Nonlinear Methods for Detecting Unmodeled Transients

Julien Sylvestre
LIGO-MIT

LIGO-G010216-00-D

University of Wisconsin - Milwaukee
April 26, 2001

Classical Linear Detection Theory

- For a given data vector, consider the standard hypothesis:

$$H_0: \quad \underline{x} = \underline{n}, \underline{n} \sim N(0, 1)$$

$$H_1: \quad \underline{x} = \underline{n} + \underline{s}, \underline{s} \in W \subseteq R^N$$

- The likelihood ratio test is:

$$\frac{\max_{\underline{s} \in W} p_{\underline{x}|H}(\underline{x}|H_1)}{p_{\underline{x}|H}(\underline{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

Classical Linear Detection Theory

- Simple algebra gives

$$\langle \underline{x}, \underline{s} \rangle \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\|\underline{s}\|^2}{2} + \lambda$$

where

$$\underline{s} = \arg \min_{\underline{\varphi} \in W} \|\underline{x} - \underline{\varphi}\|^2$$

Classical Linear Detection Theory

- Example 1: Filter Bank

$$\text{For } W \text{ such that } \forall \underline{s} \in W, \|\underline{s}\|^2 = 1,$$

$$\arg\min_{\underline{\varphi} \in W} \|\underline{x} - \underline{\varphi}\|^2 = \arg\max_{\underline{\varphi} \in W} \langle \underline{x}, \underline{\varphi} \rangle$$

$$\max_{\underline{\varphi} \in W} \langle \underline{x}, \underline{\varphi} \rangle \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} + \lambda$$

Classical Linear Detection Theory

- Example 2: Bandlimited Signal

$$\underline{s} = \underline{x}_{//} \Rightarrow \left\|\underline{x}_{//}\right\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\lambda$$

where $\underline{x}_{//}$ is the bandpass filtered data.

Classical Linear Detection Theory

- Example 3: Vector Subspace

$$\text{For a basis } \{\underline{e}_1, \ldots, \underline{e}_M\}, \underline{s} = \underline{x}_{//} = \sum_{i=1}^{M} \langle \underline{x}, \underline{e}_i \rangle \underline{e}_i \Rightarrow \left\| \underline{x}_{//} \right\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\lambda$$

$$\left\| \underline{x}_{//} \right\|^2 \sim \chi^2_M \Rightarrow \text{SNR} \cong \frac{P}{\sqrt{M}}$$

For the union of a set of subspaces, just need to consider the largest norm square over all the subspaces.

- In some sense, “efficiency scales as the inverse square root of generality” (sic)

Nonlinear Signal Estimation

- Consider again the model

$$\underline{x} = \underline{n} + \underline{s}, \underline{s} \in W \subseteq R^N$$

but from an estimation point of view.

- Consider transforming the data to some basis where the signal is concentrated on a small number of basis functions

$$y[i] = \langle \underline{x}, \underline{\Psi}_i \rangle$$

- The signal is efficiently recovered by thresholding

$$\hat{s}_\Psi[i] = \omega_i y[i]$$

Nonlinear Signal Estimation

- The goal is to minimize the mean-square error

$$\mathrm{MSE} = E[\|\underline{s} - \underline{\hat{s}}\|^2]$$

- With an Oracle that knows which basis functions are relevant,

$$\mathrm{MSE}[i] = \begin{cases} 1 \ \text{ if } \ \omega_i = 0 \\ \hat{s}_\psi[i] \ \text{ if } \ \omega_i = 1 \end{cases}$$

$$\mathrm{MSE} = \sum_{i=1}^{M} min(1, \hat{s}_\psi[i])$$

Nonlinear Signal Estimation

- Without the luxury of an Oracle, applying the threshold

$$\hat{s}_{\psi}[i] = 1_{|y[i]|>\eta} y[i], \quad \text{with} \quad \eta = \sqrt{2\log N}$$

gives a MSE

$$\text{MSE} \leq 2\log N \cdot \text{MSE}_{\text{Oracle}}$$

More importantly:

no essentially better inequality can hold universally for all signals in R^N
(Donoho & Johnstone, 1993)

The Missing Link

- The best signal in W for correlation is

$$\hat{\underline{s}} = \arg\min_{\underline{\varphi} \in W} \left\| \underline{x} - \underline{\varphi} \right\|^2$$

- For the nonlinear approach, the best signal is taken to be

$$\hat{\underline{s}} = \arg\min_{\underline{\varphi} \in R^N} E\left[\left\| \underline{s} - \underline{\varphi} \right\|^2 \right]$$

- The nonlinear approach doesn't explicitely refer to W, so it is very general
- Optimality statements are hard to make (what is the metric for an unmodeled burst?), but looks optimal at least asymptotically

The Missing Link

1. Impose a restriction on the signal character by picking an orthonormal basis.

2. Transform the data to that basis. Apply a threshold on the transformed data.

3. Inverse transform back to the time domain. This gives the signal estimation.

4. Correlate the signal estimation with the data. Threshold on the correlation to decide between H0 and H1.

Choosing the Basis

- The crux of the problem is to chose the orthonormal basis: the best MSE

$$\mathrm{MSE} = \sum_{i=1}^{M} min(1, \hat{s}_{\psi}[i])$$

is obtained in the basis where the number of coefficients above 1 is the smallest.

- One thing to do is to minimize the effect of a suboptimal basis by using a lower threshold on $|y[i]|$ but doing clustering analysis
- Another thing to do is to try to pick the best basis for the data by minimizing the number of large coefficients

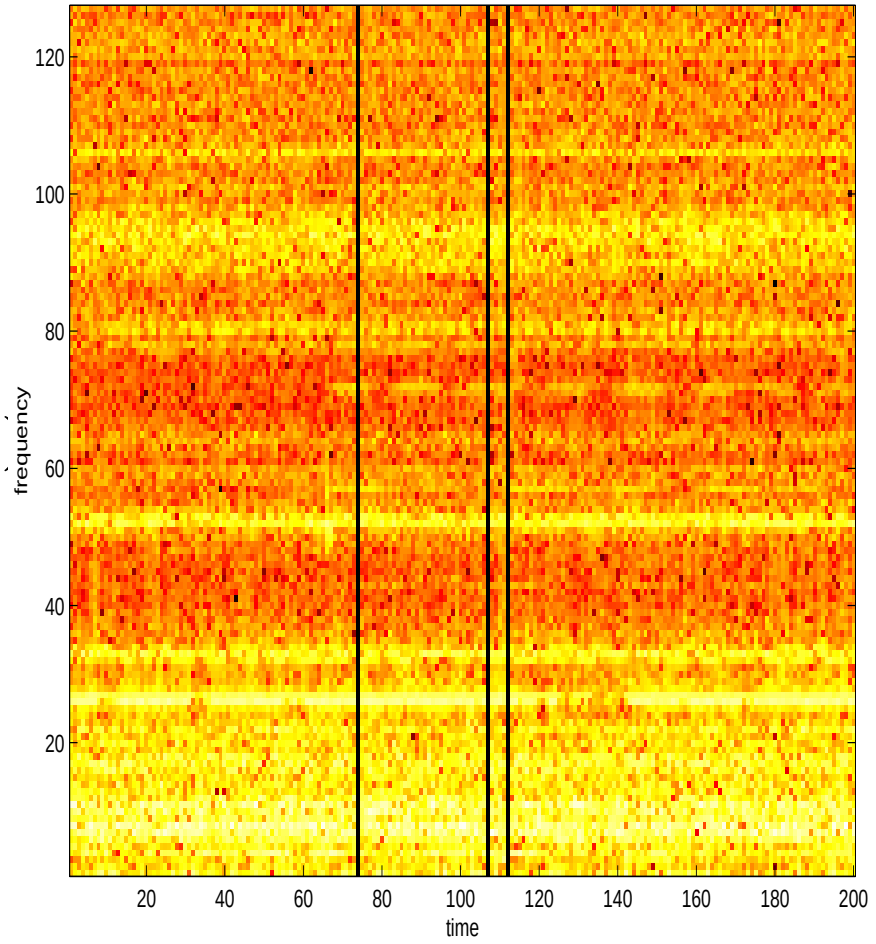
Clustering Analysis

- Spectrogram forms an orthonormal basis of R^N
- For wide sense stationary, power is independent in frequency (modulo 2π)

$$E[P[i]P[j]] \propto P^2[i]\mathrm{sinc}^2 T(i-j)$$

- For colored noise, some dependence in time, small if

$$\frac{1}{\min f_{\mathrm{line}}} \ll T$$



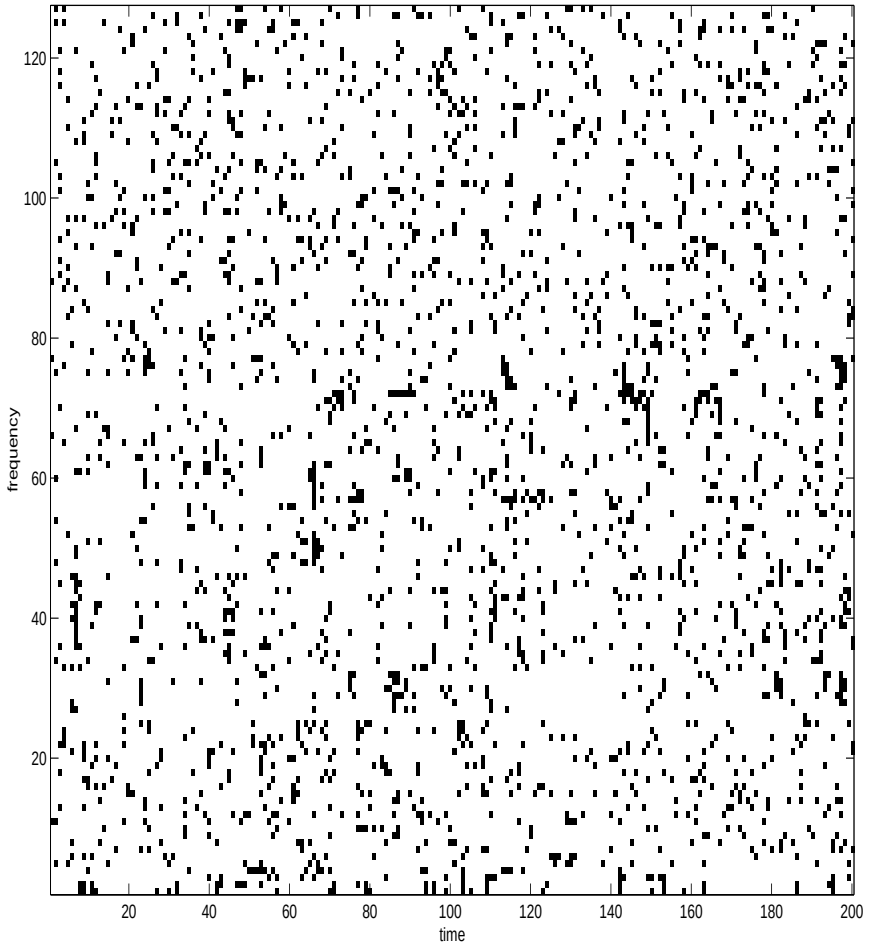
LIGO

Clustering Analysis

- For Gaussian noise, power follows the Rice distribution

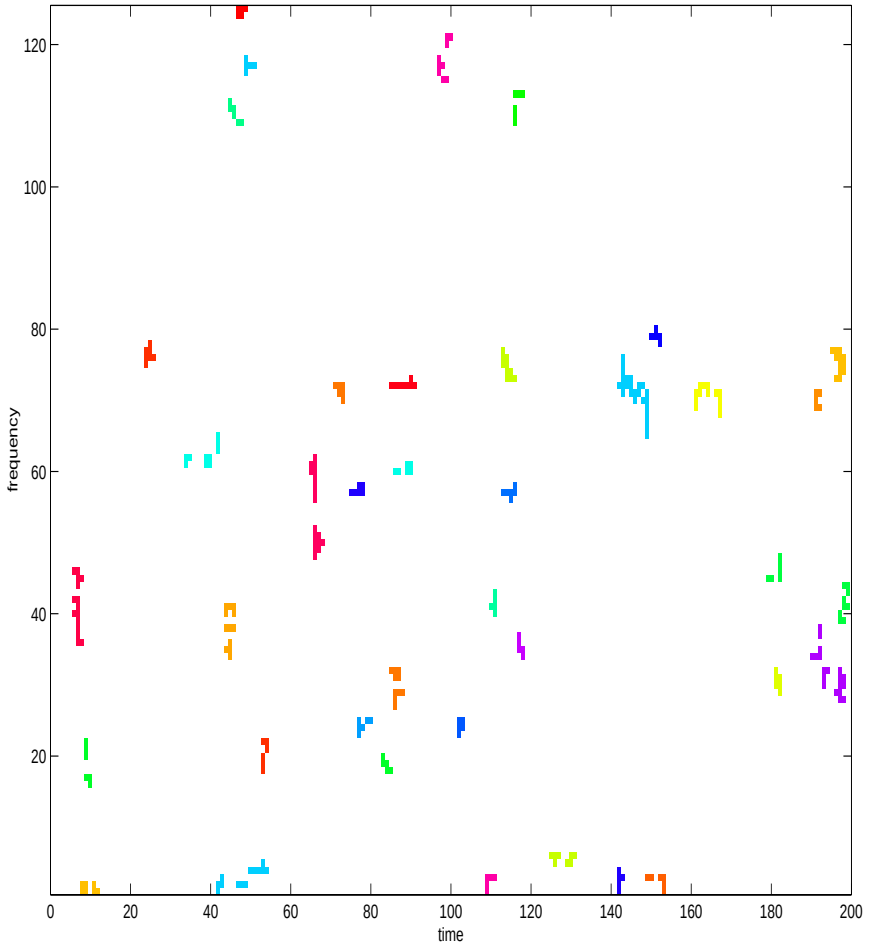
$$p(P_{ij}) = \frac{1}{S_i}\exp\left(-\frac{P_{ij}+Q_i}{S_i}\right)I_0\left(\frac{2\sqrt{P_iQ_i}}{S_i}\right)$$

- Setting a threshold on power or amplitude of real and imaginary parts gives the signal estimation
- Get a white noise picture (like TV 'snow')



Clustering Analysis

- 2D white noise doesn't like to form large clusters
- Physically, most signals are expected to form clusters
- Clean the image by thresh-olding on the cluster sizes
- Gives a list of significant events
- Power over each cluster is a χ^2, and is the statistic used for the last cut



Clustering Analysis

- Clusters on the square lattice are well known in Statistical Mechanics; they are called ‘lattice animals’
- Given a (uniform) black pixel probability p, the mean number of clusters of size s per pixel is

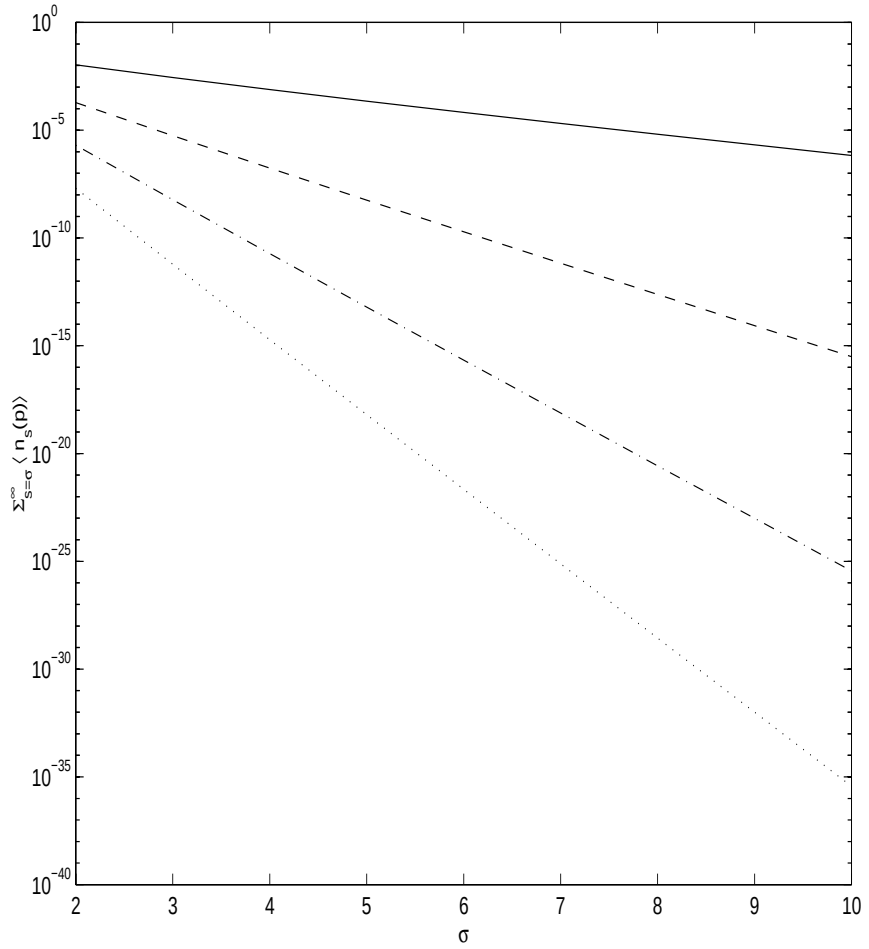
$$\langle n_s \rangle = p^s D_s(1-p)$$

- The perimeter polynomial basically counts the number of clusters of different shapes:

$$D_s(q) = \sum_t g_{st} q^t$$

Clustering Analysis

- Number of clusters of size larger than s scales almost exponentially with s.
- This is the quantitative version of '2D white noise doesn't like to form large clusters'



Clustering Analysis

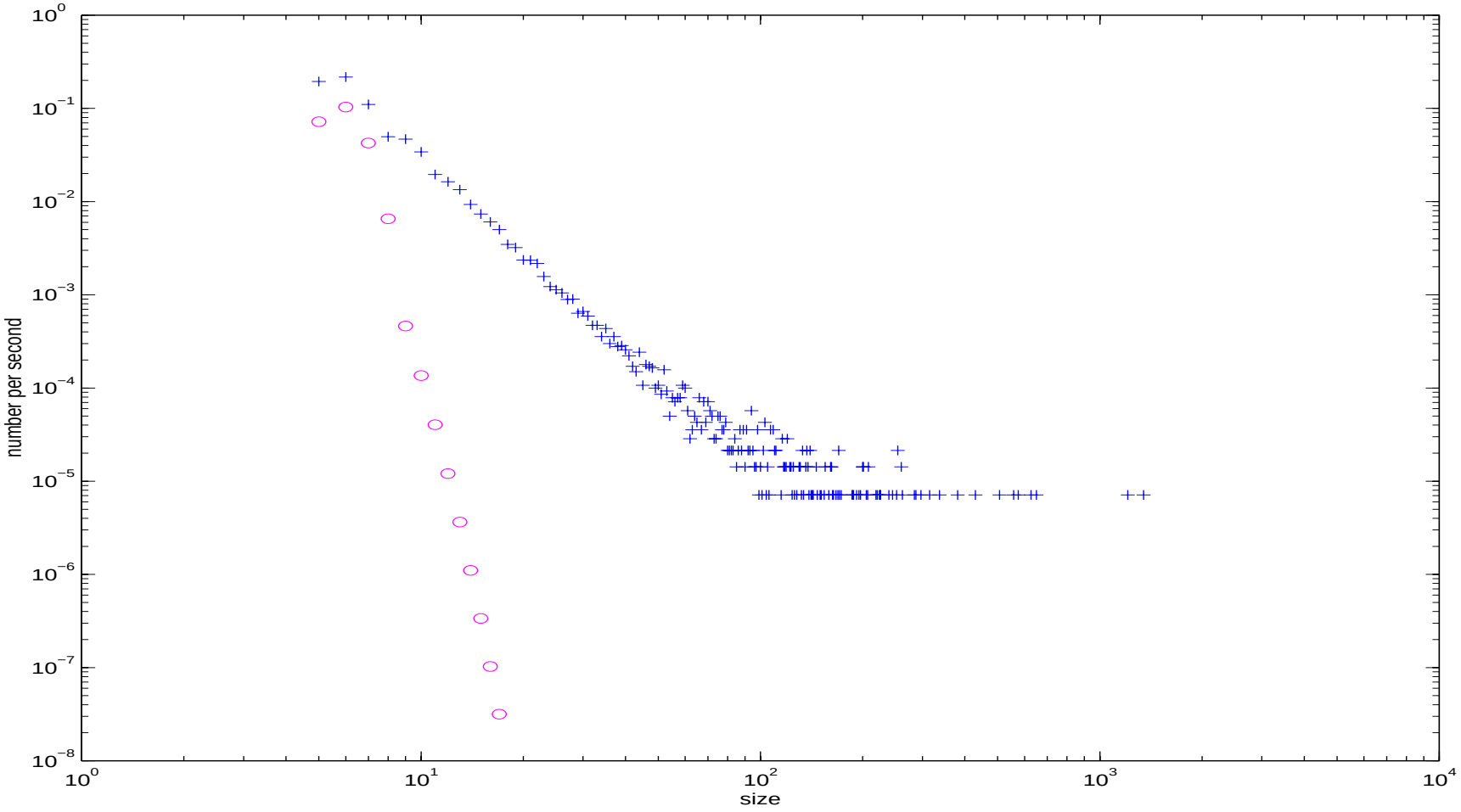
- To get the first order description of the cluster size population (i.e. the small p limit), only need to count the number of clusters of a certain size and a certain perimeter

  ›› e.g. there are 1 cluster of size 1, 2 of size 2, 19 of size 4 (Tetris), and 400795844 of size 17.

- Can also do higher order, although it's hard computationally. At this time, I know the 'two-point correlation function' for a pair of clusters up to size 4 at any distance from each other.

  ›› e.g. there are 40 ways to place two clusters of size 2 at a distance of 6, and 6004 ways to place two clusters of size 4 at a distance of 6.

Evaluating the non-gaussianity of the background noise



Best Basis Selection

- Consider a library of bases. Consider the entropy of the data in all bases. Pick the minimum of the entropy. In that basis, apply the usual threshold

$$\hat{s}_{\psi}[i] \;=\; 1_{|y[i]|>\eta} y[i], \quad \text{with} \;\; \eta \;=\; \lambda(1+\sqrt{N \log_2 N})$$

Then, with probability exceeding 1- e / $N \log_2 N$,

$$\mathrm{MSE} \le \frac{\lambda(1+\sqrt{N\log_2 N})}{1-8/\lambda} \mathrm{MSE}_{\text{Basis Oracle}}$$

Cf. the result for a fixed basis,

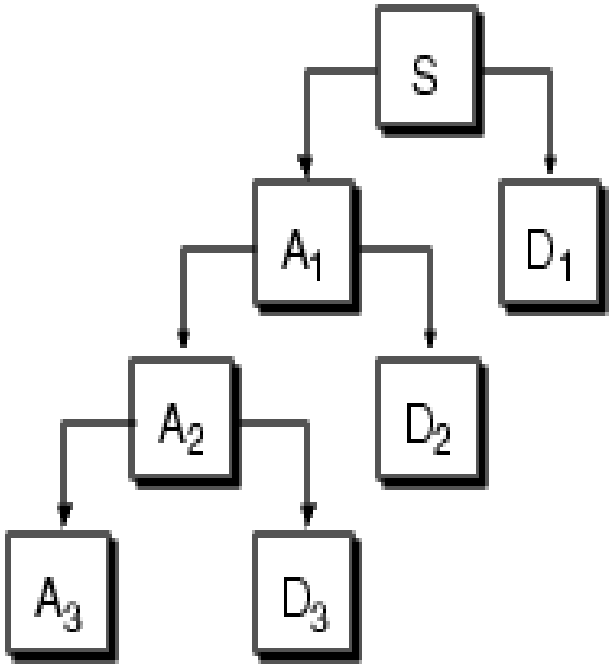
$$\mathrm{MSE} \le (2.4 + 2\log N)(1 + \mathrm{MSE}_{\text{Oracle}})$$

Best Basis Selection

- The best basis case gives a MSE at most a constant factor worse than the fixed basis MSE. However, the signal is compressed by a much larger factor.
- One interesting approach is to use wavelet packets

Wavelet Packets

- The standard wavelet transform applies iteratively a low and a high-pass filter to generate different levels of ‘details’ and ‘approximations’



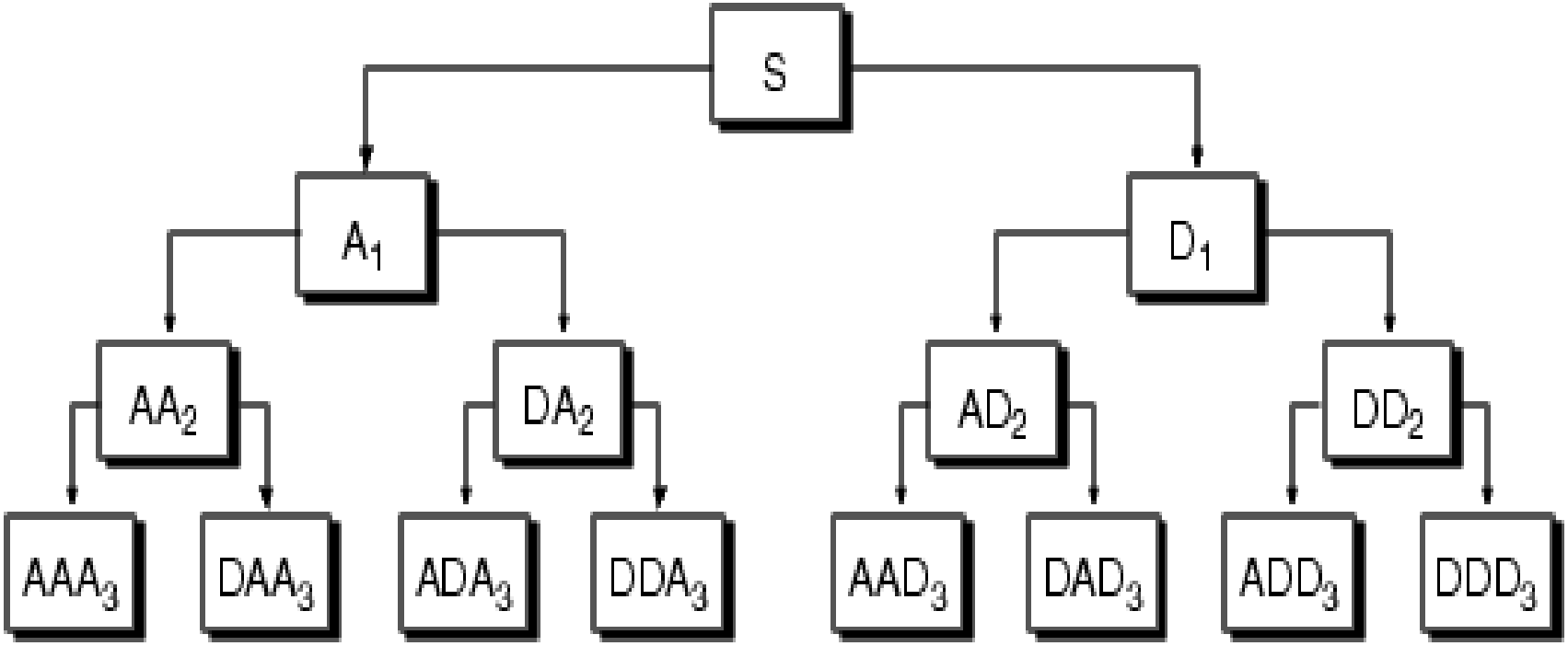
$$S = A_1 + D_1$$

$$= A_2 + D_2 + D_1$$

$$= A_3 + D_3 + D_2 + D_1$$

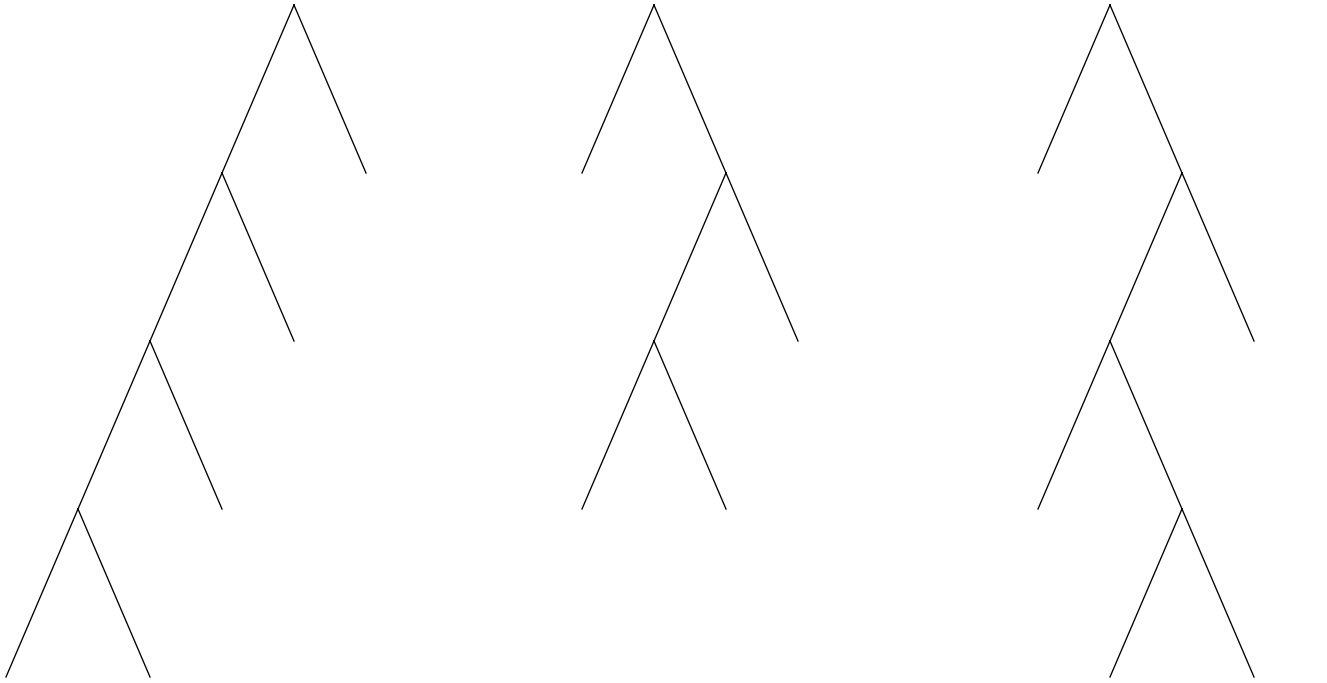
Wavelet Packets

- The wavelet basis is just one of many bases from the more general wavelet packet:



Wavelet Packets

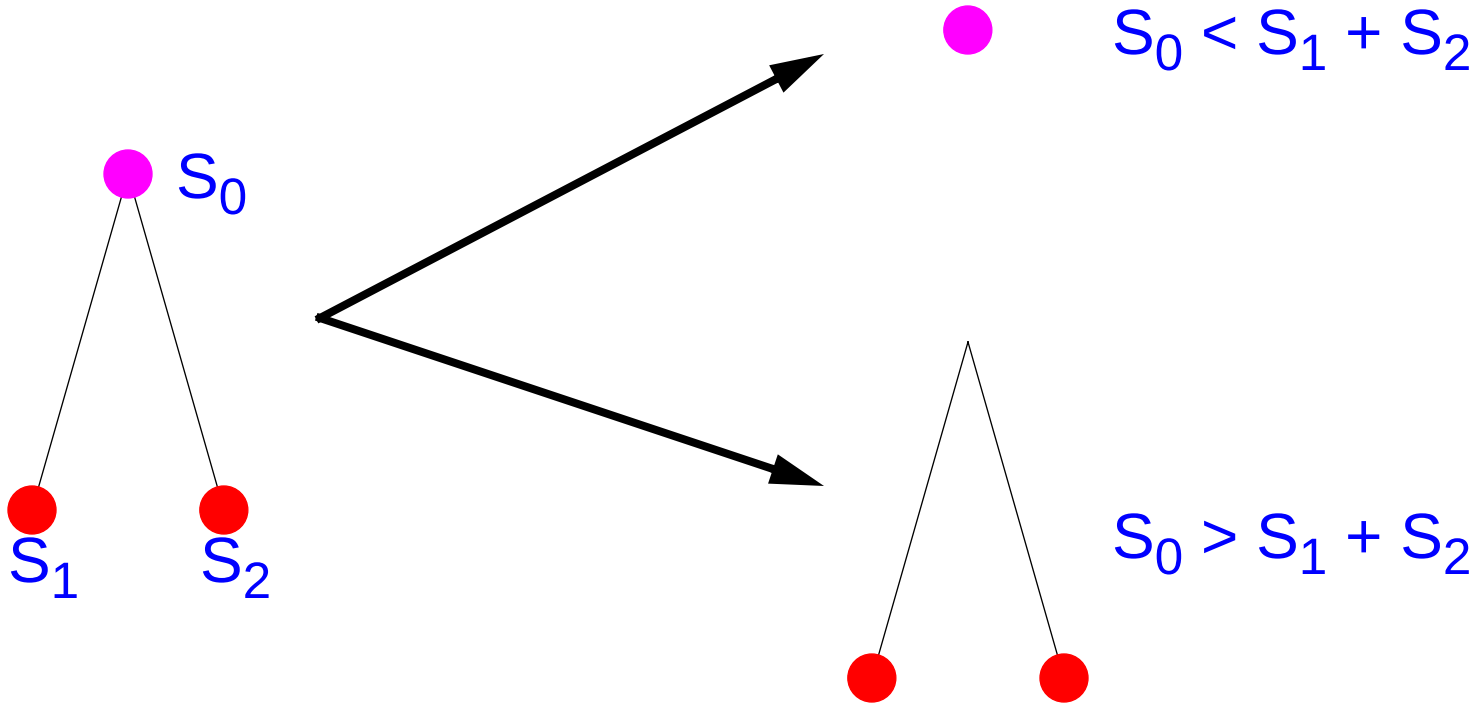
- Any subset of the binary tree is a valid basis



wavelet basis some other unnamed bases

Wavelet Packets

- For a given data segment, the best basis is picked by minimizing the entropy over the binary tree:



Wavelet Packets

1. Take a segment of data. Compute its wavelet packet decomposition.

2. Compute the entropy over each node of the wavelet packet. Go over the tree to minimize the entropy.

3. In the best basis, retain all the coefficients above some threshold.

4. Back in the time domain, this is the signal estimator. Correlate with data, apply threshold on χ^2.

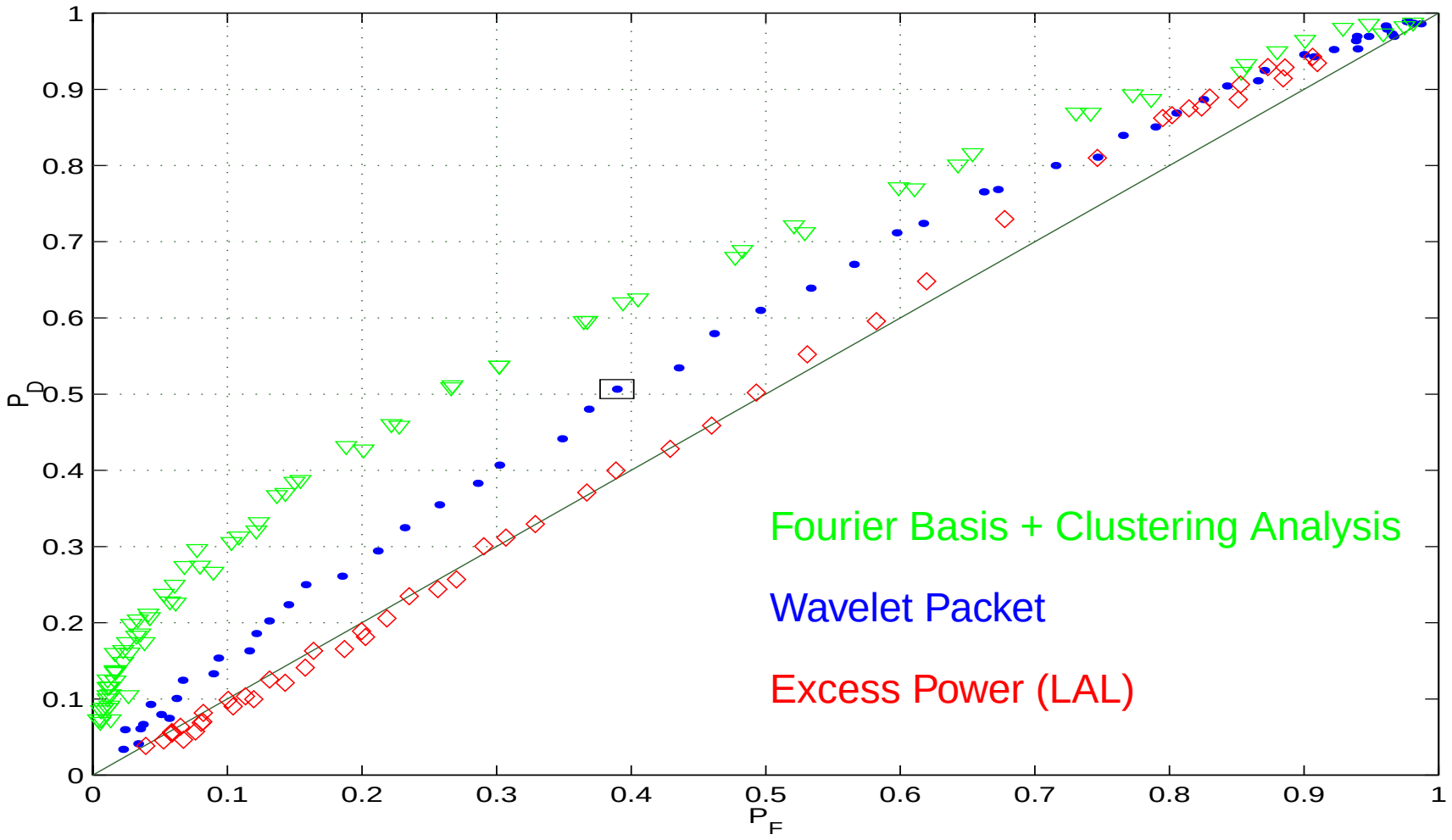
Simulations

- Working in the frequentist picture, the detectors are compared in the following way:
  1. A segment of length N of white gaussian noise is generated
  2. The detector is applied on the data segment
  3. The lower and upper limits on the confidence interval on the probability of false alarm (using a Bernouilli distribution) are computed at the 95% confidence level
  4. If (UL-LL) < 0.025, continue. Otherwise, back to 1.
  5. Repeat 1. to 4. with signals drawn from some population distribution superimposed to the noise.
- In all cases, the simulations are run a number of times, with the thresholds of the detector being varied.

I. Long signal

- Data are 16 s long at 2048 Hz
- Noise is gaussian white of unit variance
- Signal is 8 s long, from 4 s to 12 s. It is a white noise of unit variance that has been passed through a bandpass filter with corner frequencies at 100 Hz and 105 Hz.
- SNR is therefore -23 dB

I. Long signal



Isn't the excess power statistic optimal?

- Searching over a fixed shape time-frequency rectangle: excess power optimal

$$\arg\min_{\underline{\varphi} \in W} \left\| \underline{x} - \underline{\varphi} \right\|^2 = \arg\max_{t,\, f} \left\| \underline{x}_{//}(t,\, f) \right\|^2$$

$$\langle \underline{x},\, \underline{s} \rangle \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\left\| \underline{s} \right\|^2}{2} + \lambda \Leftrightarrow \max_{t,\, f} \left\| \underline{x}_{//}(t,\, f) \right\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\lambda$$

Isn't the excess power statistic optimal?

- Searching over a set of time-frequency rectangles: NOT optimal in Frequentist sense

$$\arg\min_{\underline{\varphi} \in W} \left\| \underline{x} - \underline{\varphi} \right\|^2 = \arg\max_{t,\, f,\, \Delta,\, B} \left\| \underline{x}_{//}^{\Delta,\, B}(t,\, f) \right\|^2$$

$$\arg\min_{\underline{\varphi} \in W} \left\| \underline{x} - \underline{\varphi} \right\|^2 = \arg\max_{t,\, f} \left\| \underline{x}_{//}^{\max \Delta,\, \max B}(t,\, f) \right\|^2$$

$$\langle \underline{x},\, \underline{s} \rangle \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\left\| \underline{s} \right\|^2}{2} + \lambda \Leftrightarrow \max_{t,\, f} \left\| \underline{x}_{//}^{\max \Delta,\, \max B}(t,\, f) \right\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\lambda$$

Isn't the excess power statistic optimal?

- Frequentist optimal decision rule (high confidence of no false events):

$$\frac{\max_{\underline{s} \in W} p_{\underline{x}|H}(\underline{x}|H_1)}{p_{\underline{x}|H}(\underline{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

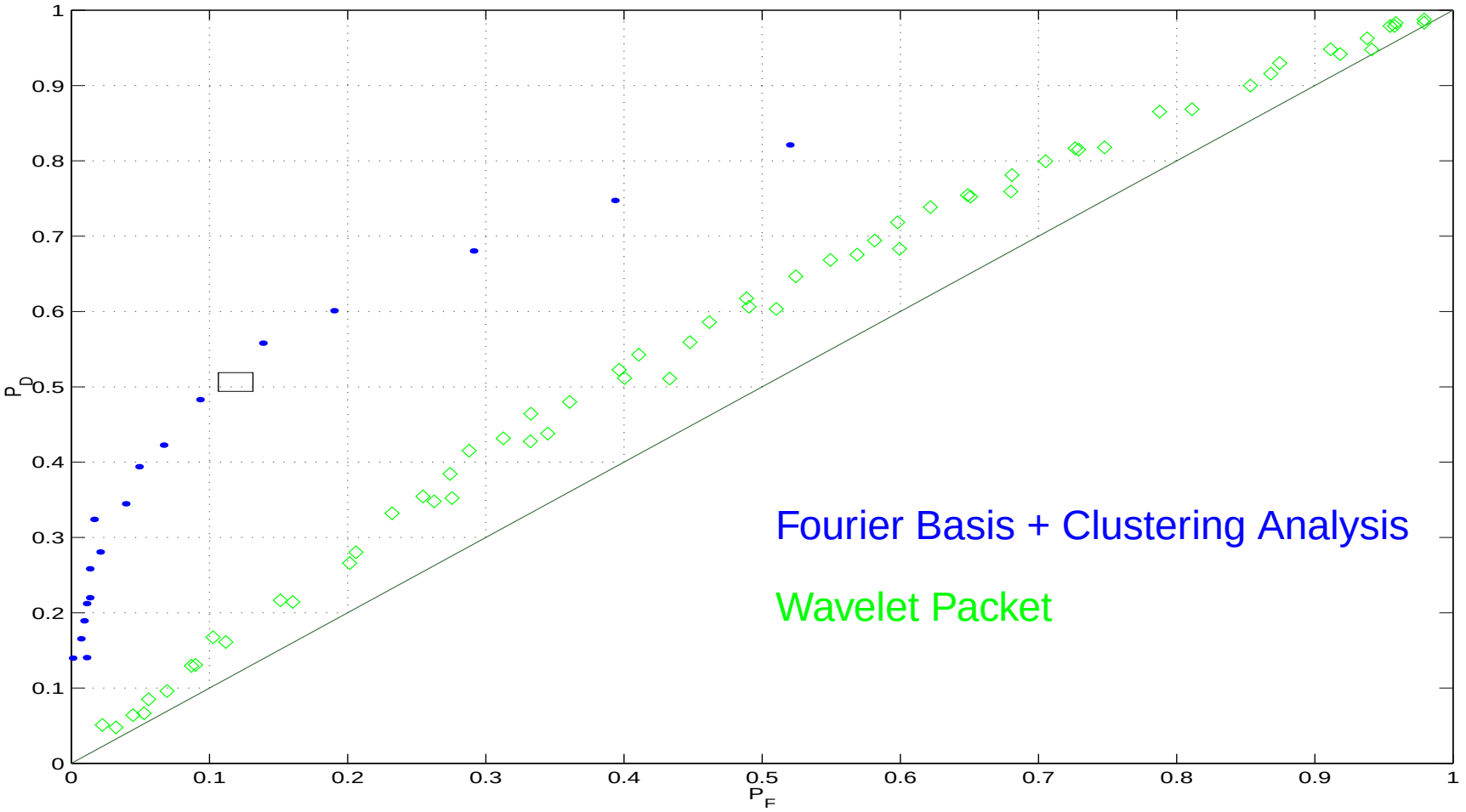
- Bayesian optimal decision rule (high confidence that events are signal):

$$\frac{\int_W p_{\underline{x}|H}(\underline{x}|H_1) p_{\underline{s}}(\underline{s}) d\underline{s}}{p_{\underline{x}|H}(\underline{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

II. Short signal

- Data are 1s long at 16384 Hz
- Noise is gaussian white of unit variance
- Signal is 0.7 s long, from 0.125 s to 0.825 s. It is an inspiral waveform (starting when f = 40 Hz) of two 100 M_{Sun} black holes, followed by a fake merger, followed by a ringdown (cf. Anderson & Balasubramanian, gr-qc/9905023)
- Energy repartition is 82% inspiral (chirp), 15% merger (broadband), 3% ringdown (narrowband)
- SNR is 19 dB

II. Short signal



Two Channels

- Consider the following hypothesis (Fourier domain):

$$\text{H0:}\quad \underline{x} = \underline{n}_1$$
$$\underline{y} = \underline{n}_2$$

$$\text{H1:}\quad \underline{x} = \underline{n}_1 + \underline{s}_1$$
$$\underline{y} = \underline{n}_2 + \underline{s}_2,\ \ \underline{s}_1, \underline{s}_2 \in S$$

$$\text{H2:}\quad \underline{x} = \underline{n}_1 + \underline{h};\underline{s}$$
$$\underline{y} = \underline{n}_2 + \underline{k};\underline{s},\ \ \underline{s} \in W$$

where by assumption $h_i > k_i$

Two Channels

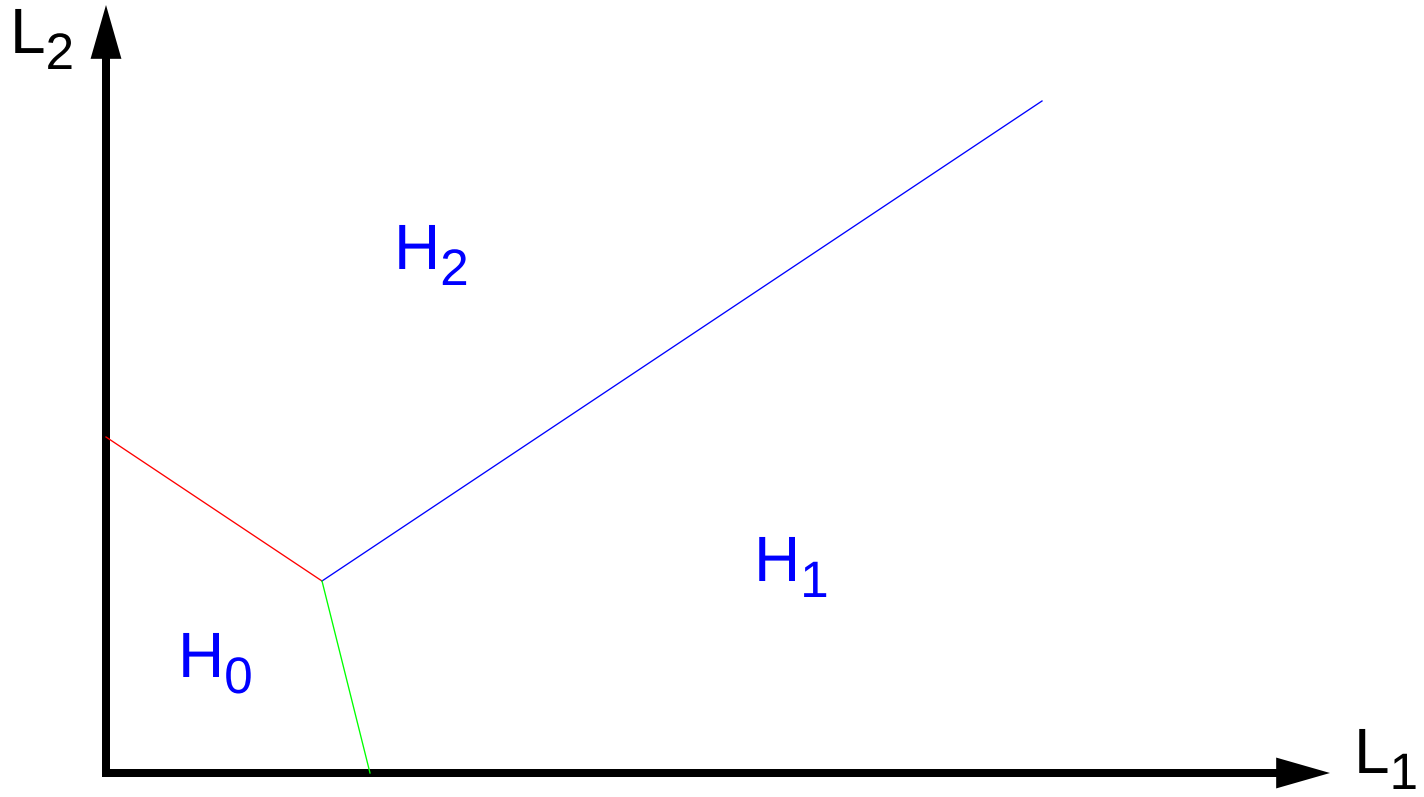
- The standard (optimal) way to run the test involves two likelihood ratios:

$$L_1 = \frac{\max_{\underline{s} \in W} p_{\underline{x},\underline{y}|H}(\underline{x},\underline{y}|H_1)}{p_{\underline{x},\underline{y}|H}(\underline{x},\underline{y}|H_0)}$$

$$L_2 = \frac{\max_{\underline{s} \in W} p_{\underline{x},\underline{y}|H}(\underline{x},\underline{y}|H_2)}{p_{\underline{x},\underline{y}|H}(\underline{x},\underline{y}|H_0)}$$

Two Channels

- The decision regions have 5 degrees of freedom



Two Channels

- Using the 'MSE' approach as before gives the following test:

  1.Construct the quantity

$$\sigma_i = \frac{|x_i|^2}{4|h_i|^2} + \frac{|y_i|^2}{4|k_i|^2}$$

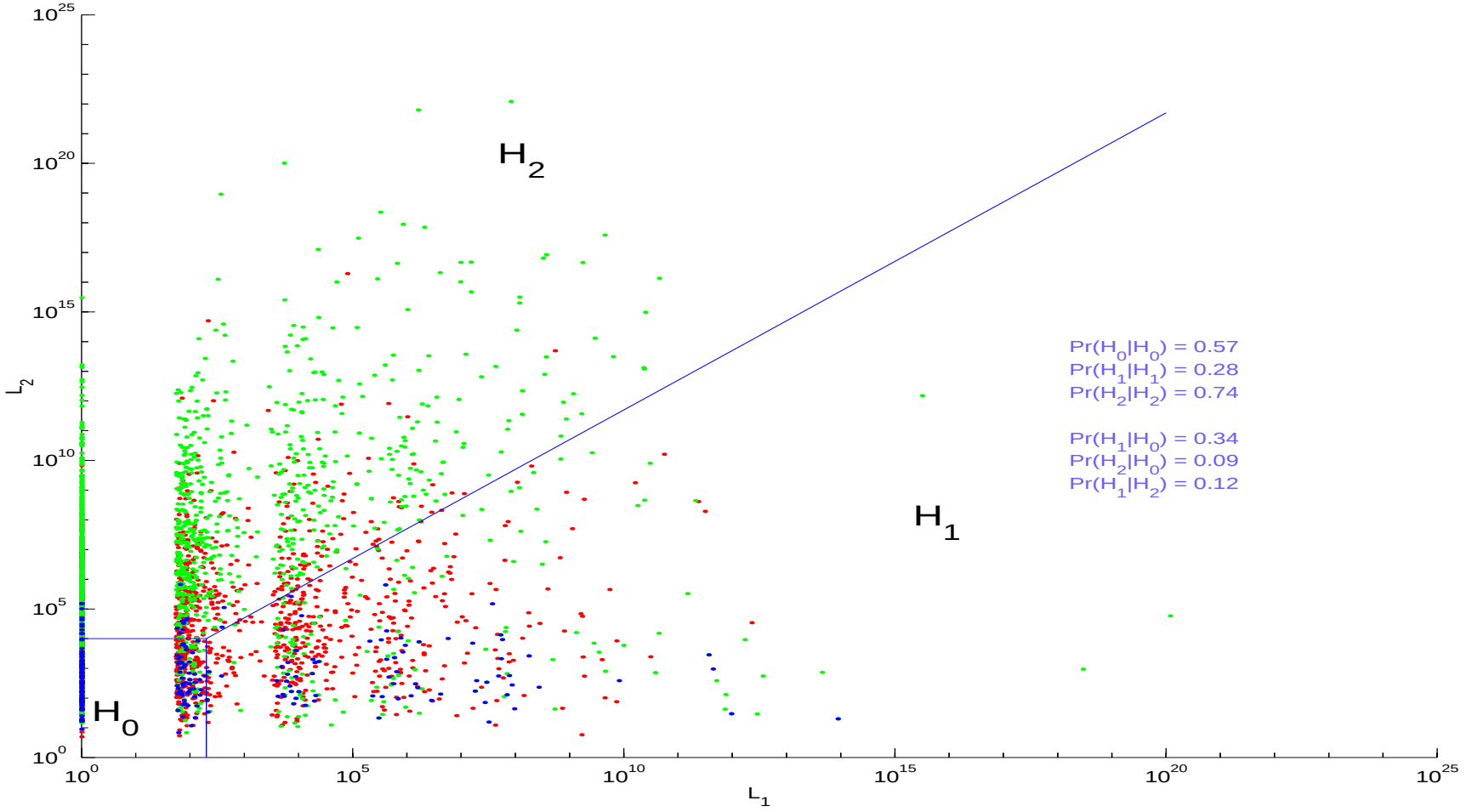
  2.Use the signal estimator

$$\hat{s}_i = \frac{x_i}{2h_i} 1_{\{\sigma_i > \lambda_1\}} + \frac{y_i}{2k_i} 1_{\{\sigma_i > \lambda_2\}} \quad , \quad \lambda_2 > \lambda_1$$

  3.Compute L_2 using the signal estimator. Compute L1 by taking the product of the 'nonlinear' likelihood ratios computed individually for the two channels as before

  4.Decide between the three hypothesis

Two Channels



Conclusion

- The main difficulty in studying the detection of unmodeled transients is to come out with an unbiased measure of the performance of the detectors
- Nonlinear methods offer great performances and generality
- For the very limited study presented here, the Fourier basis works fine
- With some optimization work, the wavelet basis will probably do even better
- The excess power statistic is optimal to build ‘Credible Intervals’ (Bayesian confidence intervals)
- Nonlinear methods are not necessarily optimal, but work better for ‘Classical confidence intervals’ (Frequentist)