An alternative view of the 'Look Elsewhere Effect' through its manifestation in the search of modulations of unknown frequency in experimental time series data

PHYSTAT 2011

Gioacchino Ranucci
Istituto Nazionale di Fisica Nucleare – Milano

19 January 2011 - CERN

General framework for this illustration of LEE: search of modulations embedded in experimental time series

Method - computation of the power frequency spectrum → periodicities hidden in the noise affecting the data would appear as **sharp peaks in the spectrum itself**

In general Power Spectrum defined as $|H(f)|^2$
$H(f)$ is the Fourier transform of the series

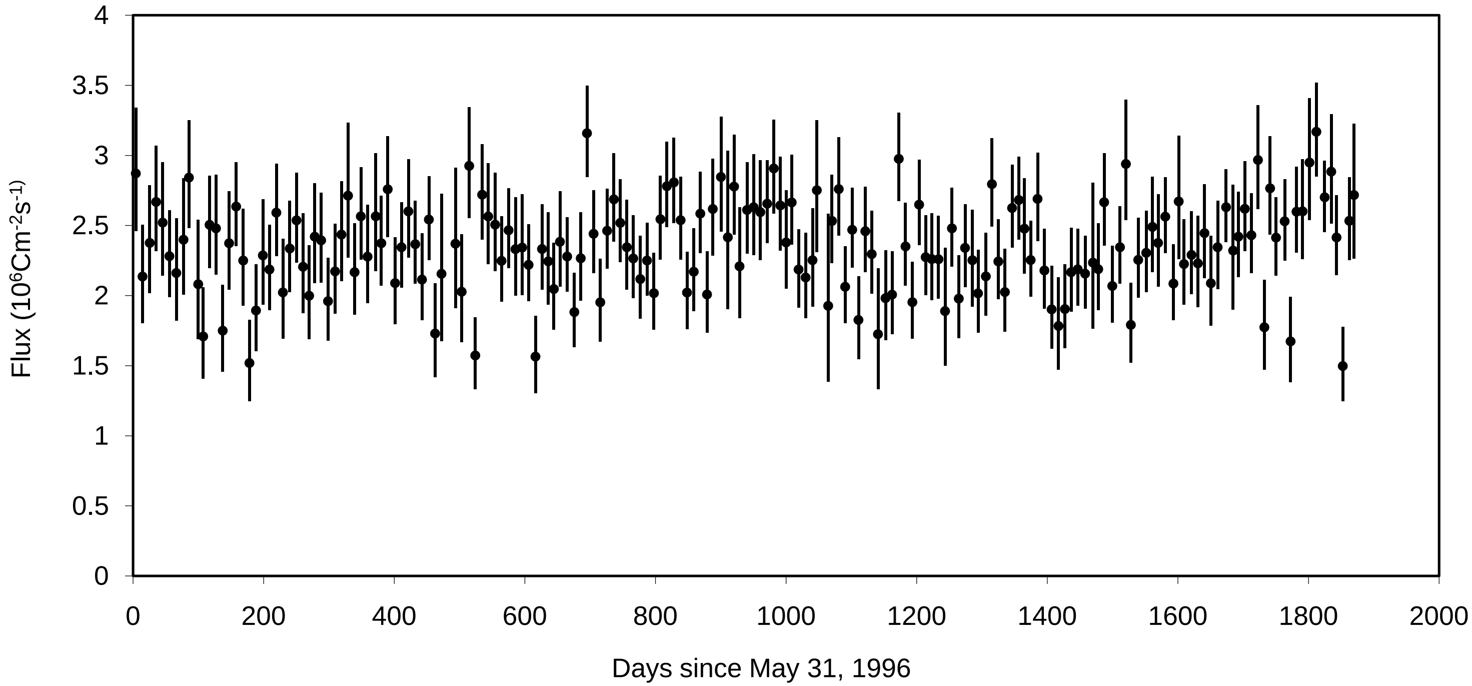
How is it practically evaluated ?

Popular approach in case of unsmoothed data (no tapering or windowing preliminarily applied) : Lomb-Scargle implementation, which takes into account the common situation in which the data are irregularly sampled or collected in time. Conventionally the computed power spectrum is referred as periodogram → **Lomb-Scargle periodogram**

It can be viewed as a generalization of the power spectrum estimate from the application of the Fourier transform to evenly sampled time series, sometime called **Schuster periodogram** — not frequently used but useful to understand the statistical features of this kind of applications

Concrete exemplification of the method

Given an experimental time series, like as concrete example the series of solar neutrino measurements from a real time solar neutrino experiment



Super-Kamiokande flux results for 10-day bins

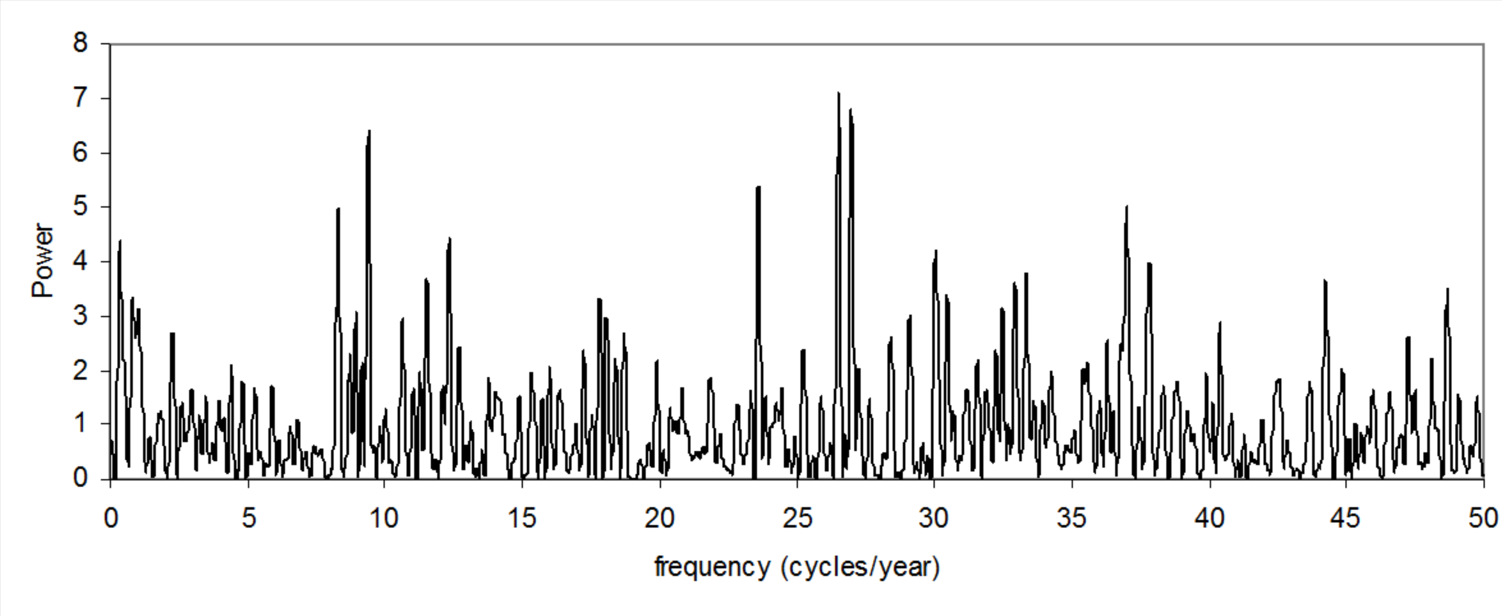
data officially released by the Collaboration

search for possible periodicities of <u>unknown</u> frequency hidden in the noise affecting the data

Note that we a-priori know in this data there is a modulation due to the Earth's orbit eccentricity of known annual period 3.3% modulation depth

Mapping of the series in the frequency domain

Power frequency spectrum (as said before often called conventionally periodogram) computed with some Fourier-related algorithm over a frequency search-band



Method naively simple to interpret → an unusual high frequency peak should indicate a modulation at that frequency embedded in the data

But noisy series → noisy spectrum featuring **random noise frequency peaks**

How to avoid to be fooled by the noise induced peaks while searching for a real signal

Crucial aspect of the analysis -> "significance" assessment of the peaks (in particular the largest) found in the estimated spectrum

"Significance" expressed by the probability that a peak as high or higher than the highest peak found in the actual spectrum, AT ANY SCANNED FREQUENCY, can be generated by chance noise fluctuations $\longrightarrow$ **p-value** intended as the probability content in the tail of the noise PDF above the actually detected value of the largest peak

Full analogy with the **Look Elsewhere Effect** while scanning a mass range for a possible particle of unknown mass: here the putative signal is searched at an **unknown frequency** location

Analytical formulation of the periodogram

t_k sampling times and x_k sampled values

$$\frac{1}{2\sigma^2}\left(\frac{\left(\sum_{k=1}^{N} x_k \cos\omega(t_k-\tau)\right)^2}{\sum_{k=1}^{N}\cos^2\omega(t_k-\tau)}+\frac{\left(\sum_{k=1}^{N} x_k \sin\omega(t_k-\tau)\right)^2}{\sum_{k=1}^{N}\sin^2\omega(t_k-\tau)}\right)$$

$$\frac{\sum_{k=1}^{N}\sin 2\omega t_k}{\sum_{k=1}^{N}\cos 2\omega t_k}=\tan 2\omega\tau$$

Lomb-Scargle periodogram
Uneven series

For series evenly sampled it reduces to **Schuster periodogram** from direct application of Discrete Fourier Transform

$$\frac{1}{N\sigma^2}\left[\left(\sum_{k=1}^{N} x_k \cos\frac{2\pi}{N}jk\right)^2+\left(\sum_{k=1}^{N} x_k \sin\frac{2\pi}{N}jk\right)^2\right]$$

The factors at denominator $2\sigma^2$ and $N\sigma^2$ are inserted for **normalization purposes**

Some general considerations

Lomb-Scargle periodogram -> **the common choice** (experimental data are usually unevenly sampled)

Schuster periodogram-> less used being valid only for strictly evenly sampled data
but of paradigmatic interest because of the clear way of describe its statistical properties very similar to those of the Lomb-Scargle

Important result: in both cases under the null hypothesis and for gaussian noise affecting the data the ordinate **z** of the spectrum at a generic frequency is distributed according to $\mathbf{e^{-z}}$ - no extra factors in the noise distribution because of the normalization terms (containing σ^2) at denominator

Gaussian noise in the time domain ⟶ noisy peaks exponentially distributed in the frequency domain

What about the errors and the variance at denominator of the formula

Drawback : the individual errors of each data point are not taken into account

The σ^2 term at denominator is inferred from the scatter of the data themselves

A more familiar for us likelihood ratio methodology originates a likelihood spectrum

⟹ generalization of the Lomb-Scargle spectrum

It accounts for the experimental errors, also if asymmetric

Maintains very similar statistical properties thus methodologically does not require an additional statistical discussion (some more mention later)

How to unravel quantitatively a signal embedded in the series

Let's consider the search of a signal at a predefined frequency - reasons to presume a signal at that frequency – **no LEE**

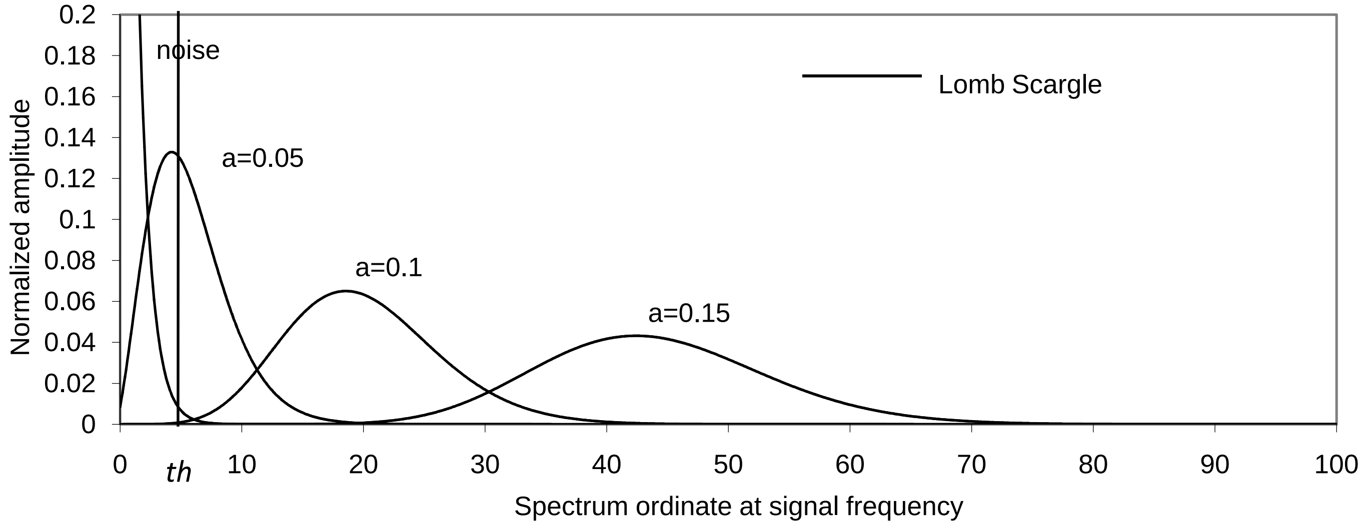
Ingredients of the detection problem

a) Distribution of the ordinate at that frequency in case of absence of the signal – null hypothesis
b) Distribution of the ordinate at that frequency in case of presence of the signal – alternative hypothesis

Case a) noise distribution simply e^{-z} - z is used to denote the spectrum ordinate

Case b) the spectrum ordinate obeys to a distribution belonging to the family of non central chi^2 distributions (2 degree of freedoms)

Predicted signal and noise distributions for a series like SK- no LEE



a modulation amplitude relative to the mean value

Once defined the detection threshold
-size of the test is the probability content of the tail of the exp. distribution above the threshold, i.e. the significance level of the detection (usually viceversa *th* determined by the desired significance) → also called type I error
-power of the test is the probability content of the non central chi2 square above the threshold → 1-power called type II error

The former unambiguously defined - only by the noise
The latter obviously no – exception annual modulation due to the Earth's orbit eccentricity

How the previous detection scenario is modified by not knowing the frequency of the putative signal

Not knowing where to expect the signal a whole frequency range is spanned for the search

The criterion to assess the presence of a signal is based upon the height of the largest peak detected in the searched frequency range ⟶ **test statistics**

Multiple hypotheses test -> **the Look Elsewhere Effect is in action !**

Once the largest peak is observed the associated significance is the "tail probability content", above the value of the peak itself, of the PDF of the highest peak generated over the search band by a pure noisy series (null hypothesis) this is the probability that a peak as high or higher than that actually observed is caused by a chance noise fluctuation

How such a PDF looks like? Example of the Schuster periodogram

Paradigmatic example : Schuster periodogram

In case of the Schuster periodogram - direct application of the DFT to an evenly sample series with N number of samples - an important frequency analysis result is that the spectrum , meaningfully computed up to the Nyquist frequency 1/(2T) (T sampling interval), comprises only N/2 independent frequencies

Therefore the spectrum is made of M=N/2 independent ordinates each distributed according to e^{-z} under the null hypothesis (pure noise series with no modulation embedded)

PDF of the height *z* of the largest peak ⟶ $M\left(1-e^{-z}\right)^{M-1}e^{-z}$

Largest among the M peak

The remaining M-1 peaks are not above *z*

Number of possible choice for the largest peak

Significance in case of M independent frequencies

Let's denote with H the largest peak actually detected in the spectrum

The significance of the signal detection is the **tail probability content (p-value)** of the previous PDF above H, i.e.

$$\int_H^{\infty} M\left(1-e^{-z}\right)^{M-1} e^{-z} dz = 1-\left(1-e^{-H}\right)^{M}$$

Bonferroni-type formula, typical of multiple hypotheses testing problems

Generalization to the PDF of all peaks, 2nd highest 3rd highest and so on

Rarely used the fact that the previous formalism extendable to peaks of any rank, PDF given by

.

$$p_i(z/M) = \frac{M!}{(i-1)!(M-i)!}\left[1-F(z)\right]^{(M-i)}\left[F(z)\right]^{i-1} p(z)$$

i order of the peak , *i=1* lowest peak and so on up to *i=M* highest peak

$$F(z) = \int_0^z p(\lambda)d\lambda$$

Valid for any form of *p(z)*

$p(z)=e^{-z}$ in the problem under evaluation

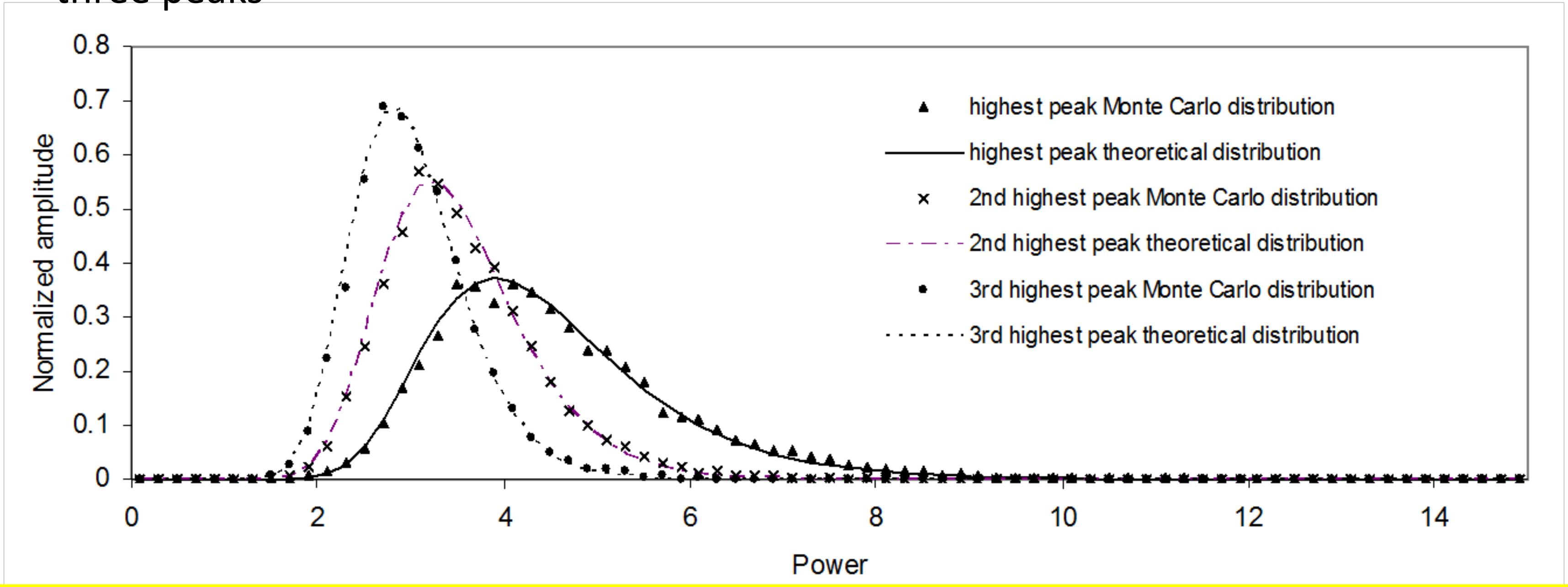
It reduces to the formula in the previous slides for *i=M*

p(z) can be for example a multi-exponential decay in case of the study of the timing statistics in photoelectron counting, where the same formula applies as well

Monte Carlo example

Monte Carlo example useful to validate the analytical model and also because MC frequently used in multiple hypotheses testing situation via the simulation of many pseudo-random experiments

- •Generation of “many” synthetic series (10000)
- •100 equally spaced sampling points → Schuster periodogram
- • Other characteristics **as the variance of each data** point chosen similar to the SK data
- •For each generation the spectrum is computed and the height of the three highest peaks recorded
- •The resulting histograms are compared with the analytical form of the PDFs of the three peaks

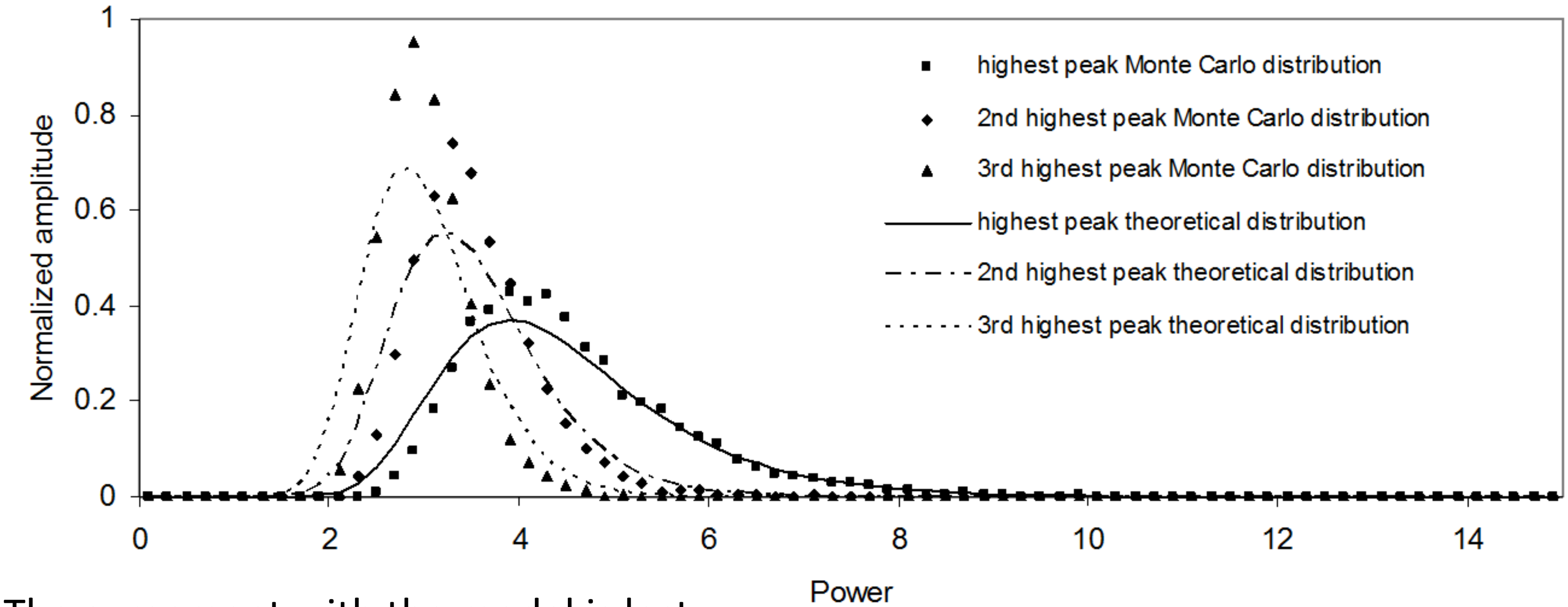


Extremely good agreement for M=50 – *Power* on the *x* axis is the *z* variable in the previous formulas - Largest peak significance assessment either analytically or MC

Role of the variance

Previous agreement if σ is a-priori known → enter the calculation of the periodogram

If instead σ is estimated from the data itself – usually the only possibility - this is the outcome



The agreement with the model is lost

Still, the model could be reasonable fine for significance assessment of the highest peak - the MC and model tails coincide - better to have a high statistics MC

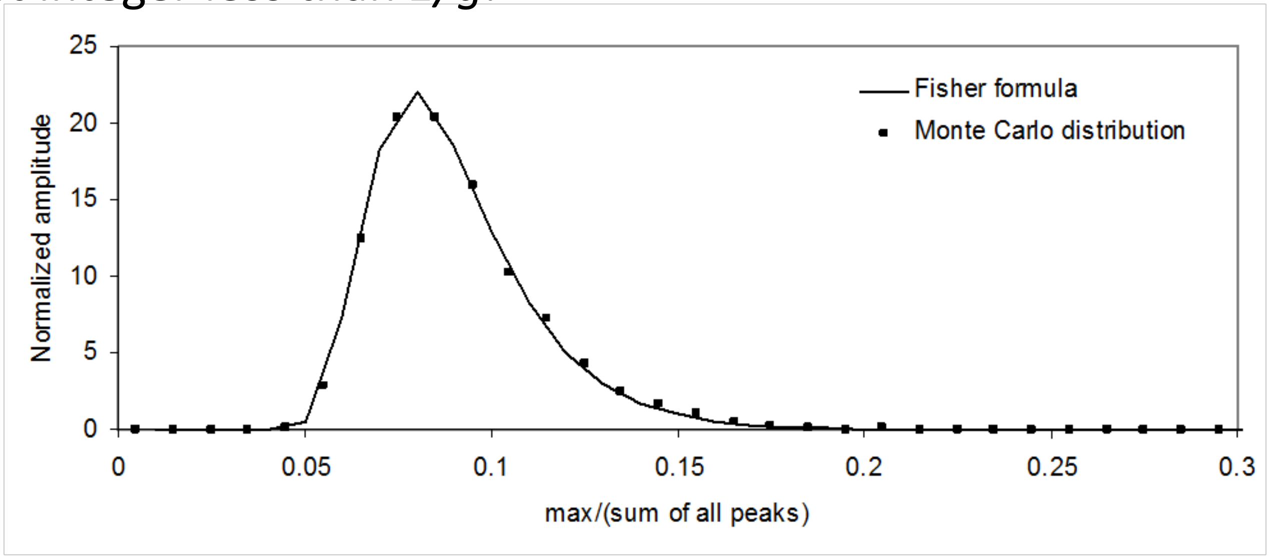
But a classical exact solution exists for the largest peak in case of unknown σ

Fisher's exact solution in case of unknown σ for the PDF of the highest noise peak in a Schuster periodogram

Test statistics *g=max/(sum of all peaks)*

$$p(g) = \sum_k \binom{M}{k} (-1)^{k+1} k(M-1)[1-kg]^{(M-2)}$$ PDF

where the sum over *k* extends up to the minimum between *M* and the highest integer less than *1/g*.



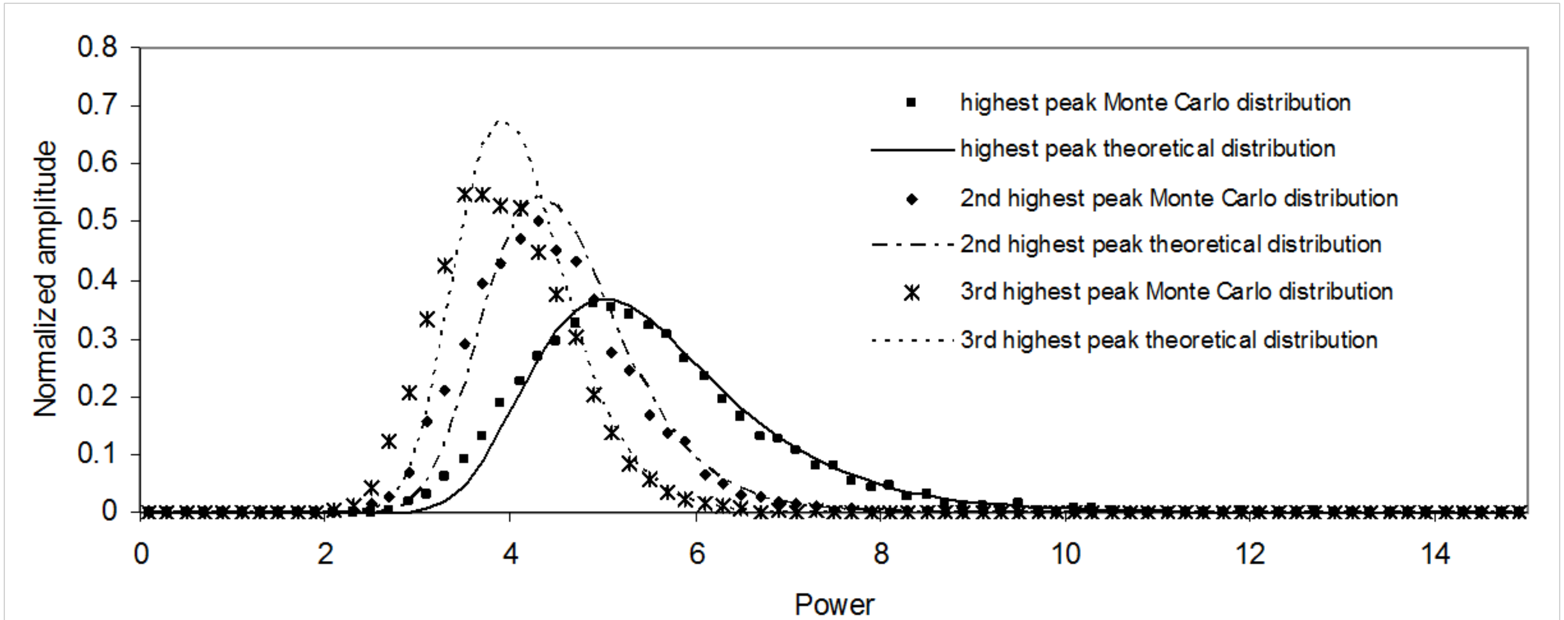
Exact answer for the significance of a signal search while scanning the M independent frequencies associated with an evenly sampled series

Some additional MC insights - simulated series with uneven sampling

In this case the number of effectively independent frequencies contained in the spectrum is not easily computable

Lomb-Scargle periodogram → two interest facts emerge from the MC test

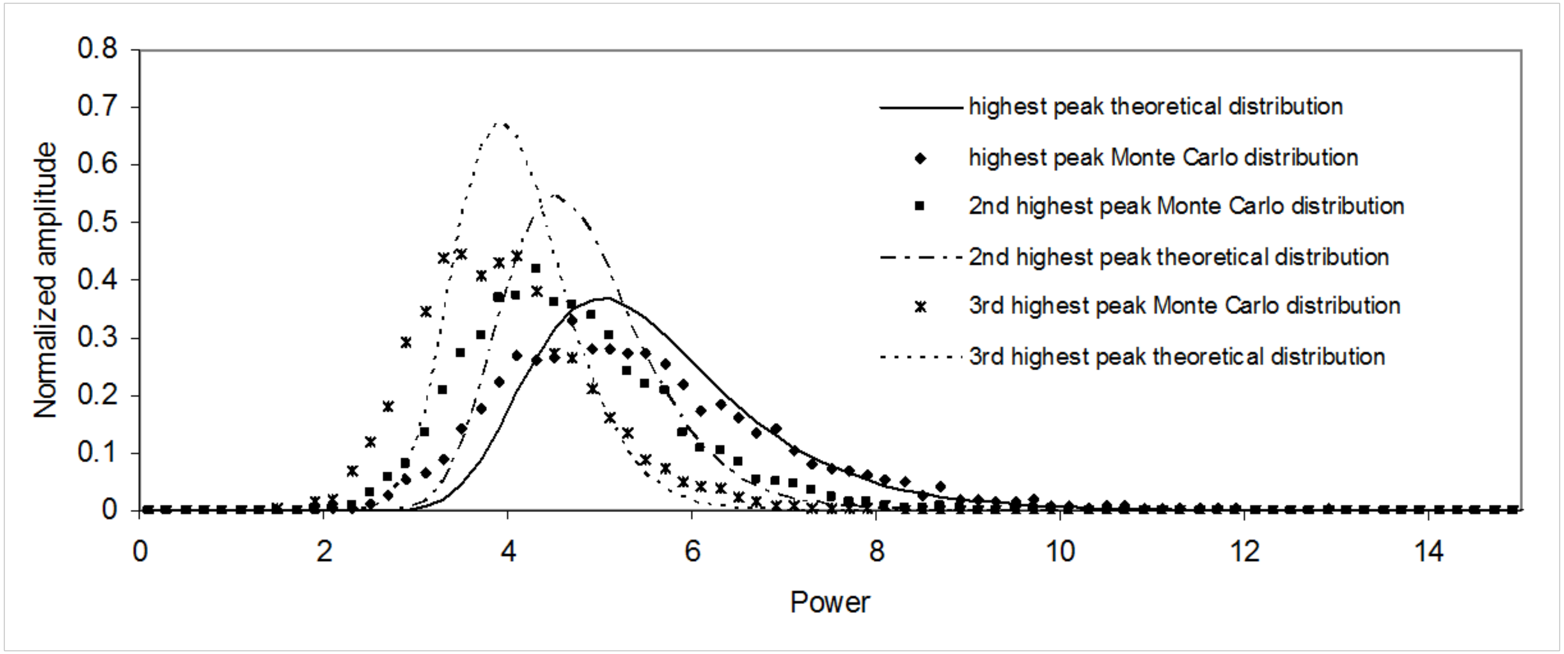
a) Unknown variance estimated from the data – realistic case



Qualitatively the model follows the MC distribution – M inferred by a fit of the largest peak MC curve to the model

Basis of the notion of "effective number of independently scanned frequencies" commonly used in the framework of the Lomb-Scargle analysis

b) case of known variance



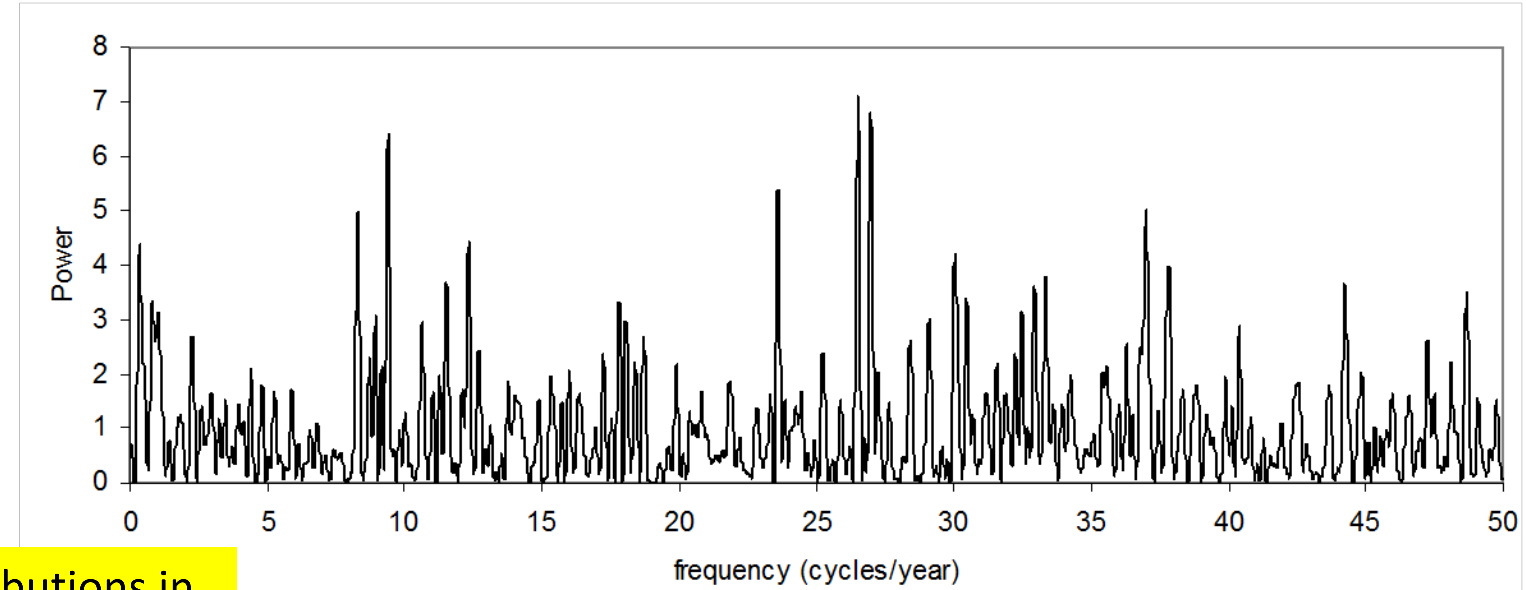
The model – MC agreement is lost

Opposite situation with respect to the Schuster periodogram

A virtue of the Lomb-Scargle approach - it behaves better in the realistic case

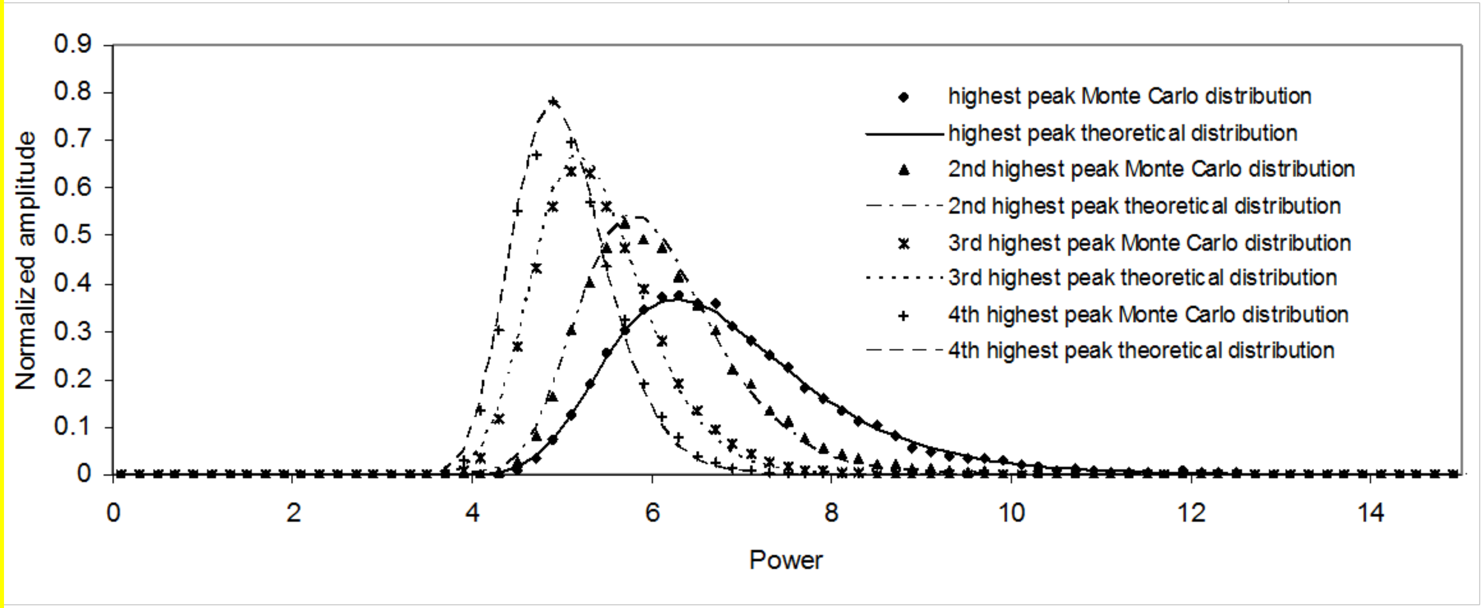
Some results of the MC based analysis for the real data

Again the Lomb-Scargle spectrum of the SK data



MC distributions in excellent agreement with the model

Highest peak fit to the model with M=529 -> number of effectively independent scanned frequencies



Numerical significances (p-values)

The usual significance assessment limited to the highest peak is extended to the first four peaks (frequency in cycles/year)

Rank	Frequency	Ordinate	Significance
1st	26.51	7.1	34.7%
2nd	26.99	6.8	15%
3rd	9.4	6.41	8%
4th	23.6	5.36	27%

Significance values directly inferred from the MC distributions – "area" of the MC distributions above the corresponding spectrum ordinate

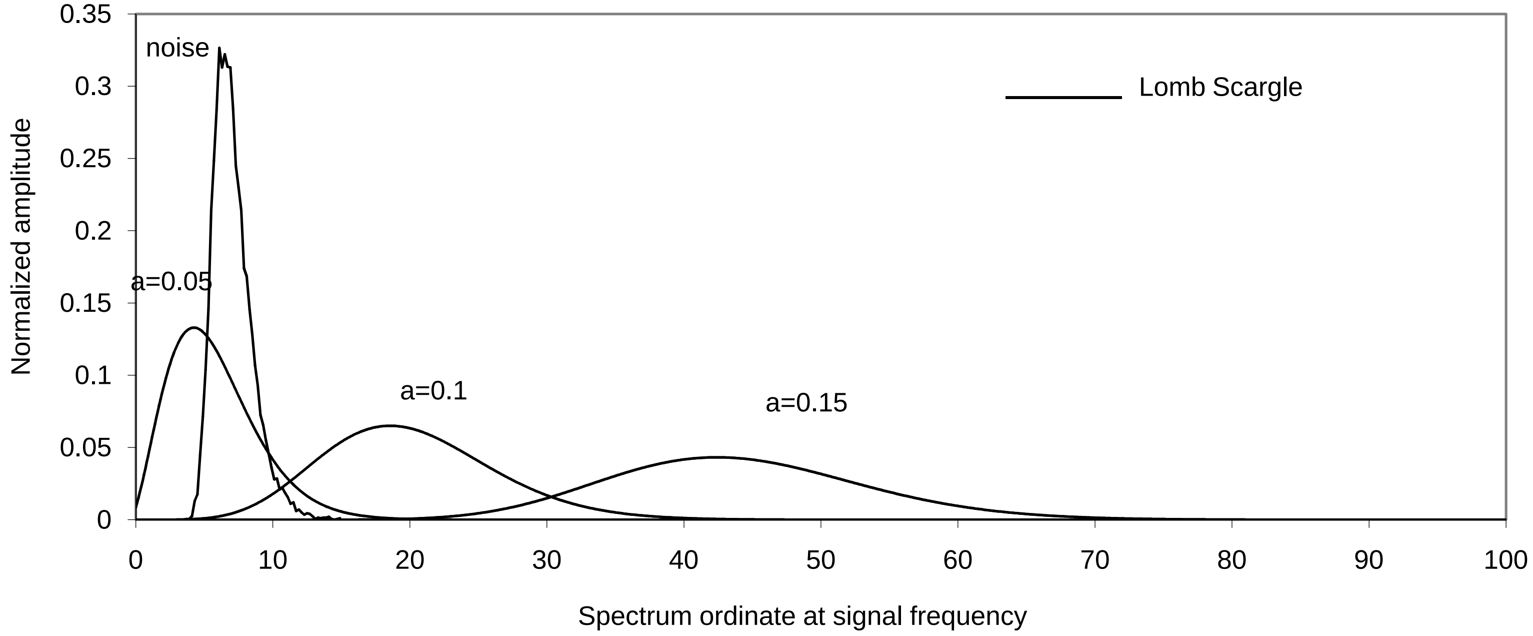
No extremely low significance value ⟶ the Lomb Scargle periodogram of the 10 days binned data is perfectly consistent with a noisy series with no periodicity embedded

The effect of the LEE is striking ⟶ for the highest 7.1 ordinate the significance for a specific-frequency signal detection from the $\mathbf{e^{-z}}$ distribution would be 0.08251 %

But what about the annual modulation?

Annual modulation

Indication in the spectrum
But too low to be identified blindly
Because of the LEE effect the noise vs. signal configuration changes as follows



The true noise distribution **stemming from the LEE**, i.e. the PDF of the highest noise peak, replaces the exponential distribution in the previous similar plot and hinders severely the sensitivity to low amplitude modulations in a blind search

The Collaboration to prove its presence binned appropriately the data on the basis of the expected period and phase of the annual signal

How to account for the errors

- Lomb-Scargle ignores the errors even if available
- The variance in the denominator of the Lomb-Scargle formula is inferred from the scatter of the data themselves
- But often the series of data are available with the errors
- Generalization of the Lomb-Scargle spectrum through the application of the **log-likelihood ratio**
- Comparing the **no signal/signal** situations

$$GLR = \frac{e^{-\frac{1}{2}\left(\sum_{k=1}^{N}\frac{x_k^{\,2}}{\sigma_k^2}\right)}}{\max(A,\varphi)\, e^{-\frac{1}{2}\left(\sum_{k=1}^{N}\frac{\left(x_k - AF\sin(2\pi f t_k + \varphi)\right)^2}{\sigma_k^2}\right)}}$$

$$-2\ln(GLR) = \sum_{k=1}^{N}\frac{x_k^{\,2}}{\sigma_k^2} - \min(A,\varphi)\sum_{k=1}^{N}\frac{\left(x_k - AF\sin(2\pi f t_{w,k} + \varphi)\right)^2}{\sigma_k^2}$$

The Wilks' theorem ensures that this quantity, under the null hypothesis, is asymptotically distributed as chi^2(2 dof)
the **likelihood spectrum**, being simply this quantity multiplied by ½, is thus asymptotically distributed as e^{-z} like Lomb-Scargle

Characteristics of the generalized likelihood spectrum

- Reduces to Lomb-Scargle ignoring the individual errors
- Under the null hypothesis same single frequency statistics as Lomb-Scargle
- Able to deal with asymmetric errors
- Advantage of slightly increased sensitivity to a signal if present

- In the multi-frequency search no agreement with the model
- Also in this case significance to be computed by MC

- **In term of illustration of the LEE in the frequency domain no differences with the outcomes of Lomb-Scargle**

Parallelism with a "prototypal" approach to the search of a particle of unknown mass

Simple paradigmatic scenario

Toy model

- Experimental mass range from 0 to 100
- Background Poisson distributed mean value 500, uniform over the entire mass range

Simple exploration of the range for a signal through a set of observation windows covering the whole interval

Basic case a set of W non overlapped contiguous windows of equal width depending upon the resolution, independent noise Poisson distribution in each of them with mean value (W/100)*500

The identification of a signal is denoted by a "significant" excess of events above background in any of the windows

But how to compute the significance?

Parallelism with the frequency search problem with M independent frequency

⟶ **M corresponds to W**

Probability function of the highest noise count over the whole set of W windows

In each windows $$P(n,B) = \frac{e^{-B} B^n}{n!}$$ B=(W/100)*500

Goal derive from this original distribution the distribution of the largest detected count N over the whole set W of observation windows

A given value N is obtained when in all the windows but one the counts are less than N, while in the residual window the count is exactly N

$$W\left(\sum_{n=0}^{N-1} \frac{e^{-B} B^n}{n!}\right)^{W-1} \frac{e^{-B} B^N}{N!} \longrightarrow M\left(1-e^{-h}\right)^{M-1} e^{-h}$$

"frequency case"

The three factors map one to one

But in the discrete case there are other configurations to consider………

Other configurations for the largest count

The same maximum count value N is also obtained when in all the windows but two the counts are less than N, while in the two residual windows they are exactly N ⟶ the probability associated to such a configuration is

$$\binom{W}{2}\left(\sum_{n=0}^{N-1}\frac{e^{-B}B^{n}}{n!}\right)^{W-2}\left(\frac{e^{-B}B^{N}}{N!}\right)^{2}$$

$\binom{W}{2}$ number of combinations for choosing two windows out of W

Generalizing such a consideration, the k_{th} configuration producing the same maximum count N is associated with the probability

$$\binom{W}{k}\left(\sum_{n=0}^{N-1}\frac{e^{-B}B^{n}}{n!}\right)^{W-k}\left(\frac{e^{-B}B^{N}}{N!}\right)^{k}$$

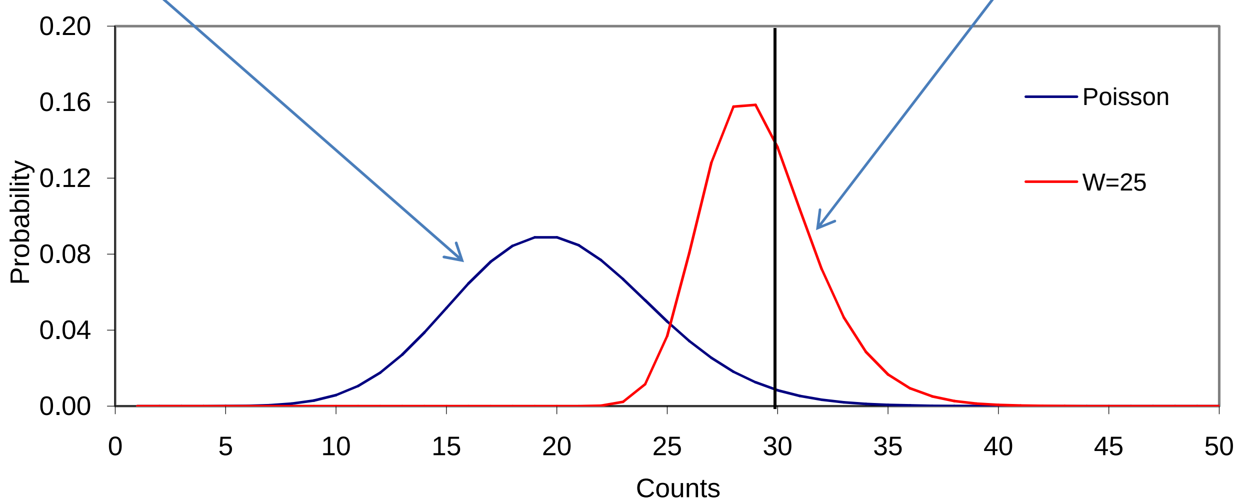
and hence the total probability for N is ➡ $P_{\max}(N)=\sum_{k=1}^{W}\binom{W}{k}\left(\sum_{n=0}^{N-1}\frac{e^{-B}B^{n}}{n!}\right)^{W-k}\left(\frac{e^{-B}B^{N}}{N!}\right)^{k}$

“Toy model” numerical example

Range 0-100, total background 500 counts uniformly distributed, 25 non overlapping windows of width 4

Single “fixed” window background distribution
Poisson

Multiple windows effective background distribution
PF of the largest count

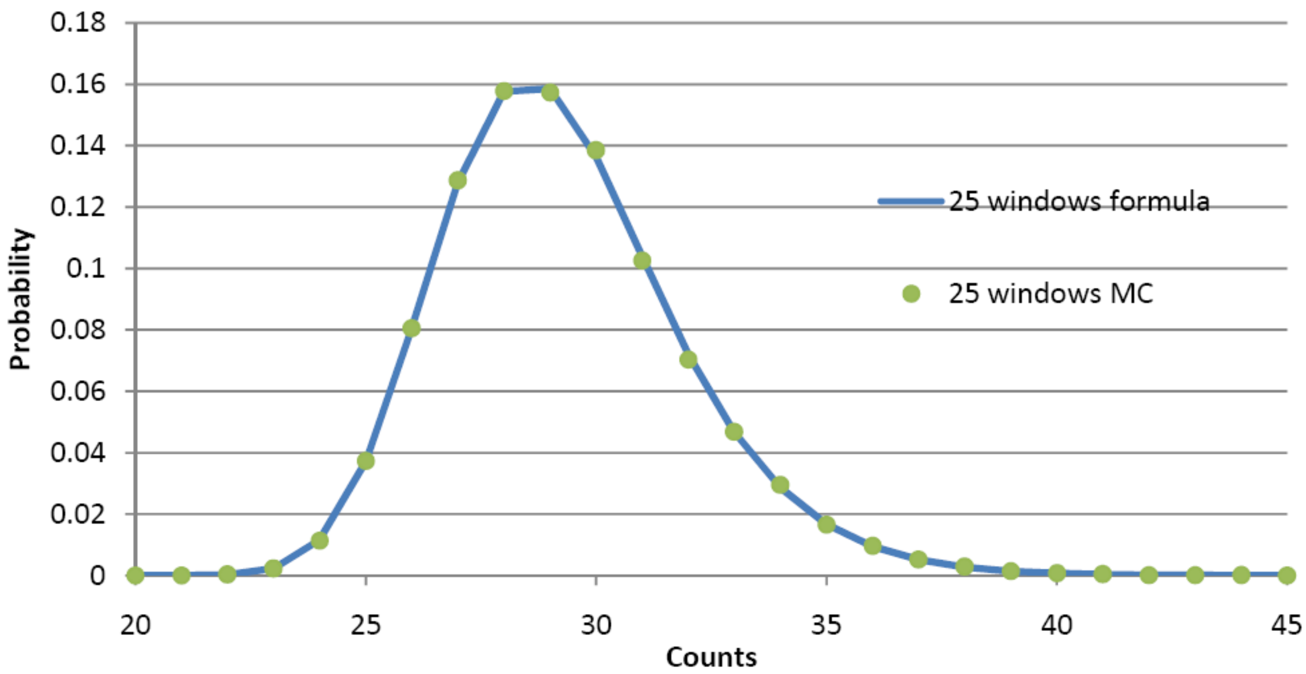


The **probability content** of the two noise distributions above an hypothetical threshold of 30 counts are enormously different -> the significance of the signal detection changes drastically as consequence of the multiplicity inherent in the search strategy

LEE in action

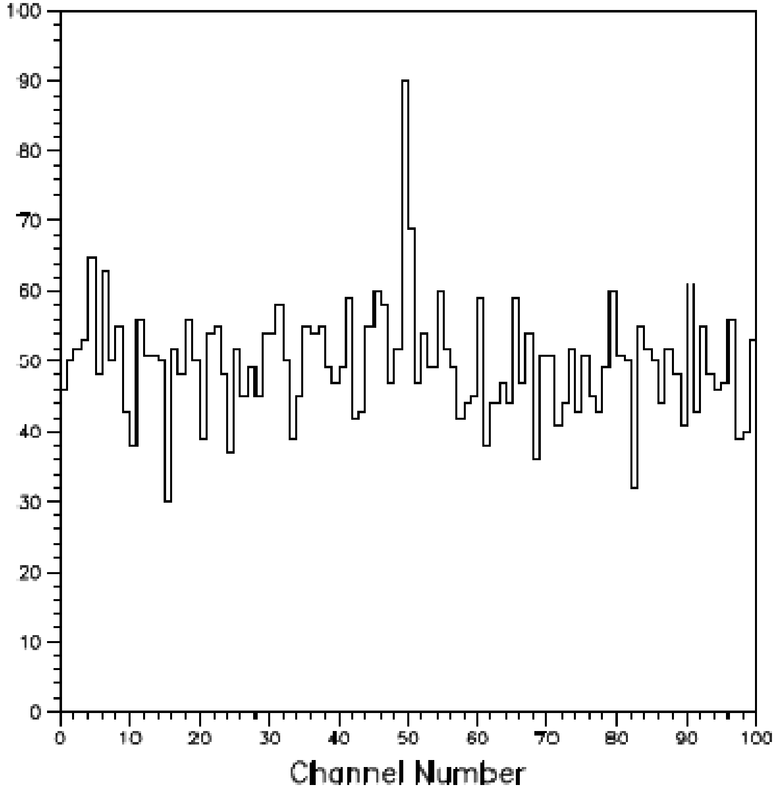
MC comparison

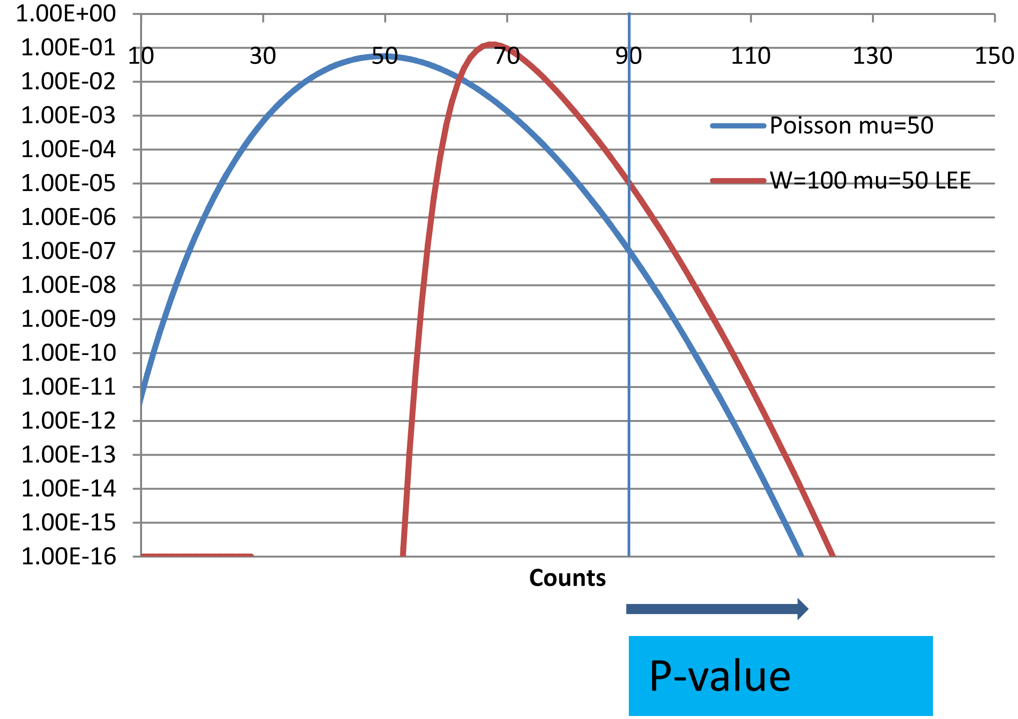
As in the frequency case the problem can be fully addressed via MC



The simulation of the toy model coincides perfectly with the formula of the probability function of the largest detected count over the whole set of windows

Valid also to describe the spikes in a finely binned interval with constant background identifying the windows with the bins – example in the talk of Luc Demortier





W=100

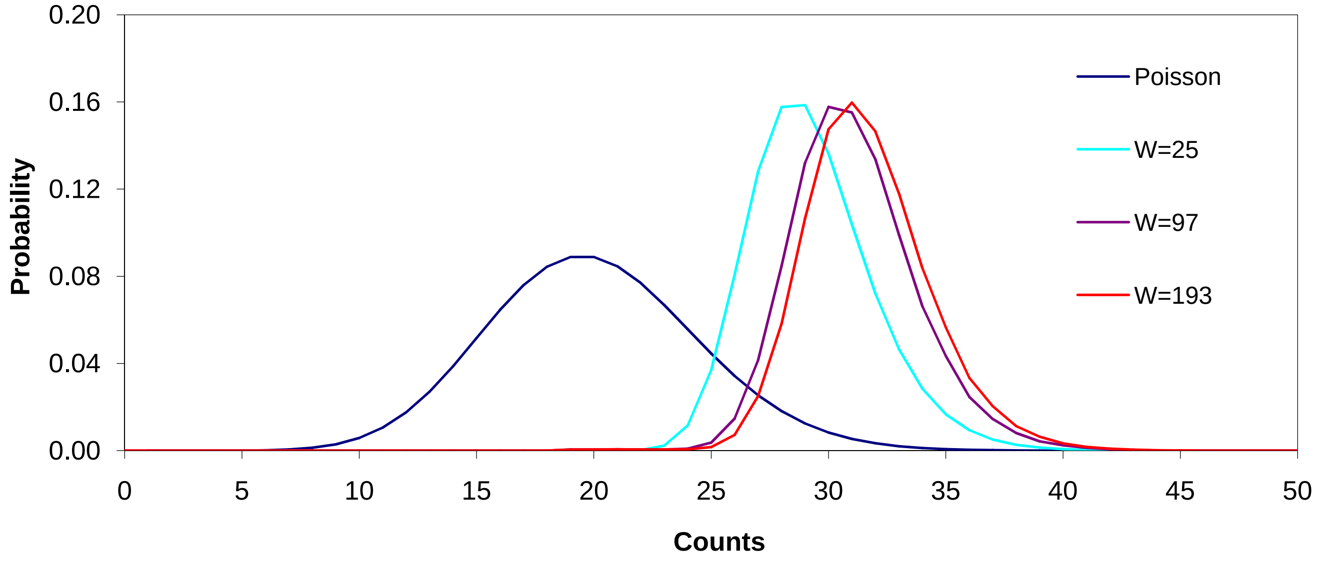
Mu=50

Poisson p-vale above 90 ⟶ $2.29\ 10^{-7}$

LEE p-vale above 90 ⟶ $2.29\ 10^{-5}$

Other parallelism with the previous frequency analysis: number of effectively scanned windows

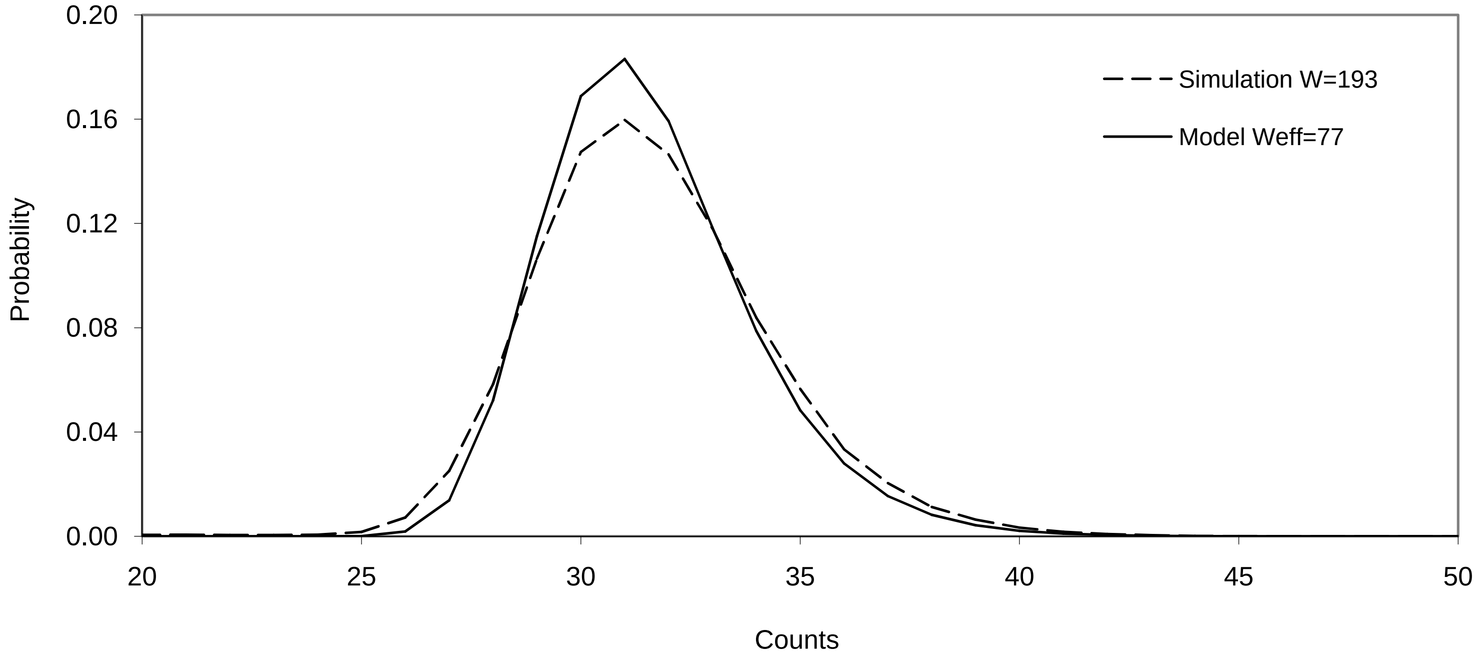
Overlapping non independent windows
In the toy model 97 (step size 1) and 193 (step size 0.5) - width 4 as for the 25 non overlapping windows – evaluation via MC



"Saturation" effect on the distribution of the highest noise peak, as if the "effective" number of scanned independent windows would converge to a definite number

Comparison of the formula for non overlapping windows with the MC output

Qualitative match for Weff=77



The notion of "effective number of scanned windows" is qualitative similar to the same concept in the Lomb-Scarge frequency analysis, though quantitatively does not lead to a perfect overlap with the model

Solution for a sliding window

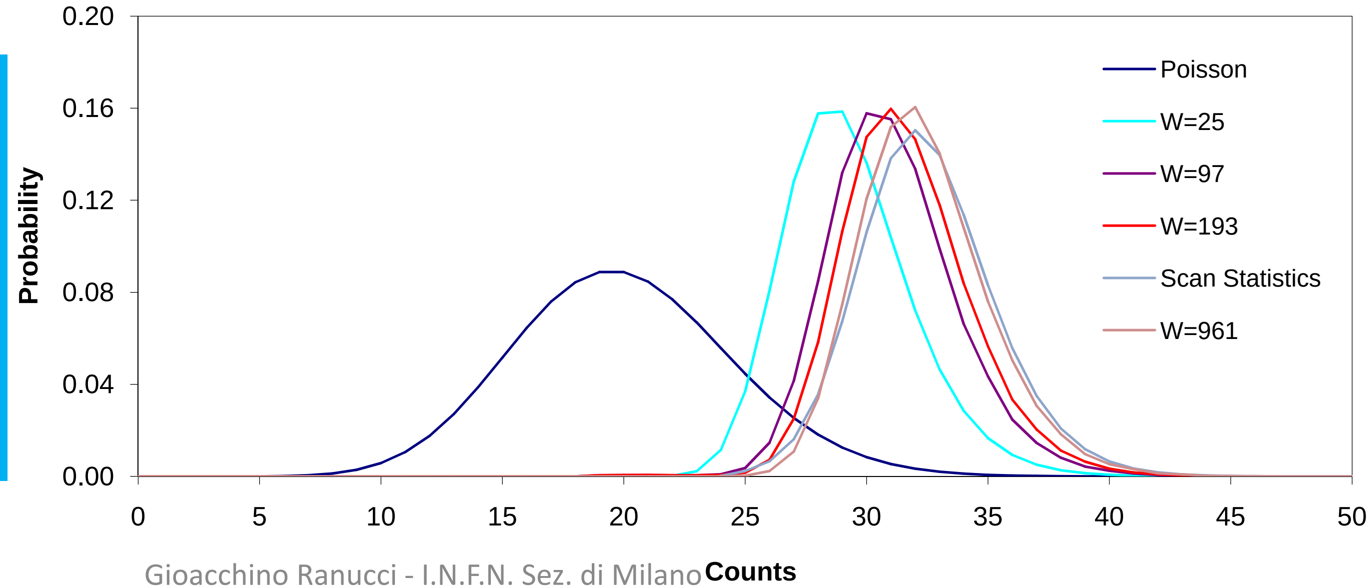
When the step size becomes very small the window scans smoothly the desired mass range

Moving/sliding window configuration

In this asymptotic case a solution does exist: "scan statistics"

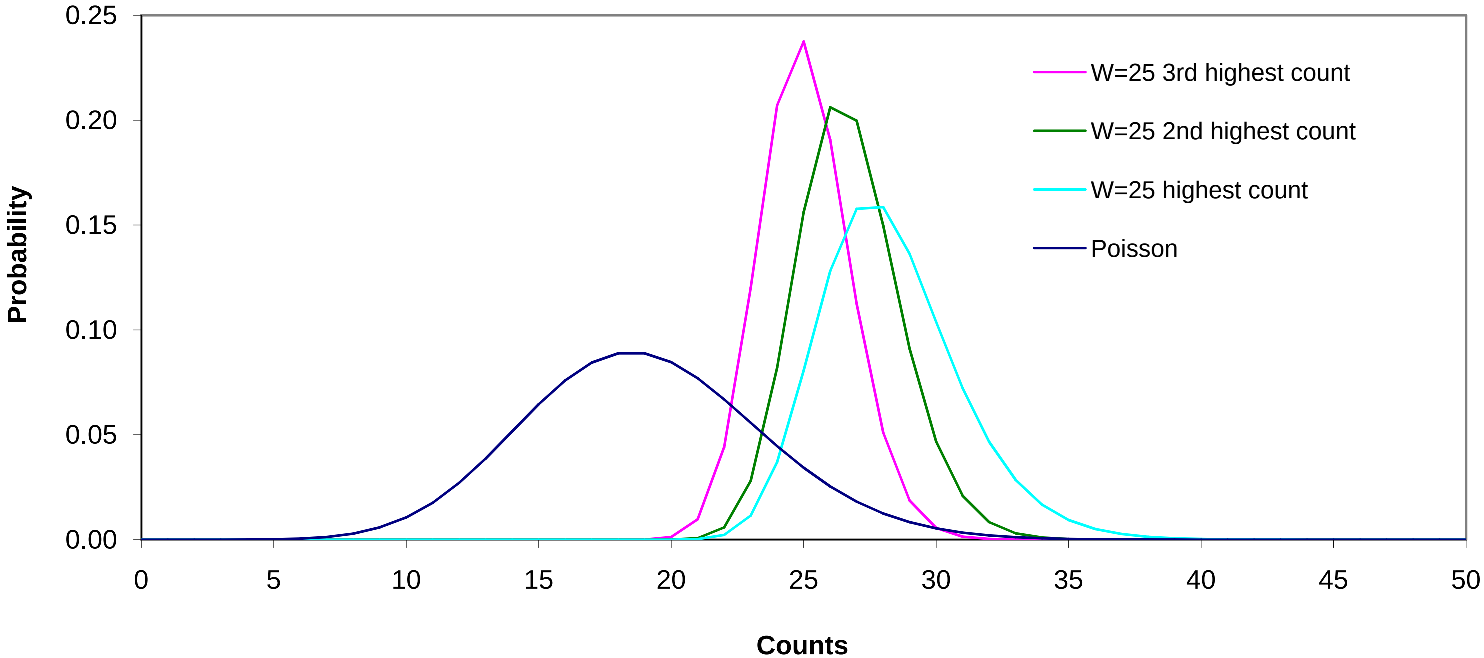
Comparison of the Alm approximation of the scan statistics with the toy model for very small step-size 0.1 → 961 windows (width 4)

The "dimensionality" of the search does not grow indefinitely
For W=961 (essentially coincident with the asymptotic scan statistics) **Weff=110**



Not only the highest count ....

With some combinatorial "gymnastic" it is possible to derive also the probability function of the 2nd , 3rd and so on highest counts , in the case of independent non overlapping windows



$$P_{\max-1}(N) = \sum_{k=1}^{W} \binom{W}{k} \left(\frac{e^{-B} B^N}{N!} \right)^k \sum_{m=1}^{W-k} \binom{W-k}{m} \left(\sum_{n=0}^{N-1} \frac{e^{-B} B^n}{n!} \right)^{W-k-m} \sum_{i=N+1}^{\infty} \left(\frac{e^{-B} B^i}{i!} \right)^m$$

2nd highest count

and this result completes the parallelism with the frequency analysis

Epilogue: what modulation in the solar neutrino data

✓Super-Kamiokande Lomb-Scargle analysis fully compatible with a pure constant series with no modulation embedded

✓**Annual modulation** picked-up taking into account the known periodicity and phase (published by the Collaboration – important systematic check)

✓A more sensitive analysis based on finer binning (5 days) and on the more sensitive likelihood spectrum gave a (very vague) hint of a modulation of 9.43 cycles/year (but in any case significance of only 2%) subjected to some debate

✓SNO (Sudbury Neutrino Observatory) data cancelled that hint – spectrum compatible with no modulation - **annual signal** identified through its known characteristics

✓With two datasets available what can be done to shed more light on the annual signal that must be present in the data ? ⟶ Combine them !

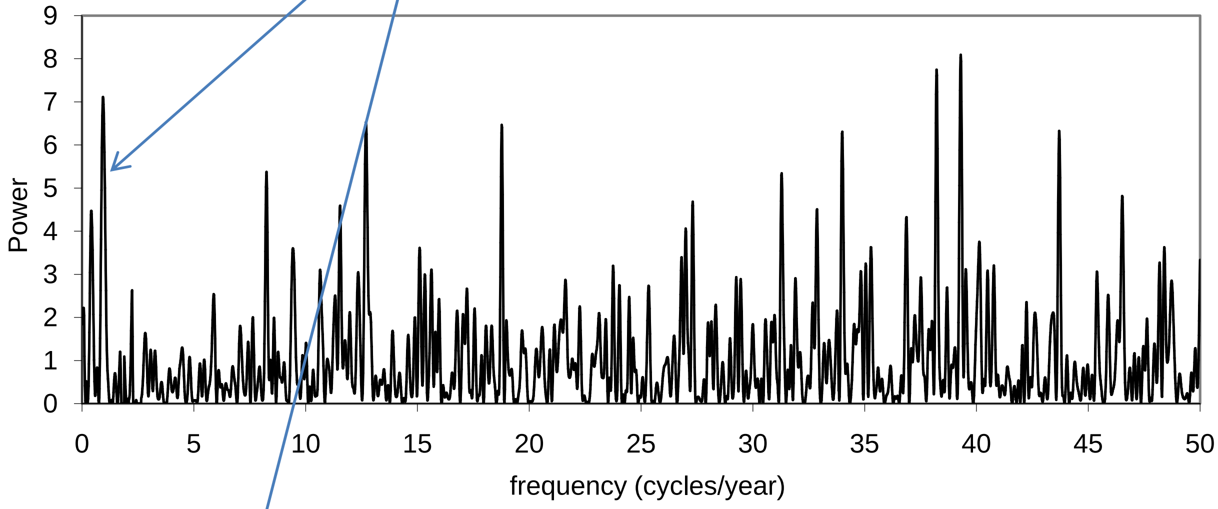
Cumulative SK and SNO likelihood spectrum

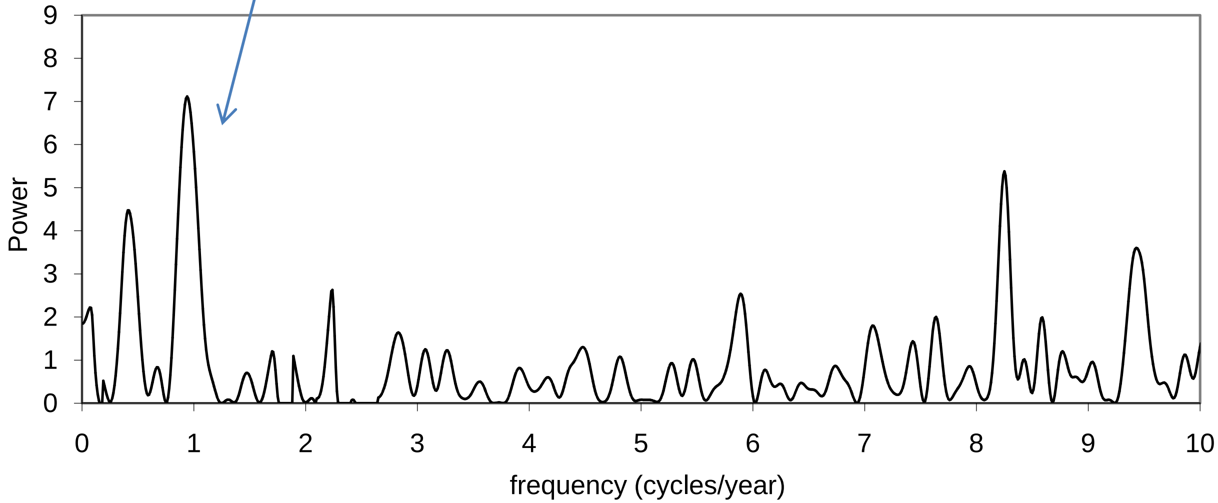
Neither in the Super-Kamiokande spectrum nor in the SNO spectrum such an expected line is so prominent

good systematic check of the methodology chosen to write down the cumulative spectrum: with more statistics the manifestation in the spectrum of the annual modulation signal is indeed enhanced, as it should

Legitimate to ignore the LEE and calculate the significance for fixed location search 0.08251 % i.e. 3.34 σ - error on the period only 5%

Annual line 3rd highest peak in the cumulative spectrum





Conclusions

The correct incorporation of the **Look Elsewhere Effect** is vital while searching for a modulation hidden in time series, otherwise **“look long enough, find anything!”** (Numerical Recipes)

The **LEE** change completely the detection scenario passing from the single frequency to the multiple frequency strategy search

The sensitivity to low modulation amplitude is severely affected

Same situation as the search of a particle of unknown mass

A parallelism can be established between the frequency search and a “prototype” approach to the scan of a mass range, also **through similar analytical formalisms**

Finally what modulation in the solar neutrino data? So far, only the **annual modulation**

References

J.Yoo et al., "Search for periodic modulations of the solar neutrino flux in Super-Kamiokande-I", Physical Review D, vol. 68, Issue 9, November 2003, id. 092002

D.O. Caldwell and P.A. Sturrock., "Evidence for solar neutrino flux variability and its implications", Astroparticle physics, vol. 23, (2005), 543-556

P. A. Sturrock et al., "Power-spectrum analyses of Super-Kamiokande solar neutrino data: variability and its implications for solar physics and neutrino physics ", hep-ph/0501205

N.R. Lomb, "Least-squares frequency analysis of unequally spaced data", Astrophysics and Space Science, vol. 39, Feb. 1976, p. 447-462

J.D. Scargle, "Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data", Astrophysical Journal, Part 1, vol. 263, Dec. 15, 1982, p. 835-853

E. J.,Groth, " Probability distributions related to power spectra", Astrophysical Journal Supp. Serise 286, 29, (1975), 285-302

F.J.M. Barning, "The numerical analysis of the light-curve of 12 Lacertae ", Bull. Astron. Inst. Neth., Volume 17, no. 1, pp 22-28 (1963)

J.D. Scargle, "Studies in astronomical time series analysis. III – Fourier Transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data ", Astrophysical Journal, vol. 343, Aug. 15, 1989, p. 874-887

A. Schuster, “On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena”, Terr. Mag. Atoms. Elect., 3,13-41, 1898

S.S. Wilks, "The large-sample distribution of the likelihood ration for testing composite hypotheses ", Ann. Math. Stat., 9, (1938) 60-62

W.T. Eadie et al. "Statistical methods in experimental physiscs", North-Holland Editions (1971)

W.H. Press et al., "Numerical recipes in Fortran ", Cambridge University Press (1991), sect. 12.1 and 13.8

G. Ranucci, "Time statistics of the photoelectron emission process in scintillation counters", Nucl. Instr. and Meth. A335, 15 October 1993, p. 121-128

A. Papoulis, "Probability, random variables, and stochastic processes ", McGraw-Hill (1984), 2nd edition

Horne, J, H..and Baliunas, S., L., “A prescription for period analysis of unevenly sampled time series”, Astrophysical Journal, vol. 302, March 15, 1986, p. 757-763

R.A. Fisher, “Tests of significance in harmonic analysis”, Proceedings of the Royal Society of London, A, 125 (1929) 54-59

A. F. Siegel,“Testing for periodicity in a time series”, Journ. Am. Stat. Assoc., vol. 75, No 370 (1980), 345-348

W. X. Zhou and D. Sornette, “Statistical significance of periodicity and log-periodicity with heavy-tailed correlated noise”, cond-mat/0110445

G. Ranucci, "On the significance of signal search through the “sliding window” algorithm", Nucl. Instr. and Meth. A562, June 2006, p. 433-438

F. Terranova, “Peak finding through Scan Statistics”, Nucl. Instr. and Meth. A 519, March 2004, p. 659 -666