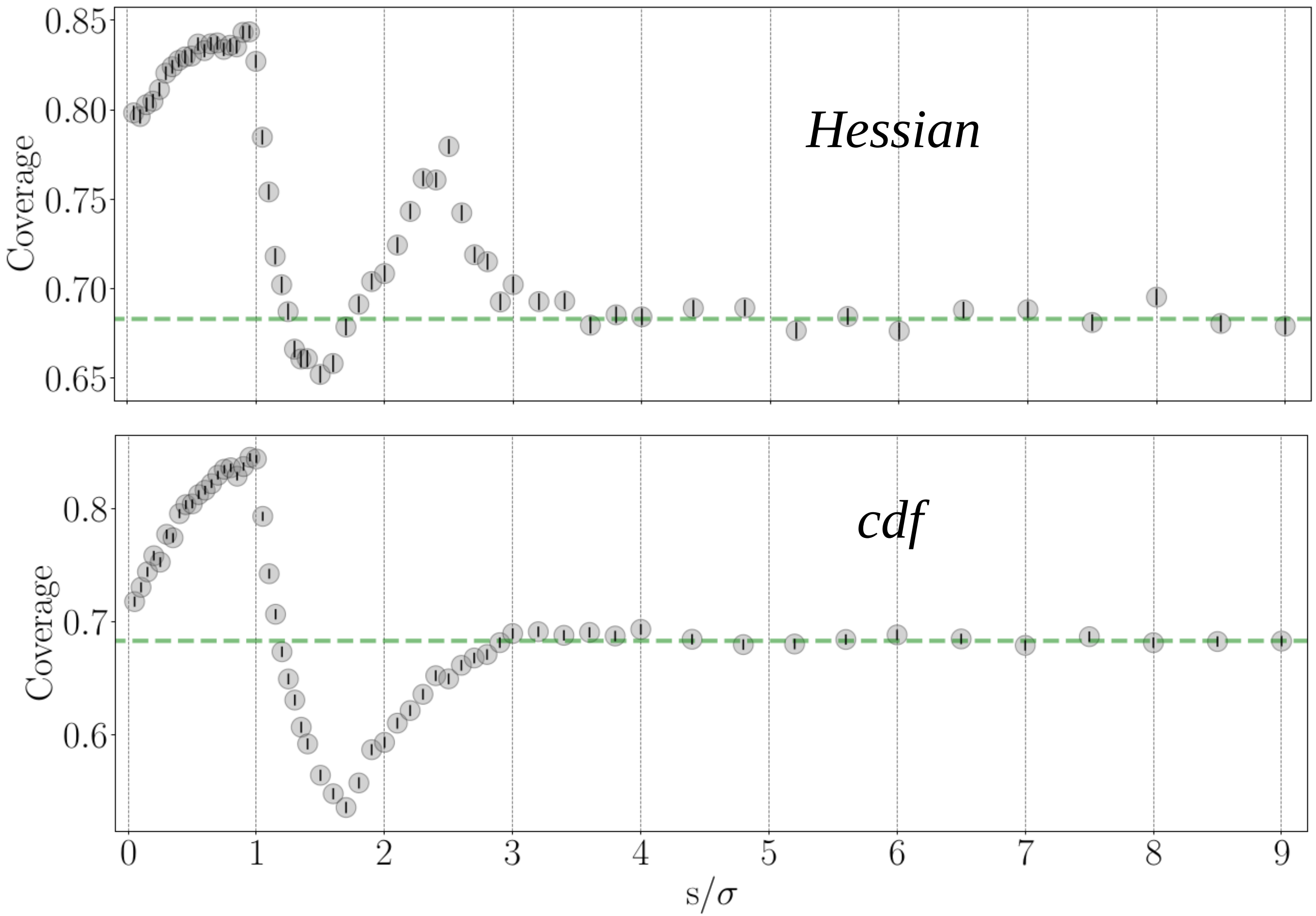
Non-Gaussian approximation





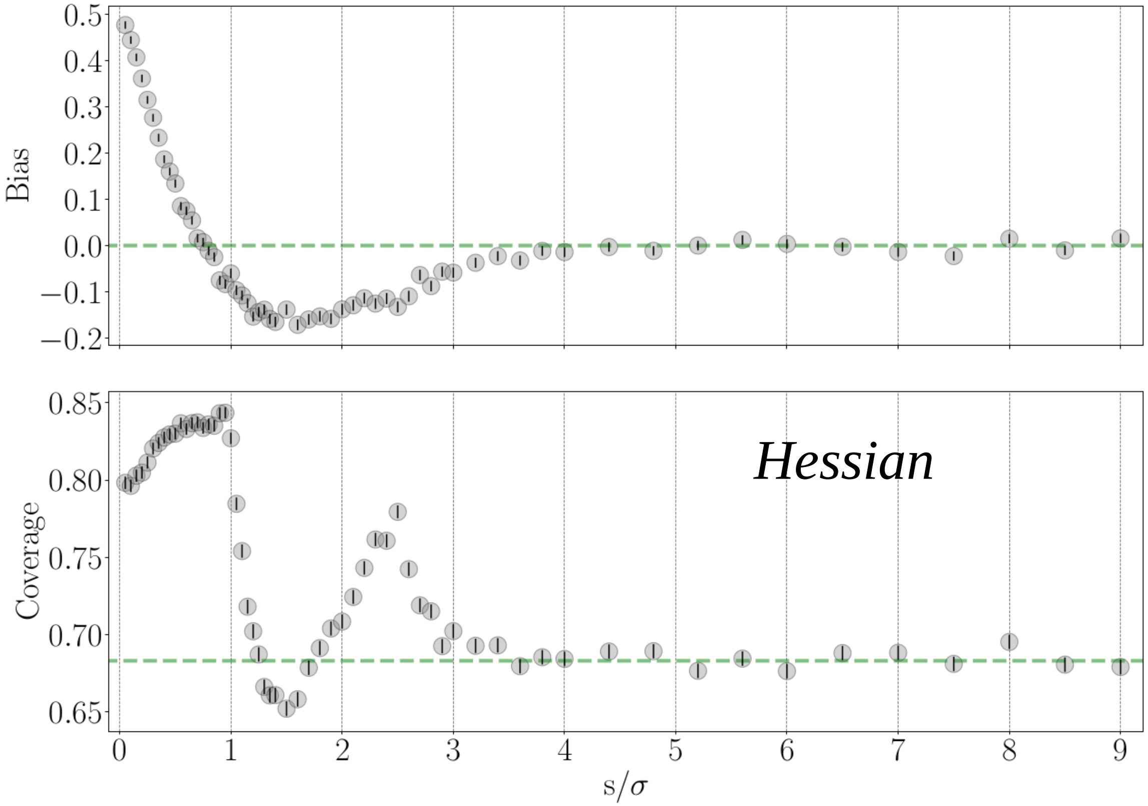
The two methods diverge

Estimators uncertainty: Hessian vs cdf

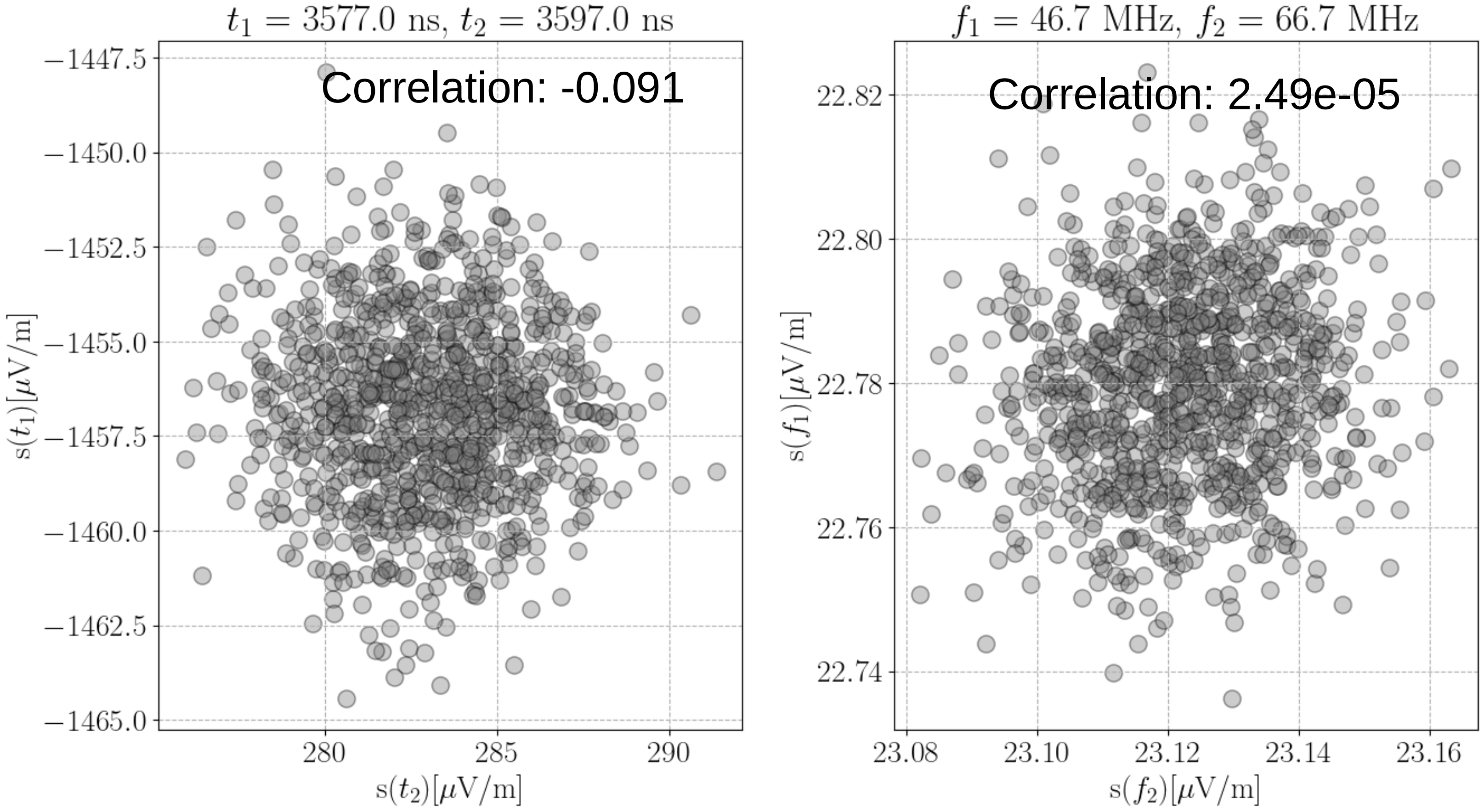


Estimators uncertainty

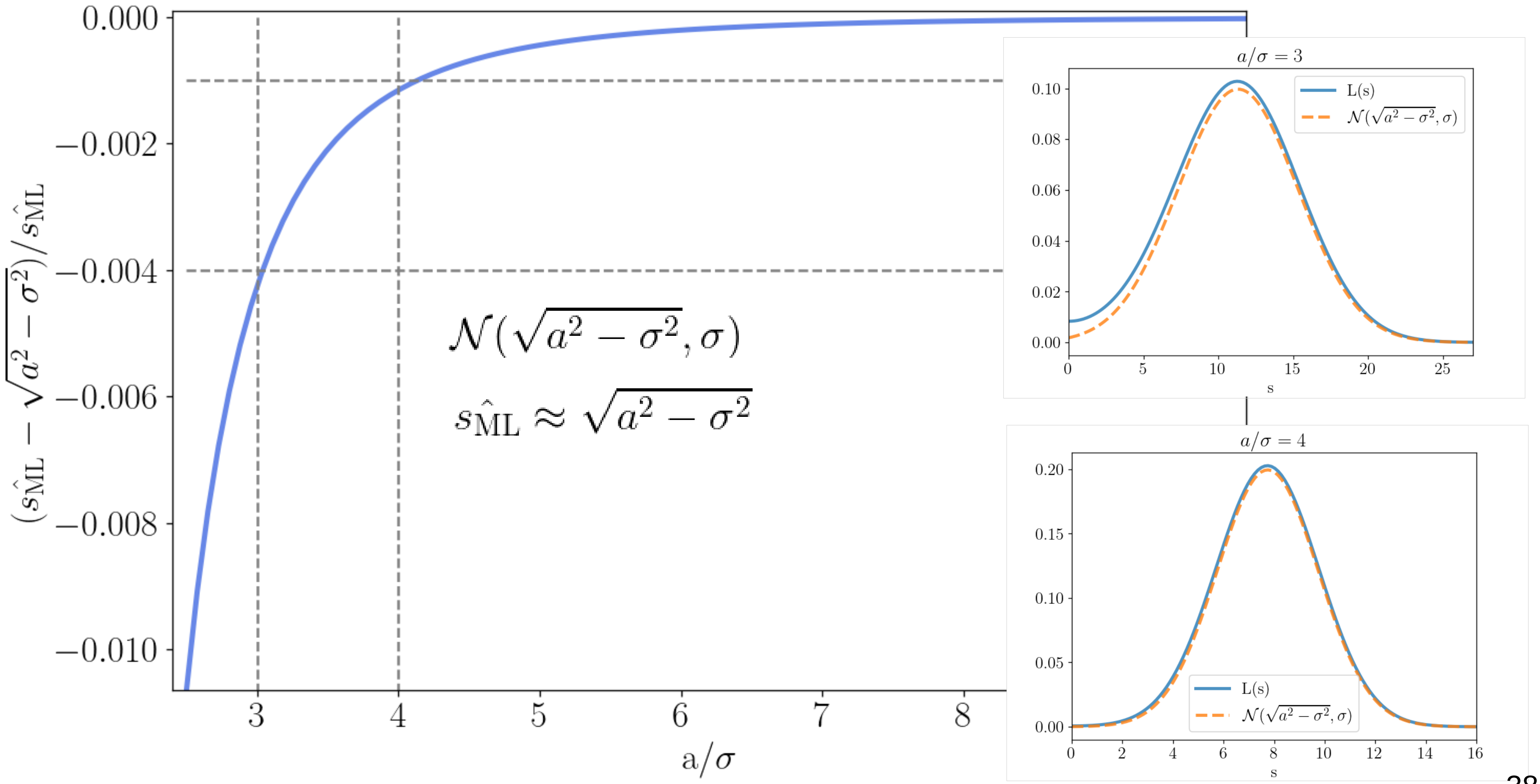
- The coverage fluctuate around 68% where the Gaussian approximation is valid. Large bias values mostly correspond to an overestimation of the error.



Correlation of the amplitudes of the signal

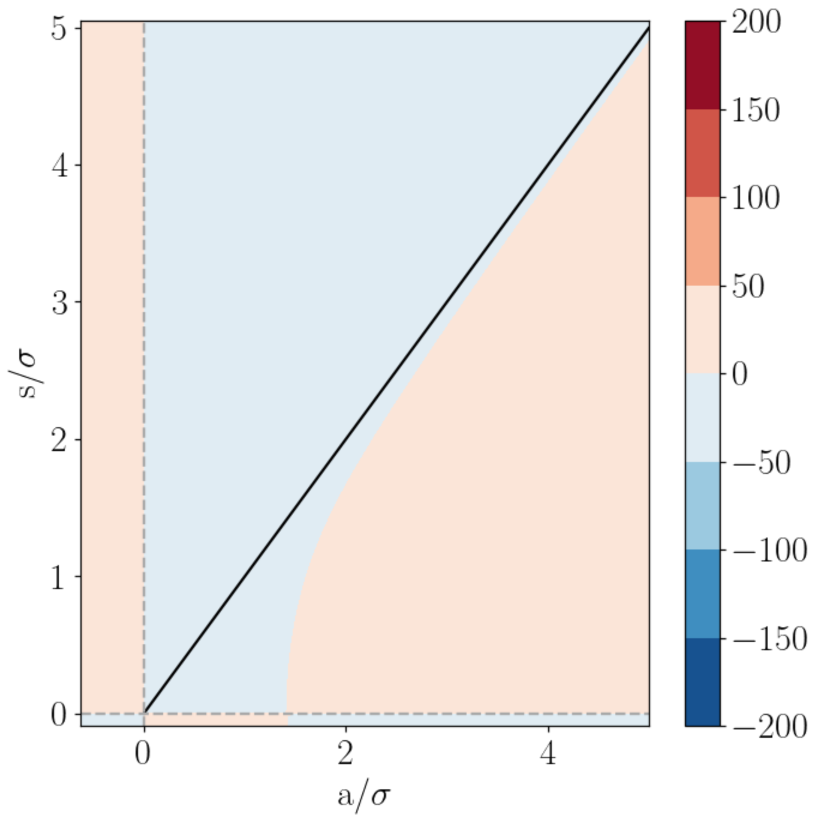


Gaussian approximation of L(s)



Rice method: the algorithm(s)

1. root finder between positive and negative value of dL/ds
2. minimizing -L with bounded method between smin, smax
3. run one of the previous methods, fit s vs a → in principle faster (less accurate...)



For the three of them, we need to study the solutions space

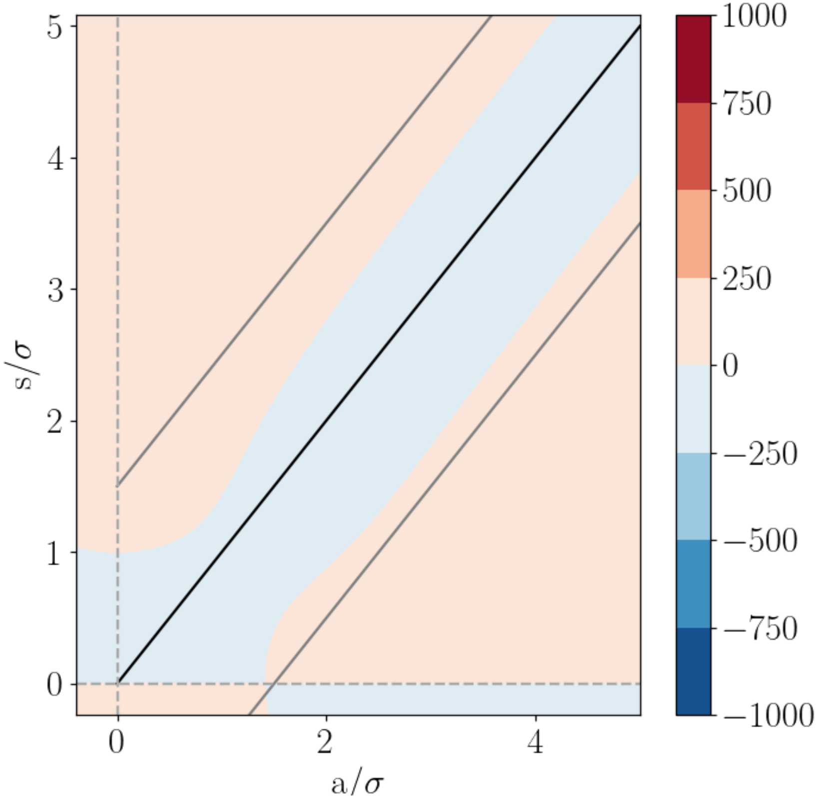
$$\frac{a^2 e^{\frac{-a^2-s^2}{2\sigma^2}} I_1\left(\frac{as}{\sigma^2}\right)}{\sigma^4} - \frac{ase^{\frac{-a^2-s^2}{2\sigma^2}} I_0\left(\frac{as}{\sigma^2}\right)}{\sigma^4}$$

$$a/\sigma \lesssim 1.4 \rightarrow \hat{s_{\mathrm{ML}}} = 0$$

$$a/\sigma \gtrsim 3 \rightarrow \hat{s_{\mathrm{ML}}} \approx \sqrt{a^2 - \sigma^2}$$

Rice method: the algorithm(s)

Second derivative to find min/max of the first derivative



$$\frac{a\left(a^2 s I_0\left(\frac{as}{\sigma^2}\right) - 2as^2 I_1\left(\frac{as}{\sigma^2}\right) - a\sigma^2 I_1\left(\frac{as}{\sigma^2}\right) + s^3 I_0\left(\frac{as}{\sigma^2}\right) - s\sigma^2 I_0\left(\frac{as}{\sigma^2}\right)\right) e^{-\frac{a^2+s^2}{2\sigma^2}}}{s\sigma^6}$$

$$a - 1.5\,\sigma < \hat{s_{\mathrm{ML}}} < a + 1.5\,\sigma$$