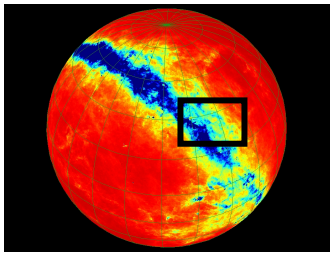
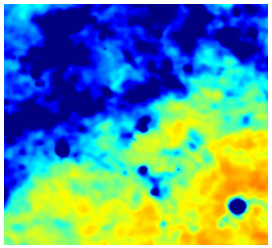
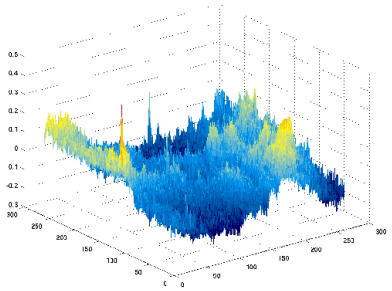
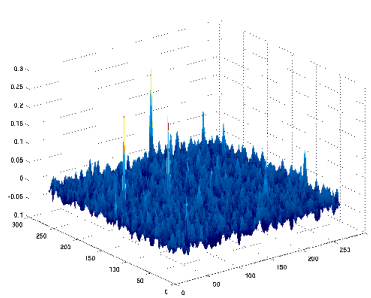
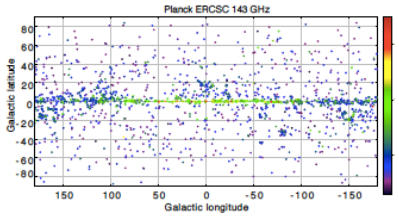
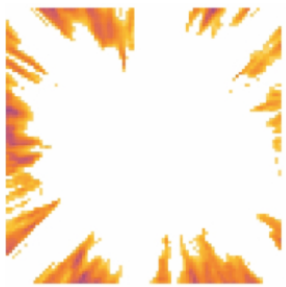
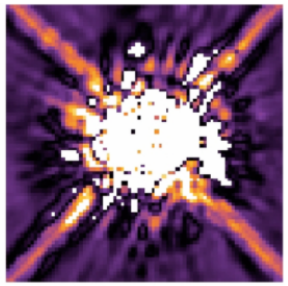
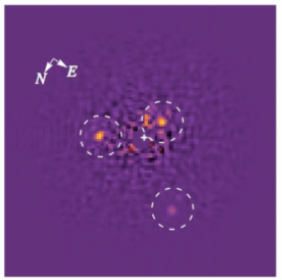
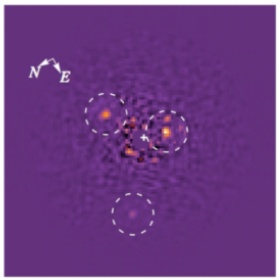
Bayesian High Contrast Imaging Algorithms



Graça Rocha

Dimitri Mawet, Bertrand Mennesson, Tiffany Meshkat, Gautam Vasisht

Michael Bottom, Jeff Jewell, Marie Ygouf

JPL-MPIA Workshop, 10th of April 2018

JPL, Caltech & IPAC

Motivation

- **My Motivation:**
  - To help enhance the detectability of faint exoplanets at small orbital separations from the host star
    - Both ground-based and space-based instruments have not yet achieved the contrast gain needed to detect mature planets with masses lower than 1 Jupiter mass at separations smaller than 0.5".
    - The difficulty arises from the residual glare of starlight at small orbital separations due to diffraction, scattered light, and speckles caused by defects in the optical system
  - New approach → unify source detection and characterization (Position, Flux or Intensity and hence accurate spectrum extraction) into one single rigorous mathematical framework, the Bayesian framework, enabling an adequate hypothesis testing given the S/N of the data.
  - The method will be applied in combination with other post-processing techniques (best suited for this approach), for example KLIP, but now recast in a Bayesian perspective.

- To extend PowellSakes, PwS, a Bayesian approach, to direct imaging data analysis
  - PwS has successfully been applied to detect compact sources immersed in a diffuse background in Planck maps - *Carvalho, Rocha & Hobson, MNRAS, 393, 681C, 2009; Carvalho, Rocha, Hobson & Lasenby, MNRAS, 427, 2011; Bayesian Methods in Cosmology' – CUP, December 2009:; Planck catalog (of compact sources and SZ clusters) papers*

Bayesian – what does it really mean?

❑ What does it mean to recast the problem of planet detection and characterization into a Bayesian perspective?

A. The Bayesian framework entails defining the following key ingredients:

a data model + *a Likelihood shape* + *model parameter priors*

$D(x)=s(x)+n(x)$ $\quad L(d)=P(d|\Theta,H)$ $\quad \Pi(\Theta)=P(\Theta|H)$

B. Next apply Bayes Theorem – to retrieve the probability distribution of the model parameters:

Bayes Theorem ⟶ *Posterior distributions of the model* + *Best Fit models*

$P(\Theta|d,H)$

eg. Maximum Likelihood

Inference → Parameter estimation → $P(\Theta|d,H) = P(d|\Theta,H)\, P(\Theta|H) / P(d|H)$ Ignored

Model Selection → Evidence is crucial $\quad E=P(d|H) = Z$

Expectation of the likelihood over the prior → $\mathcal{Z} = \int L(\Theta)\pi(\Theta)d^D\Theta,$

Bayesian Inference
basic tools

- In contrast to parameter estimation problems → in model selection the evidence takes the central role and is simply the factor required to normalize the posterior:

Evidence

$$\mathcal{Z} = \int L(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})d^D\boldsymbol{\Theta},$$

Evaluation of this multidimensional Integral is a challenging numerical task – resort to sampling techniques: MCMC, Multinest, (Sivia &Skilling 2006; Feroz et al. 2009), etc. or model the posterior as a multivariate Gaussian centered at its peak(s) and apply the Laplace formula (Hobson, Bridle & Lahav 2002).

- The evidence automatically implements Occam's razor:

A simpler theory with compact parameter space will have a larger evidence than a more complicated one, unless the latter is significantly better at explaining the data.

- Model selection between two models H_0 and H_1 can be decided by comparing their respective posterior probabilities given the observed data set d:

$$\frac{\Pr(H_1|d)}{\Pr(H_0|d)} = \frac{\Pr(d|H_1)\Pr(H_1)}{\Pr(d|H_0)\Pr(H_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\frac{\Pr(H_1)}{\Pr(H_0)},$$

$\Pr(H_1)/\Pr(H_0)$ = prior probability ratio for the models

Bayesian Inference
Basic tools

- Prior on the models: The prior ratio $Pr(H_1)/ Pr(H_0)$ on the models is often neglected (i.e. assumed to equal unity), but plays a very important role in the PwS detection criterion

  - let us imagine we know in advance all the true values of the parameters that define an object, which translates into delta-function priors, then we obtain the inequality:

$$\mathrm{SNR} \underset{H_0}{\overset{H_1}{\gtrless}} \sqrt{2\left[\xi + \ln\left(\frac{\Pr(H_0)}{\Pr(H_1)}\right)\right]}$$

  - interpret the term $\ln(Pr(H_0) /Pr(H_1))$ as an extra 'barrier' added to the detection threshold
    - because we are expecting more fake objects than the objects of interest, due to background fluctuations
  - Assuming Poisson statistics for the number of sources and the number of likelihood maxima resulting from the background fluctuations:

$$\frac{\Pr(H_1 \mid N_s)}{\Pr(H_0 \mid N_s)} = \left(\frac{\lambda_1}{\lambda_0}\right)^{N_s}$$

λ_0 =expected number of maxima per unit area resulting from background fluctuations above the minimum limit of detection of the experiment

λ_1 = number density of sources above the same limit - derived from their differential counts

Bayesian Inference
decision theory

- Probability theory defines only a state of knowledge: the posterior probabilities.

There is nothing in probability theory per se that determines how to make decisions based on these probabilities.

- To deal with such difficulties, apply decision theory - one must first define the loss/cost function L(D, E) for the problem at hand.
  - D = set of possible decisions; E = set of true values of the variables to infer.
  - DT can be applied equally well to both parameter estimation and model selection

- Loss function - maps the 'mistakes' in our estimations/selections, D, into positive real values L(D, E), thereby defining the penalty one incurs when making wrong judgments.

Bayesian approach to DT → minimize the expected loss with respect to D:

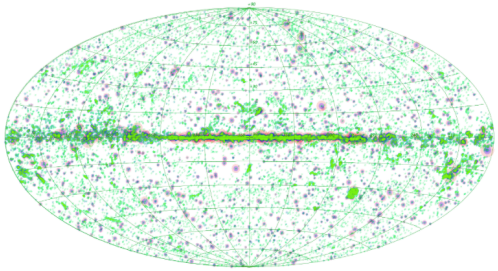
$$\langle L(D, E)\rangle = \iint L(D, E)\,\Pr(D, E)\,\mathrm{d}D\,\mathrm{d}E.$$

'decisions' D = parameter estimates $\Theta\hat{}$; 'entities' E = true values Θ^* of the parameters

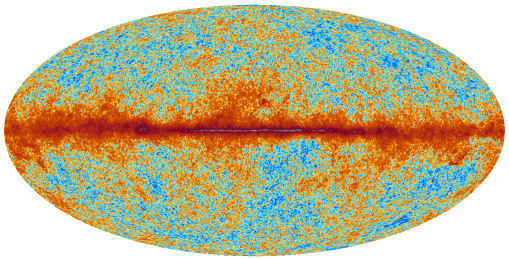
$\varepsilon \equiv \Theta\hat{} - \Theta^*$ → ε^2, $|\varepsilon|$, unity if $||\varepsilon > \Delta$ and zero if $|\varepsilon| < \Delta$

Bayesian Object Detection & characterization

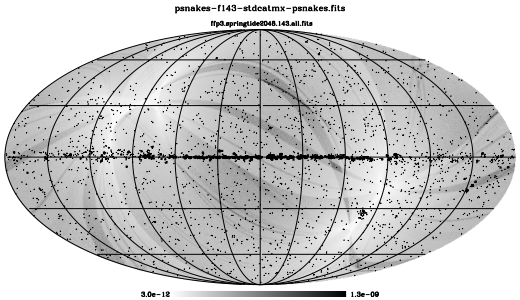
- **Case I** – Detection of point sources in Planck maps ie detection of discrete objects immersed in a diffuse background: white noise, correlated noise, anisotropic, non-gaussian emission, destriping residuals, etc…



Point Sources

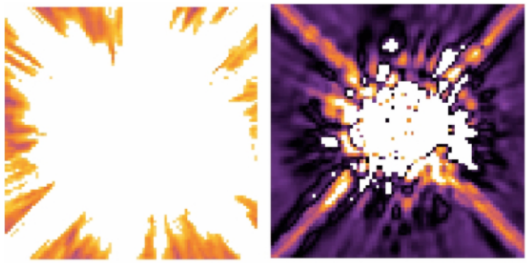


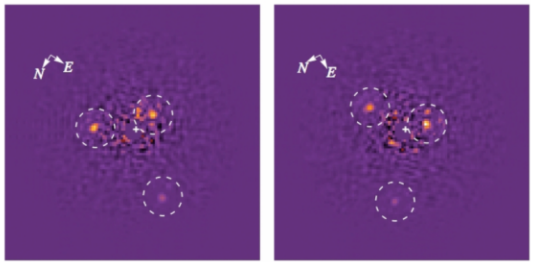
CMB fluctuations
+
Galactic emission



Noise

- **Extra complications:**
  - The Cosmic Microwave Background, CMB, emission fluctuations varies on a characteristic scale of order ~10 arcmin, similar to that of extragalactic 'point' (i.e. beam-shaped) sources or the Sunyaev–Zel'dovich (SZ) effect in galaxy clusters, the objects we are interested in
- **Case II**: Detection of planets around a star in direct imaging data





Frequentist approach (common)

- Apply a linear filter Ψ(x) to the original image d(x) and analyse the filtered field:

$$d_{\rm f}(x) = \int \psi(x-y)\, d(y)\, {\rm d}^2 y.$$

- the filtering process as 'optimally boosting' (in a linear sense) the signal from discrete objects, with a given spatial template, and simultaneously suppressing emission from the background.
- If the original image contains N_{obj} objects at positions X_i with amplitudes A_i:

$$d(x) \equiv s(x) + n(x) = \sum_{i=1}^{N_{\rm obj}} A_i t(x - X_i) + n(x),$$

signal

generalised noise

- It is straightforward to design an optimal filter function ψ(x) such that the filtered field (1) has the following properties:
  - (i) $d_f(X_k)$ is an unbiased estimator of A_i;
  - (ii) the variance of the filtered noise field $n_f(x)$ is minimized;

- the corresponding function ψ(x) is the standard matched filter (eg. Haehnelt & Tegmark 1996).

$$\psi(k) \propto \frac{B(k)}{P(k)} \qquad \text{where} \qquad P(k) = \langle N(k) N(k) \rangle \quad \text{and B(k) is the beam or PSF}$$

Standard approach
Matched Filter

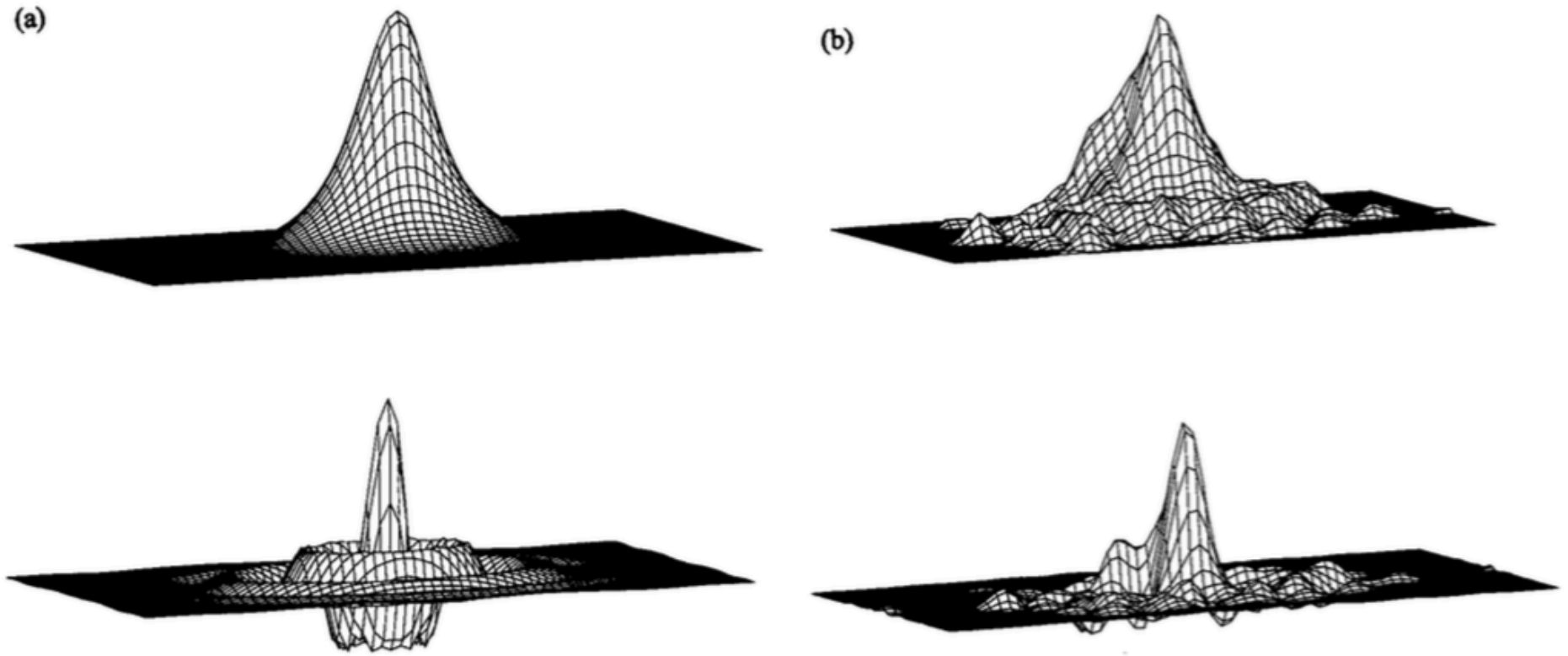


Figure 2. (a) Typical filter function for an axisymmetric cluster in arbitrary units. The cluster profile is shown for comparison. The angular scales are the same. (b) Same as (a) for a non-axisymmetric cluster. The cluster is shown at the top, the corresponding filter function at the bottom.

Bayesian vs frequentist

- Some points to consider:

- The approaches outlined have been shown to produce good results, BUT
  - The filtering process is only optimal among the rather limited class of linear filters and is logically separated from the subsequent object detection step performed on the filtered map(s).
  - The detection threshold is empirically established - while the threshold is a logical byproduct of the framework in the Bayesian approach (*Carvalho, Rocha & Hobson, MNRAS, 393, 681C, 2009; Carvalho, Rocha, Hobson & Lasenby, MNRAS, 427, 201*)
  - It is well known that MFs are excellent at finding and locating sources, but not as good at estimating fluxes
  - Do not capitalize on previous knowledge both theoretical (modeling) and observational - for example using prior information we can enhance the probability of detecting very faint sources reliably (see references above)

- Bayesian approach
  - Hobson & McLachlan (2003) first introduced this approach, but it was too slow and slower than the traditional approaches.
  - New efficient approach with PowellSnakes - PWS

Bayesian Object Detection

Case I: Detection of Point sources in Planck maps - PWS

- *Data Model*

Suppose we want to extract the amplitude A of a signal with a known spatial distribution t(x) from a measured signal d (x) which is contaminated by noise n(x) → Start by defining your data model:

Amplitude of the signal – what we want to know

Data Model $\rightarrow$ $d(x) = s(x) + n(x) = At(x) + n(x)$

Signal *Generalized Noise* *Signal spatial template - known*

$$\boldsymbol{d}(\boldsymbol{x}) = \sum_{j=1}^{N_s} \boldsymbol{s}_j(\boldsymbol{x}; \boldsymbol{\Theta}_j) + \boldsymbol{b}(\boldsymbol{x}) + \boldsymbol{n}(\boldsymbol{x}),$$

B(x) = Background sky emission: Galactic emission, CMB fluctuations, ..

$$s(x;a) = A \exp\left[-\frac{(x-X)^2 + (y-Y)^2}{2R^2}\right]$$

Signal (Source)

$$\boldsymbol{s}_j(\boldsymbol{x}; \boldsymbol{\Theta}_j) = A_j \boldsymbol{f}(\boldsymbol{\phi}_j)\boldsymbol{\tau}(\boldsymbol{x} - \boldsymbol{X}_j; \boldsymbol{a}_j),$$

f=emission coefficients at each frequency

$$\boldsymbol{\Theta}_j = \{A_j, \boldsymbol{X}_j, \boldsymbol{a}_j, \boldsymbol{\phi}_j\}$$

Source parameters

A_j=amplitude, X_j=position,
a_j= shape parameters,
Φ_j=emission law parameters

Carvalho, Rocha & Hobson 2009, & Lasenby 2011, Rocha in prep

Bayesian Object Detection

Case I: Detection of Point sources in Planck maps – PWS

- *Data Model*

$$d(x) = s(x;\theta) + AY(x) + n_{inst}(x) + F(x)$$

Diffuse galactic foregrounds ignored here

CMB fluctuations → $\frac{\Delta T}{T} \equiv \sum_{\ell m} a_{\ell m} Y_{\ell m}(\theta, \phi)$

$$\langle a_{\ell m} a_{\ell' m'} \rangle = C_\ell \, \delta_{\ell\ell'} \delta_{mm'}$$

Statistical isotropy

Rotational invariance

$$p(a_{\ell m}) = N(a_{\ell m}; 0, C_\ell)$$

$$C = \sum_{\ell=2}^{\ell_{mas}} \frac{2\ell+1}{4\pi} C_\ell P_\ell(\alpha) \qquad N = C + N_{inst} \qquad N_{inst} = \left\langle n_{inst} n_{inst}^T \right\rangle$$

- The CMB fluctuations are a statistically homogeneous Gaussian random field → The correlation matrix is circulante hence diagonal in Fourier or Harmonic space – the spherical harmonics or Fourier basis are our PCA basis

Carvalho, Rocha & Hobson 2009; & Lasenby 2011, Rocha in prep

Bayesian Object Detection

Case II: Exoplanets in direct imaging data (a)

- *Data Model - a* *Data model* $\rightarrow$ $d(x) = s(x;\theta) + A.w\,(x) + n_{inst}(x)$

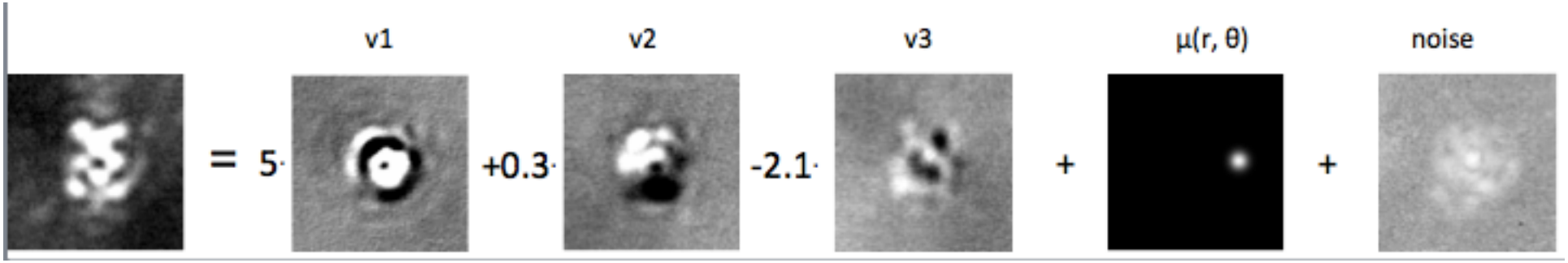
planet Speckles Noise-instrument

$$y = A \cdot w + \mu(\theta) + \text{noise}$$

- The systematics design matrix *A* is a collection the systematics vectors {V}, these could be:
  - The principal components of the data
  - A sparse decomposition of the data
  - The non-negative matrix factorization
- *A.w* describes the systematics in a single data frame
- μ(r, θ) describes the signal

$$A = \left[\uparrow \; V1 \; \downarrow \quad \uparrow \; V2 \; \downarrow \quad \uparrow \; V3 \; \downarrow \quad \cdots \right] \qquad w = \begin{bmatrix} 5 \\ 0.3 \\ -2.1 \\ \vdots \end{bmatrix}$$

Credit Mike Bottom



Bayesian Object Detection

Case II: Exoplanets in direct imaging data (b)

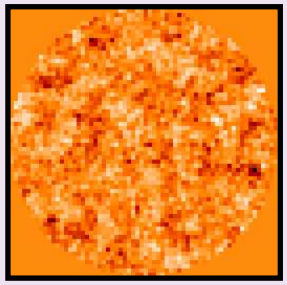
MEDUSAE

Data Model-b

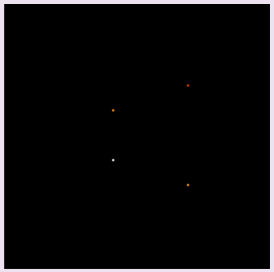
Input parameters = Unknowns

Star flux f_λ^*

Aberrations ϕ



Object o_λ

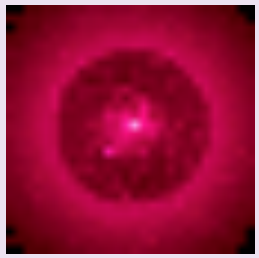


- Start with a guess
- Minimize criteria
- Iterate

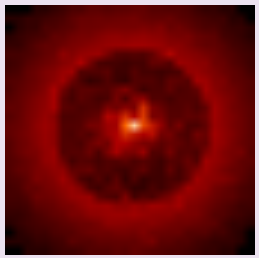
Inversion

Data model $\rightarrow i_\lambda = f_\lambda^* \cdot h_\lambda^c(\phi) + o_\lambda * h_\lambda^{nc}(\phi) + n$

Data

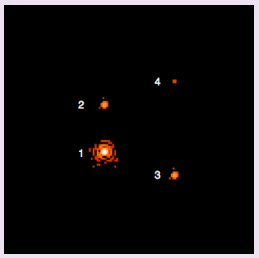


=



Star

+



Planet image

Difference: model of the coronographic PSF - from instrument model *Ygouf et al. 2013, A&A*

Bayesian Object Detection

Case I: Detection of Point sources in Planck maps - PWS

- *Likelihoods*

The form of the likelihood is determined by the statistical properties of the generalized noise:

This case Gaussian distributed noise (white + correlated noise) → Multivariate Gaussian Likelihood

$$p(d \mid \Theta, H) = L(\Theta) = \frac{\exp\left\{-\frac{1}{2}\left[d - s(\Theta)\right]^T N^{-1}\left[d - s(\Theta)\right]\right\}}{\sqrt{(2\pi)^{N_{pix}}\left|N\right|}} \qquad N = C + N_{inst}$$

$$\langle T_{i_1} T_{i_2} \rangle = \underbrace{\sum_{\ell=2}^{\ell_{\max}} \frac{2\ell+1}{4\pi} \hat{C}_\ell P_\ell(\theta_{i_1 i_2})}_{C} + \underbrace{\mathbf{N_{i_1 i_2}}}_{N_{inst}},$$

- This is equivalent to considering the Likelihood expression:

$$p(d \mid \Theta, A, H) = L(\Theta, A) = \frac{\exp\left\{-\frac{1}{2}\left[d - AY - s(\Theta)\right]^T N_{inst}^{-1}\left[d - AY - s(\Theta)\right]\right\}}{\sqrt{(2\pi)^{N_{pix}}\left|N_{inst}\right|}}$$

- \+ marginalize over A

Carvalho, Rocha & Hobson 2009; & Lasenby 2011, Rocha in prep

Bayesian Object detection

Case I: Detection of Point sources in Planck maps

- *Priors*
  - The Jeffreys (Jeffreys 1961) rule for constructing ignorance priors for the one-dimensional case read:

$$\pi(\theta) \propto \mathcal{J}^{1/2}(\theta), \qquad \text{where} \qquad \mathcal{J}(\theta) \equiv -\left\langle \frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2} \right\rangle$$

Fisher information

- Prior on position: if the sky patches used are sufficiently small, our locally uniform model can easily cope with clustering when the gradient of the density of sources is small across the patch boundaries.

$$\Pr\left(\boldsymbol{X}^{N_s} \mid N_s,\ N_{\rm pix}\right) = \frac{1}{{N_{\rm pix}}^{N_s}},$$

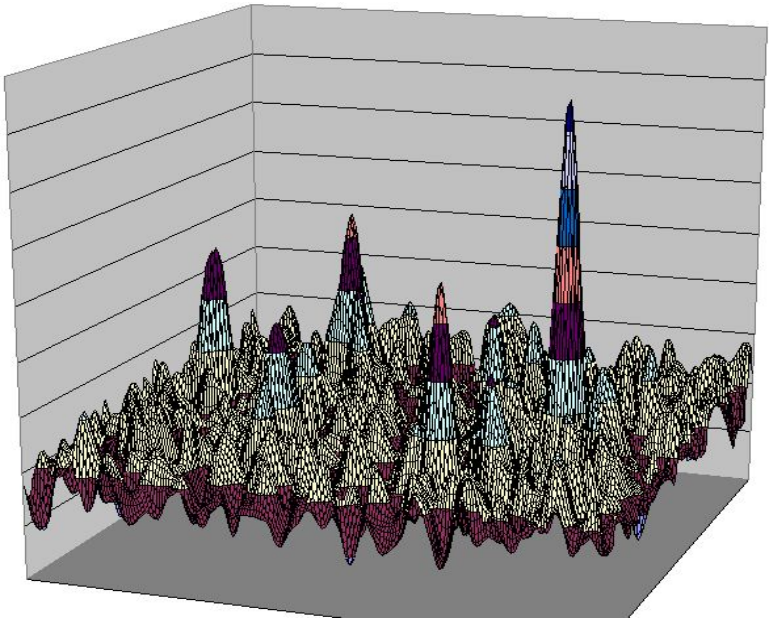
Npix is the number of pixels in each patch and N_s is the number of sources in that patch

Carvalho, Rocha & Hobson 2009; & Lasenby 2011, Rocha in prep

Likelihood manifold
and Posteriors - PWS

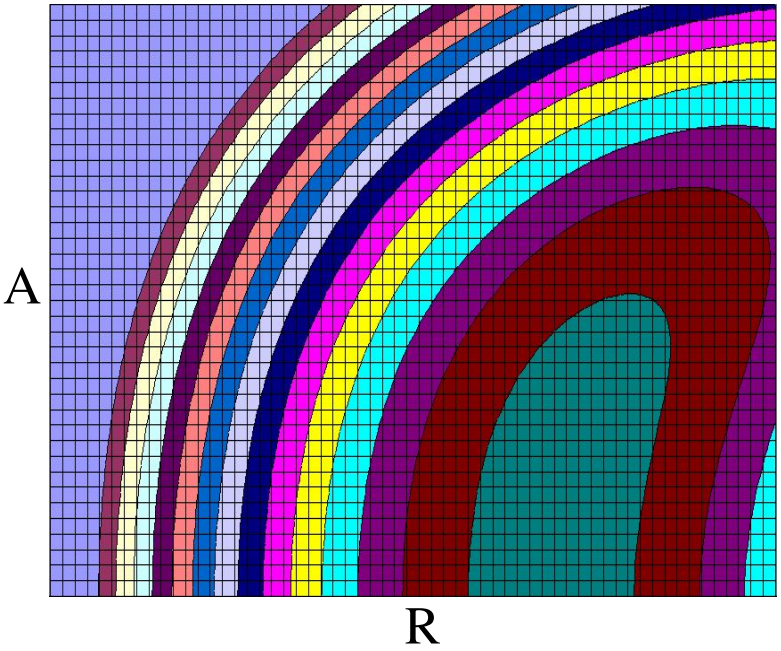
- Bayesian perspective → The filtered field is the projection of the likelihood manifold onto the sub-space of position parameters X_j
- Ie the Likelihood (after marginalizing over the CMB harmonic coefficients) as function of (X,Y)

$$p(d \mid \Theta, H) = L(\Theta) = \frac{\exp\left\{-\frac{1}{2}\left[d - s(\Theta)\right]^T N^{-1}\left[d - s(\Theta)\right]\right\}}{\sqrt{(2\pi)^{N_{pix}}\,|N|}}$$



Position subspace (X,Y) ; High res antenna

2d-contour plots of the posterior distributions of the A and R of the source



(A,R) subspace ; (X_0, Y_0) of a maximum

Likelihood

Case II: Exoplanets in direct imaging data (a)

- *Likelihoods*

$$p(y|\theta, w) = \mathcal{N}(y; Aw + \mu(\theta), C)$$

$$= \frac{\exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{A}\cdot\mathbf{w} - \boldsymbol{\mu}(\boldsymbol{\theta}))^{\mathrm{T}} \boldsymbol{C}^{-1}(\mathbf{y} - \mathbf{A}\cdot\mathbf{w} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right)}{\sqrt{(2\pi)^k |C|}}$$

- C = Covariance matrix of the residual noise after speckles removal, white noise, normally distributed, hence a diagonal matrix
- The speckles are described by some combination **w** of feature vectors **A**
- Marginalize over w, assuming $p(w) = \mathcal{N}(w; 0, \Lambda)$

$$p(y|\theta) = \int_{-\infty}^{\infty} p(w)p(y|\theta, w)dw = \mathcal{N}(y; \mu(\theta), C + A\Lambda A^T)$$

$$p(d \mid \Theta, H) = L(\Theta) = \frac{\exp\left\{-\frac{1}{2}\left[d - \mu(\Theta)\right]^T (C + A\Lambda A^T)^{-1}\left[d - \mu(\Theta)\right]\right\}}{\sqrt{(2\pi)^k \left|C + A\Lambda A^T\right|}}$$

- **Note**: this is exactly the procedure PWS follows for point sources detection and is here applied for detection and characterization of Exoplanets

Credit Mike Bottom

Priors and Posteriors

Case II: Exoplanets in direct imaging data

- *Priors, Posteriors*

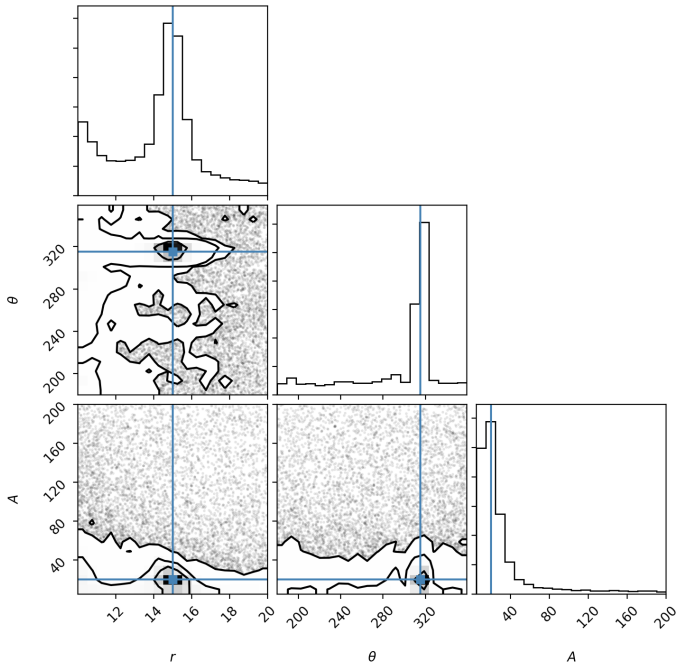
Shifted-scaled PSF model $\mu(x_0, y_0, a) = a \cdot \mathrm{PSF}[x - x_0, y - y_0]$

Uniform prior on position $p(x_0), p(y_0) \propto 1$

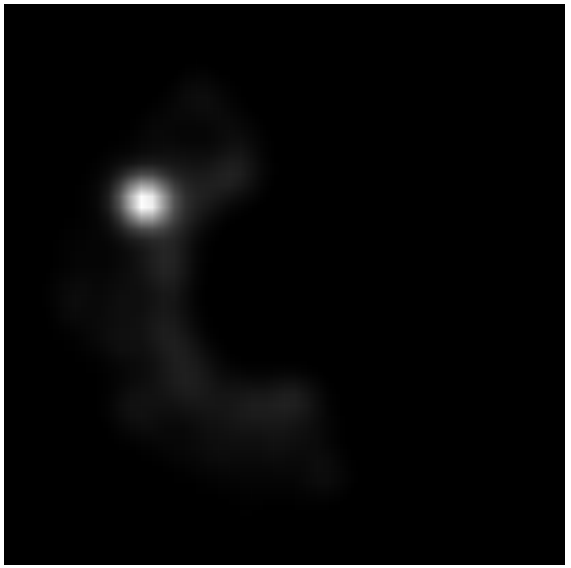
Scale-invariant prior on amplitude $p(a) \propto 1/a$

Priors

Get $p(\theta|y) \propto p(y|\theta)p(\theta)$



For 20 PCAs



Probability map

Inference only → No model selection yet

Credit Mike Bottom

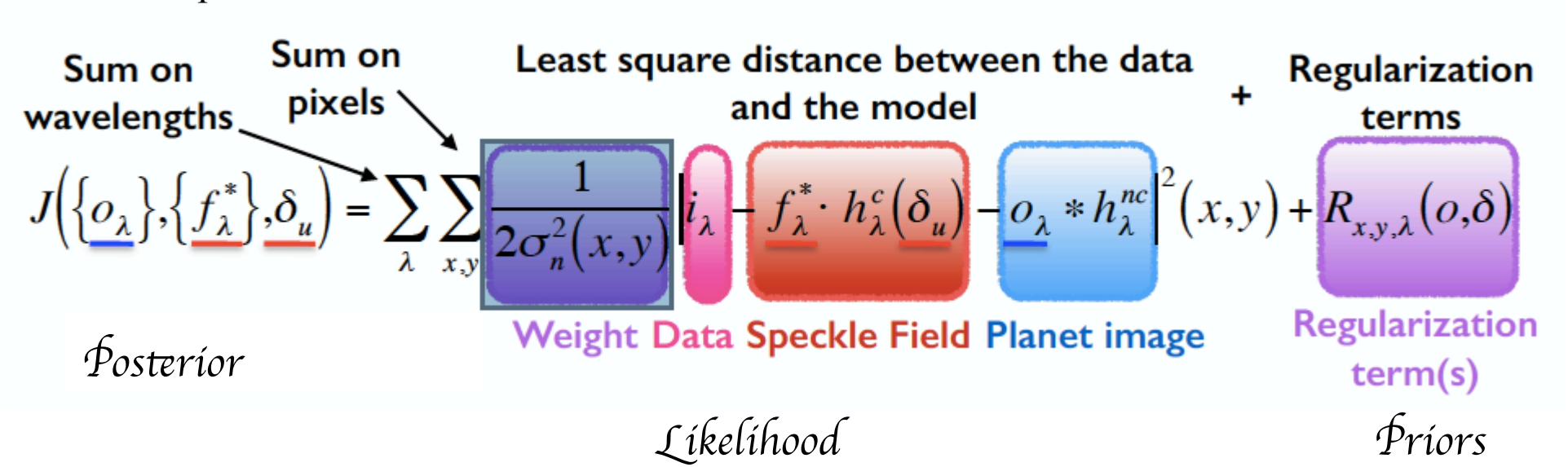
Bayesian Object Detection

Case II: Exoplanets in direct imaging data (b)

MEDUSAE

- *Likelihood, priors and posteriors*

Apply iterative scheme – start from a guess aberration map and planet(s) map – maximize posterior - replace by new aberration and planet(s) map if J is larger – stop when the 'stop condition' is achieved



Currently

- Estimates the maximum of the posterior distribution :
  - Best fit model of the aberration map + planet(s) map
- Does not estimate the posterior distribution of the parameters yet

Ygouf et al. 2013, A&A

Bayesian Object Detection

How frequentist approach naturally emerges within the Bayesian framework

- For model selection: we are interested in the likelihood ratio between the hypothesis H_s that objects (of a given source type s) are present and the null hypothesis H_0 that there are no such objects (= corresponds to setting the sources signal s(x;Θ) to zero):

$$\ln\left[\frac{\mathcal{L}_{H_s}(\boldsymbol{\Theta})}{\mathcal{L}_{H_0}(\boldsymbol{\Theta})}\right] = \sum_{\eta} \tilde{d}^{t}(\eta)\mathcal{N}^{-1}(\eta)\tilde{s}(\eta;\boldsymbol{\Theta}) - \frac{1}{2}\sum_{\eta} \tilde{s}^{t}(\eta;\boldsymbol{\Theta})\mathcal{N}^{-1}(\eta)\tilde{s}(\eta;\boldsymbol{\Theta}),$$

tilde denotes a Fourier transform
K=2πη mode wavenumber ;
N(η)= generalized noise cross power spectra

- Maximizing the likelihood ratio, with respect to the source amplitudes A_j → we recover the expression for the MMF

$$\widehat{A}_j = \frac{\mathcal{F}^{-1}\left[\mathcal{P}_j(\eta)\tilde{\tau}(-\eta;\widehat{a}_k)\right]_{\widehat{X}_j}}{\sum_{\eta}\mathcal{Q}_{jj}(\eta)\,|\,\tilde{\tau}(\eta;\widehat{a}_j)\,|^2} \qquad \hat{A}(\underline{X},R) = \frac{t^{\mathrm{T}}(\underline{X},R)N^{-1}d}{t^{\mathrm{T}}(\underline{X},R)N^{-1}t(\underline{X},R)}.$$

$$\psi(k) \propto \frac{B(k)}{P(k)} \qquad t(X,R) = \exp\left[\frac{X^2+Y^2}{2R^2}\right]$$

- Substituting this maximum-likelihood (ML) estimate onto the Likelihood ratio expression we get for the jth object:

$$\max\left[\ln\left(\frac{\mathcal{L}_{H_s}}{\mathcal{L}_{H_0}}\right)\right] = \frac{1}{2}\sum_{\eta}\mathcal{Q}_{jj}(\eta)\,|\,\tilde{\tau}(\eta;\widehat{a}_j)\,|^2\widehat{A}_j^2 = \frac{1}{2}\widehat{\mathrm{SNR}}_j^{\,2}$$

$\widehat{\mathrm{SNR}}_j$ SNR (at the peak) of the jth source

- The traditional approach to catalogue making, in which one compares the maximum SNR of the putative detections to some threshold = performing a generalized likelihood ratio test GLRT

PWSIII

- At the Post-Processing stage there are a number of Image Processing techniques that aim at modeling and subtracting the stellar Point Spread Function, PSF, to allow the planet to become detectable, in effect increasing the contrast achievable next to a bright star:

  - Angular Differential Imaging, ADI, (Marois et al. 2008)
  - LOCI, (Lafreniere et al. 2007);
  - Reference Differential Imaging RDI
  - Principal Component Analysis, PCA, (Amara & Quanz 2012, Meshkat et al. 2014)
  - KLIP (Soummer et al. 2012) which uses the Karhunen-Loeve, KL, transform to model the PSF
  - Stochastic speckle discrimination, SSD, (Gladysz & Christou 2008)
  - Enhanced faint companion photometry and astrometry using wavelength diversity (Burke & Devaney 2010)
  - KLIP-FM (Pueyo 2016)

  etc…

PWSIII

(1) Construct the Likelihood,. L ((x, y), t, λ, I):

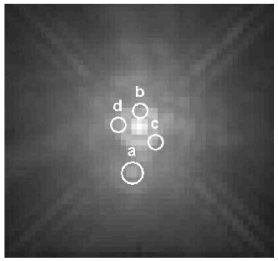
- Assuming subspaces are independent we can recast it as:

$$L((x, y), t, \lambda, I) = L((x, y), I) \times L(t) \times L(\lambda)$$

  - Spatial likelihood - Multivariate Gaussian – *discussed here*
    - Connection to to optimal adaptive matched filter (MF) done : current study
  - Temporal Likelihood - mild Non- Gaussian likelihood specified by the first 3 moments of the pdf

$$p(x) = |\psi|^2 = e^{-x^2/(2\sigma_0^2)} \left| \sum_n \alpha_n C_n H_n \left(\frac{x}{\sqrt{2}\sigma_0} \right) \right|^2 ,$$

Rocha et al. 2001, 2005



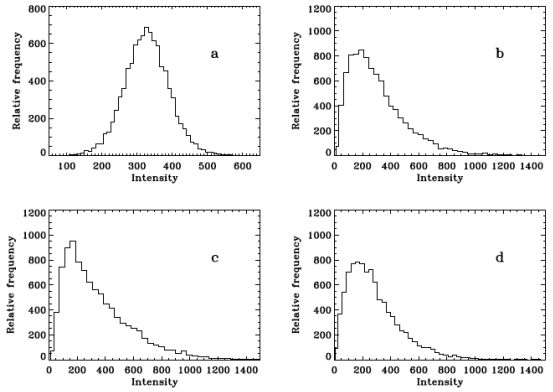


FIG. 4.—Histograms of intensity at the locations corresponding to (*a*) the companion, and (*b*, *c*, *d*) the static speckles.

Gladysz & Christou 2008

PWSIII

- Multi-wavelength Likelihood
  - Related to estimating the covariance of the data, estimate the PSF and construct a new Multi-Matched filter, MMF, a whitening filter (akin to the Hotelling observer, *Burke & Devaney 2010*)

- Priors of the model parameters – constructed from external information, previous observations, relevant information that helps distinguishing the signal from the noise

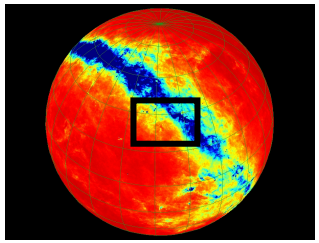
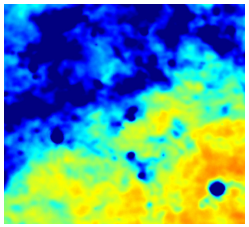
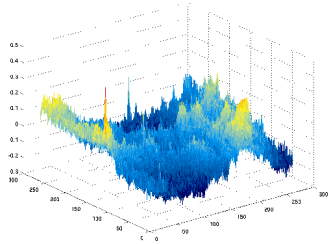
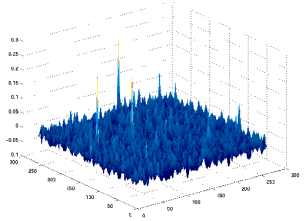
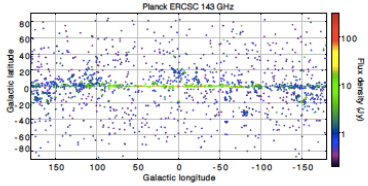
(2) Estimate the posterior distributions of the model parameters + the evidence ratios of the competing models

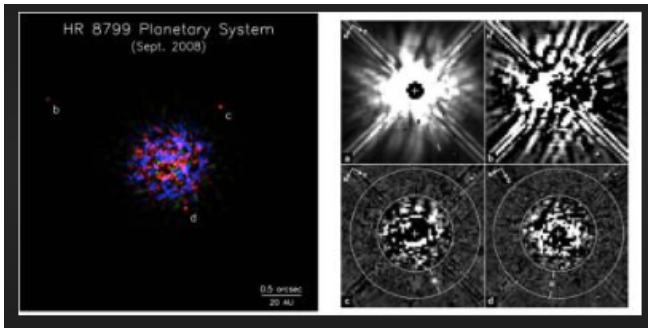
(3) Potentially iterate previous steps

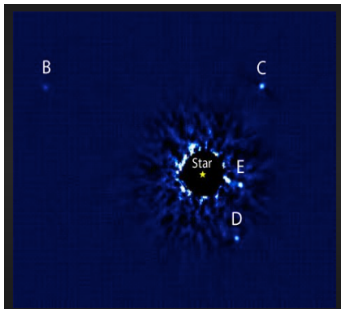
- Quantify performance of PWSIII using simulations with injected planets

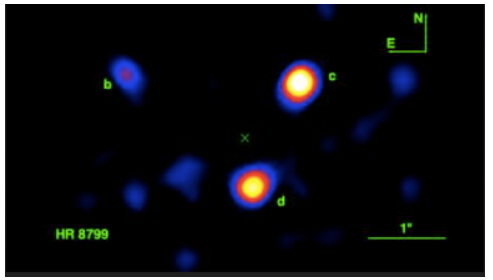
Conclusions

Bayesian High Contrast Imaging Algorithms







Graça Rocha

Dimitri Mawet, Bertrand Mennesson, Tiffany Meshkat, Gautam Vasisht

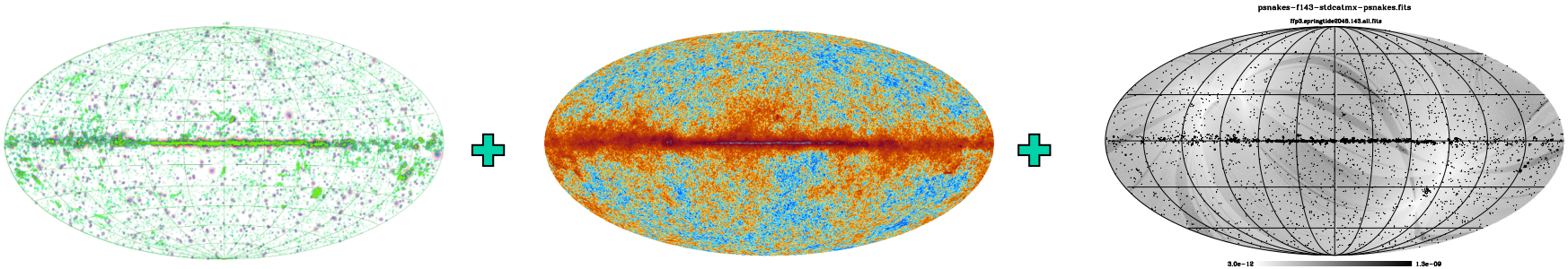
Michael Bottom, Jeff Jewell, Marie Ygouf

JPL-MPIA meeting, KISS, 10th of April 2018

JPL, Caltech & IPAC

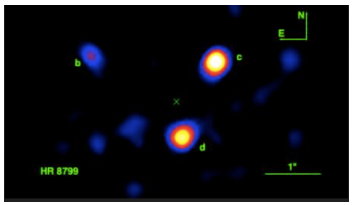
Bayesian Object Detection & characterization

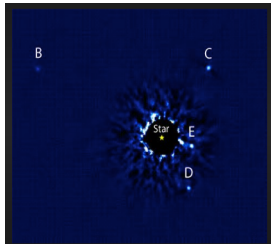
- **Case I** – Detection of point sources in Planck maps ie detection of discrete objects immersed in a diffuse background: white noise, correlated noise, anisotropic, non-gaussian emission, destriping residuals, etc…

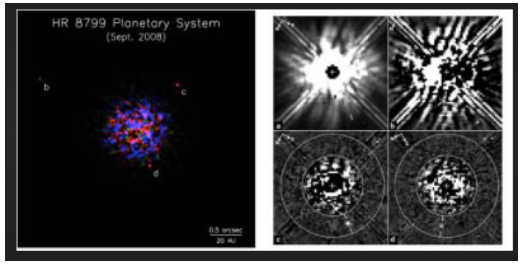


Point Sources + CMB fluctuations + Galactic emission + Noise

- **Extra complications:**
  - The Cosmic Microwave Background, CMB, emission fluctuations varies on a characteristic scale of order ~10 arcmin, similar to that of extragalactic 'point' (i.e. beam-shaped) sources or the Sunyaev–Zel'dovich (SZ) effect in galaxy clusters, the objects we are interested in
- **Case II**: Detection of planets around a star in direct imaging data







Bayesian Inference
decision theory

- Probability theory defines only a state of knowledge: the posterior probabilities.

There is nothing in probability theory per se that determines how to make decisions based on these probabilities.

- To deal with such difficulties, apply decision theory - one must first define the loss/cost function L(D, E) for the problem at hand,
  - where D is the set of possible decisions and E is the set of true values of the entities one is attempting to infer.
  - DT can be applied equally well to both parameter estimation and model selection

- Loss function - maps the 'mistakes' in our estimations/selections, D, into positive real values L(D, E), thereby defining the penalty one incurs when making wrong judgments.

The Bayesian approach to DT is simply to minimize, with respect to D, the expected loss:

$$\langle L(D, E)\rangle = \iint L(D, E)\ \Pr(D, E)\,\mathrm{d}D\,\mathrm{d}E.$$

'decisions' D = parameter estimates $\Theta^{\wedge}$; 'entities' E = true values Θ^{*} of the parameters

PWSIII

(1) Construct the Likelihood,. L ((x, y), t, λ, I):

- As the subspace are independent we can recast it as: L ((x, y), t, λ, I)= L ((x, y), I) x L (t) x L (λ);
  - Spatial likelihood - Multivariate Gaussian
  - Temporal Likelihood - mild Non- Gaussian likelihood specified by the first 3 moments of a PDF *(Rocha et al. 2001, Rocha et al. 2005)*
- The priors for the model parameters will be constructed based on previous observations and any other relevant information that helps distinguishing the signal from the noise
- Construct an optimal adaptive matched filter (MF): based on the spatial estimation of the noise (using KLIP for example) and a spatial model for the planet (e.g. a Airy function) and/or – current study
- Use multi-wavelength data to estimate the covariance of the data, estimate the PSF and construct a new Multi-Matched filter, MMF, a whitening filter (akin to the Hotelling observer),

(2) Estimate the posterior distributions of the model parameters + the evidence ratios of the competing models

(3) To improve detection characterization - repeat the previous step - this time as a temporal analysis of the peaks in the previous Likelihood manifold (filtered map in the positional subspace):

- (a) Construction of a potentially Non-Gaussian Likelihood based and construction of priors for the moments of the distribution, (b) estimating the posterior distributions of these moments and (c) the evidence ratios of the competing probability distributions

- Quantify performance of PWSIII using simulations with injected planets

Bayesian Object detection

Ingredients: Data Model, Likelihood and Priors

- Prior on the models:
  - Only background - the density of maxima, λ_0, resulting from the filtering procedure that creates the likelihood manifold can be estimated using the 2D Rice formula:

$$n_b(\nu,\kappa,\epsilon) = \frac{8\sqrt{3}\tilde{n}_b}{\pi\sqrt{1-\rho^2}}\,\epsilon(\kappa^2-4\epsilon^2)\,\mathrm{e}^{-\frac{1}{2}\nu^2-4\epsilon^2-\frac{(\kappa-\rho\nu)^2}{2(1-\rho^2)}},$$

  - Where: $\nu=A/\sigma$ is the 'normalized peak amplitude'; k='normalized curvature'; ε='normalized shear'

$$\rho = \sigma_1^2/(\sigma_0\sigma_2), \qquad \sigma_n^2 = (2\pi)^{1+2n}\int_0^\infty \eta^{1+2n}\,|\,\mathcal{P}(\eta)\,|^2\,d\eta$$

Marginalizing over all parameters – we obtain the expected density of maxima of a Gaussian filtered field:

$$\tilde{n}_b = \frac{\sigma_2^2}{8\pi\sqrt{3}\sigma_1^2}.$$

- But we are only interested in the peaks above a certain level ν_0 - as PwS pre-selects the putative detections by imposing a minimum SNR level before attempting the evidence evaluation - The main reason for adopting this early selection is computational efficiency

Bayesian Object detection

Ingredients: Data Model, Likelihood and Priors

- Prior on the models - considering this flux cut:

$$n_b(\nu) = \frac{\tilde{n}_b\sqrt{6}}{2\sqrt{\pi}\rho_1}\left\{\left(1+\mathrm{erf}\left(\frac{\rho}{\rho_1\rho_2}\nu\right)\right)\mathrm{e}^{-\nu^2\left(\frac{1}{2}+\left(\frac{\rho}{\rho_2}\right)^2\right)}\left(\frac{\rho}{\rho_2}\right)\right.$$
$$+\left(1+\mathrm{erf}\left(\frac{\rho}{\rho_1}\nu\right)\right)\mathrm{e}^{-\frac{\nu^2}{2}}(\nu^2-1)\rho^2\rho_1$$
$$\left.+\frac{\nu\mathrm{e}^{-\nu^2\left(\frac{1}{2}+\left(\frac{\rho}{\rho_1}\right)^2\right)}}{\sqrt{\pi}}\rho\rho_1^2\right\},$$

$$\lambda_0 = \int_{\nu_0}^{\infty} n_b(\nu)\mathrm{d}\nu,$$

$$\rho_1 = \sqrt{2(1-\rho^2)}$$
$$\rho_2 = \sqrt{2(\tfrac{3}{2}-\rho^2)}.$$

- The expected number count of targeted objects above a certain flux threshold S, $\lambda_1 \equiv \langle N(>S)\rangle$, may be easily derived from their differential counts.
- the expected differential counts for a certain population type of galaxies per flux interval at a certain frequency always follow a power law:

$$dN\varphi/dS = A\varphi\, S^{-b} \quad \text{(de Zotti et al. 2005)}$$

$$\lambda_1 = N(>S_0) = \int_{S_0}^{\infty}\frac{\mathrm{d}N_\phi}{\mathrm{d}S}\mathrm{d}S = A_\phi\,(1-b)^{-1}\,S_0^{1-b},\ b\neq 1$$

With $\{A_\Phi, b\}$ free - provided by the user to target a specific type of object and/or nstrumental setup

Object detection strategy

- So far, we have only developed the logic and probabilistic underpinnings of PwS.
  - It is now time to bring all the pieces together into a consistent strategy for the detection and characterization of discrete objects:

The single object approach

Estimating the posterior odds ratio is a daunting task - find an effective solution, make assumptions:

(i) the objects of interest are 'well separated';

(ii) all variables pertaining to each individual source are mutually independent

Separate the integrals associated with each source - deal with each source independently, one at a time - 'single object approach' - replaces a single N_{param} x N_s - dimensional integral with a sequence of N_s integrals, each of dimension N_{param}

- The odds of the model H_1 (for a given source type), given N_s such sources, reads:

$$\frac{\Pr(H_1 \mid \boldsymbol{d}, N_s)}{\Pr(H_0 \mid \boldsymbol{d}, N_s)} = (N_{\rm pix}\Delta_p)^{-N_s} e^{-\lambda_1} \frac{\lambda_1^{N_s}}{N_s!} \left(\frac{\lambda_1}{\lambda_0}\right)^{N_s} \prod_{j=1}^{N_s} \mathcal{Z}_{1j},$$

- Where the 'partial evidence' for each individual source:

$$\mathcal{Z}_{1j} \equiv \int \frac{\mathcal{L}_1(\boldsymbol{\Theta}_j)}{\mathcal{L}_0} \pi(\boldsymbol{\Theta}_j)\, d\boldsymbol{\Theta}_j.$$

Object detection strategy

The single object approach

- Taking logarithms and rearranging:

$$\ln\left[\frac{\Pr(H_1 \mid \boldsymbol{d}, N_s)}{\Pr(H_0 \mid \boldsymbol{d}, N_s)}\right] = \sum_{j=1}^{N_s} \ln(\mathcal{Z}_{1j}) - N_s P_s, \qquad (45)$$

- Where we defined the 'penalty per source', P_s:

$$P_s \equiv \ln \Lambda_s^{-1} + \ln\left(\frac{\lambda_0}{\lambda_1}\right) + \frac{1}{N_s}\left[\lambda_1 + \ln N_s!\right].$$

- Thus, the total ln (odds) for a single patch is the sum of the partial ln (evidence) for each source, plus an extra global penalty term that contributes, in the majority of the cases, negatively to the final balance and does not depend on any particular source, but exclusively on the ensemble properties

Object detection strategy

- One possible procedure to select the optimal set of sources is as follows:

> (i) evaluate $\mathcal{Z}_j$ for each source;
> (ii) partition the candidate detections into the pre-defined homogenous zones. For each zone:
> (a) sort the candidate detections in descending order of $\mathcal{Z}$ and number them ($j = 1\ldots$);
> (b) with N_s, iterate down the list of catalogue lines evaluating formula (45);
> (c) stop when moving from $N_s = k$ to its successor ($N_s = k + 1$) makes expression (45) decrease.
> (d) This means, $\widehat{N_s}$ (the value of N_s that maximizes the evidence ratio) has been found and the 'proto-catalogue' is formed from the first k lines.

we have only selected the set of detections that maximizes the odds.

- the ln(odds) for each object plays a pivotal role in catalogue making:

$$\ln(\text{odds})_j \equiv \ln\left[\frac{\Pr(H_1 \mid d)}{\Pr(H_0 \mid d)}\right]_j = \ln(\mathcal{Z}_{1j}) - \widehat{P}_s,$$

Ps is the penalty per source evaluated at Ns or the catalogue penalty per source),

Object detection strategy

- Evaluation of the odds ratio
  - 'Brute force' evaluation of the evidence integrals is still not feasible
  - Use MCMC methods and thermodynamic integration - can fail when the posterior distribution is very complex
  - use 'Nested Sampling' (Sivia &Skiling 2006),which is much more efficient, although not without its difficulties; MultiNest' (Feroz et al. 2009) efficient implementation of the nested sampling algorithm, which is capable of exploring high-dimensional multi-modal posteriors; other simpler nested sampling scheme (Mukherjee, Parkinson & Liddle 2006) perform well.

- Or use another approach (as in PwsI):
  - PwS I started a Powell minimization chain (hence the name 'PowellSnakes') in many different locations of the manifold in an attempt to find all the maxima - where the Brent line minimizer was 'enhanced' with an ancillary step to allow it to 'tunnel' from one minimum to the next.
  - Explore the fact that we can separate the position variables from all others- so first locate maxima in position space, then start a four-dimensional PwS optimization at each such location to find the ML parameters for that particular peak

Object detection strategy

- Exploring the posterior distribution

  - Our initial step provides the ML estimates and the SNR of each detection candidates
  - Only a much smaller sub-set is chosen based on an SNR threshold.
  - This shorter list is then sorted in descending order of SNR and one-by-one the maxima are sent to the nested sampler,
  - The nested sampler returns an evidence estimate and a set of weighted samples that we use to model the full joint posterior distribution
  - The final catalogue is almost completely independent of the SNR threshold if this is not too high
  - From these samples we can compute any parameter estimate, draw joint distribution surfaces, predict HPD intervals of any content over the marginalized distributions to infer the parameter uncertainties

Catalog making

- The last step of PwS is to assemble the final catalogue from a list of candidates

> (i) maps flat sky patches back on to the sphere at the positions of the putative detections;
> (ii) applies a detection mask, if any;
> (iii) merges multiple detections of the same source obtained in different patches into a single candidate detection; and
> (iv) makes the final catalogue by rejecting those lines that do not meet the pre-established criterion of purity or loss.

- The last step is critical to the success of our methodology

- If the selection criterion is based on losses, then we just need to trim the 'proto-catalogue' further by applying the decision rule - it is much more common in astronomy to require a catalogue to have an expected contamination ratio or that the contamination does not exceed a prescribed value

- The Bayesian logic framework can give us exactly that ->

Catalog making

- The Bayesian logic framework can give us exactly that:
  - The number of false positives in a catalogue may be represented as a sum of Bernoulli variables
- Assuming all catalogue entries are statistically independent, the sum of N of those variables is distributed as a Poisson–binomial distribution:

(57) $$\mu = \sum_{i=1}^{n} p_i, \quad \sigma^2 = \sum_{i=1}^{n} p_i(1 - p_i), \qquad p_i = \Pr_i(\widetilde{H_{j*}} \mid \boldsymbol{d}).$$

p_i=the probability of source i being a false positive

(i) sort the list of candidate detections in ln (odds) descending order (p_i ascending order);
(ii) for each candidate, accumulate p_i until μ (see formula 57) exceeds the prescribed contamination $\alpha \equiv$ (spuriousdetections)/(totallinesincatalogue) times the total number of lines already included; and
(iii) discard the last line.

- CLT=> the number of spurious detections in the catalogue is $\sum_{i=1}^{N} p_i \pm \sqrt{\sum_{i=1}^{N} p_i(1 - p_i)},$
- An estimate of the fraction of spurious detections in the catalogue, α, reads: $\left(\widehat{\alpha} = \frac{\sum_{i=1}^{N} p_i}{N}\right) \pm \frac{\sqrt{\sum_{i=1}^{N} p_i(1 - p_i)}}{N}.$

Catalog making

- Finally, we are now in position to answer the key question all the frequentist methods must at some point face:

  '*what threshold should one use for accepting the candidates for inclusion in the final catalogue?'*

- The answer is just: "*the ln (odds) estimate of the last line of the final catalogue*"
  - since the initial list of putative detections was sorted in descending order of ln (odds) and all those with a higher or equal ln (odds), and only those, were selected for inclusion.

➢ Note the question is no longer relevant in our Bayesian approach, since it is an output of our catalogue-making method, rather than an input

Bayesian Inference
basic tools

- In contrast to parameter estimation problems → in model selection the evidence takes the central role and is simply the factor required to normalize the posterior:

Evidence

$$\mathcal{Z} = \int L(\Theta)\pi(\Theta)d^D\Theta,$$

Evaluation of this multidimensional Integral is a challenging numerical task – resort to sampling techniques: MCMC, Multinest, (Sivia &Skilling 2006; Feroz et al. 2009), etc. or model the posterior as a multivariate Gaussian centered at its peak(s) and apply the Laplace formula (Hobson, Bridle & Lahav 2002).

- The evidence is the expectation of the likelihood over the prior → is central to Bayesian model selection between different hypothesis H_i

- The evidence automatically implements Occam's razor:

A simpler theory with compact parameter space will have a larger evidence than a more complicated one, unless the latter is significantly better at explaining the data.

Bayesian Inference
basic tools

$$\mathcal{Z} = \int L(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})d^D\boldsymbol{\Theta},$$

The evidence is the expectation of the likelihood over the prior, and hence is central to Bayesian model selection between different hypothesis H_i

- The question of model selection between two models H_0 and H_1 can then be decided by comparing their respective posterior probabilities given the observed data set d

$$\frac{\Pr(H_1|d)}{\Pr(H_0|d)} = \frac{\Pr(d|H_1)\Pr(H_1)}{\Pr(d|H_0)\Pr(H_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\frac{\Pr(H_1)}{\Pr(H_0)},$$

where $\Pr(H_1)/\Pr(H_0)$ is the a priori probability ratio for the models

Bayesian Object Detection

Case I: Detection of Point sources in Planck maps – PWS

- *Data Model*

$$d(x) = s(x;\theta) + AY(x) + n_{inst}(x) + F(x)$$

Diffuse galactic foregrounds ignored here

CMB fluctuations →

$$\frac{\Delta T}{T} \equiv \sum_{\ell m} a_{\ell m} Y_{\ell m}(\theta, \phi)$$

$$p(a_{\ell m}) = N(a_{\ell m};0,C_\ell)$$

$$\langle a_{\ell m} a_{\ell' m'} \rangle = C_\ell \, \delta_{\ell\ell'} \delta_{mm'}$$

Statistical isotropy

Rotational invariance

$$\langle T_{i_1} T_{i_2} \rangle = \underbrace{\sum_{\ell=2}^{\ell_{\max}} \frac{2\ell+1}{4\pi} \hat{C}_\ell P_\ell(\theta_{i_1 i_2})}_{C} + \underbrace{\mathbf{N_{i_1 i_2}}}_{N_{inst}},$$

$$N_{inst} = \left\langle n_{inst} n_{inst}^T \right\rangle$$

$$N = C + N_{inst}$$

- The CMB fluctuations are a statistically homogeneous Gaussian random field → The correlation matrix is circulante hence diagonal in Fourier or Harmonic space – the spherical harmonics or Fourier basis are our PCA basis

Carvalho, Rocha & Hobson 2009; & Lasenby 2011, Rocha in prep

Standard approach
Matched Filter

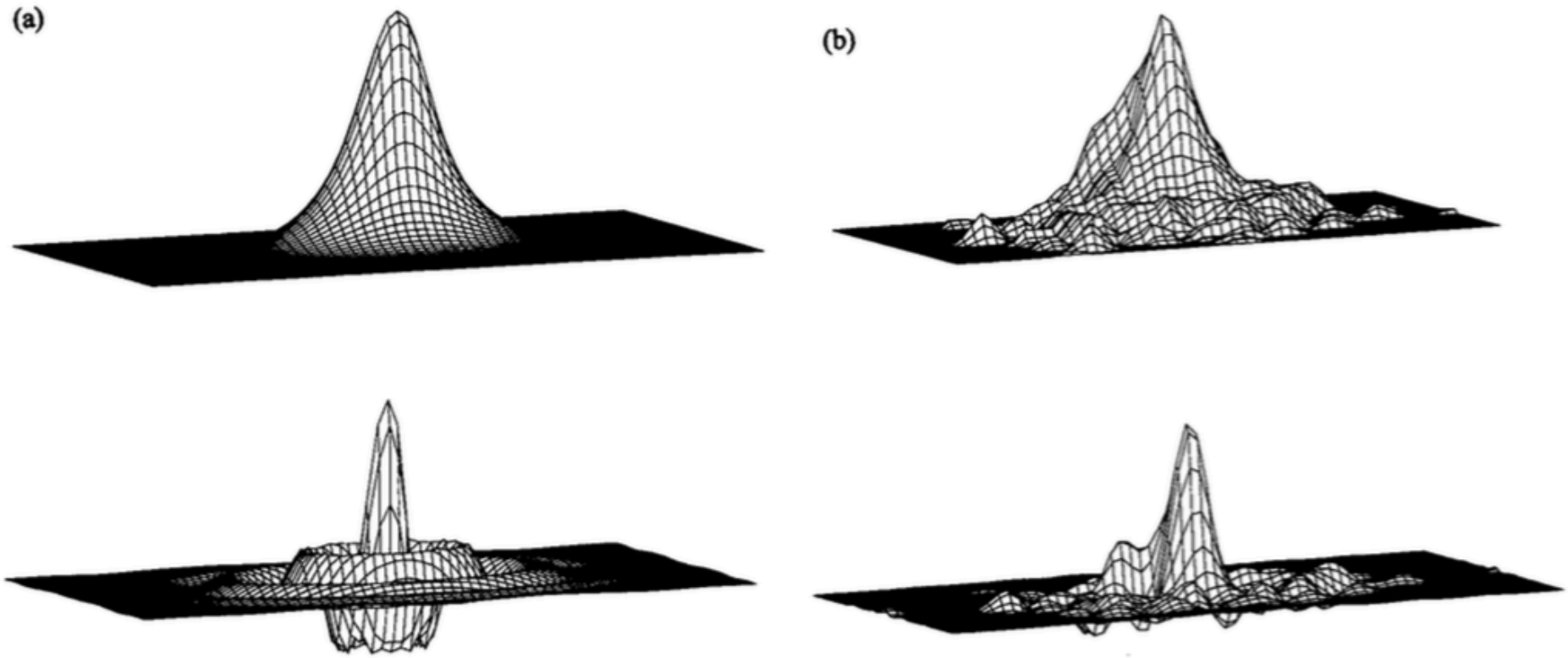


Figure 2. (a) Typical filter function for an axisymmetric cluster in arbitrary units. The cluster profile is shown for comparison. The angular scales are the same. (b) Same as (a) for a non-axisymmetric cluster. The cluster is shown at the top, the corresponding filter function at the bottom.