Detecting galaxies by marked point process in a Bayesian framework : how to control detection errors ?

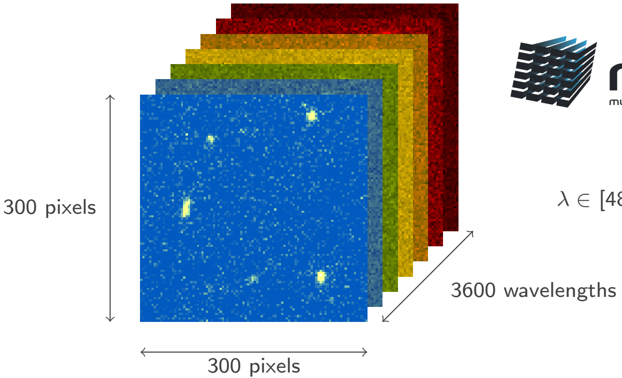
Céline Meillier

14 November 2014

gipsa-lab

PhD supervisors:
Florent CHATELAIN, Olivier MICHEL, Hacheme AYASSO

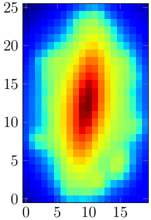
Hyperspectral data



$\lambda \in$ [480 nm, 930 nm]

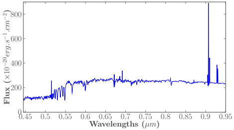
MUSE data collected by 24 3D spectrographs combined with one of the four telescopes of VLT (Chile).

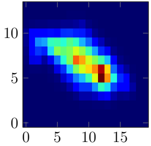
Typical observed galaxies



Bright and spread galaxy

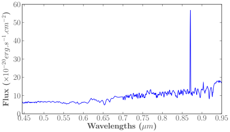
SNR ≃ 30dB

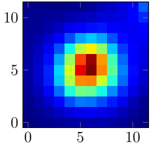




Nonregular shape galaxy

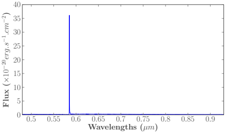
SNR ≃ 10dB





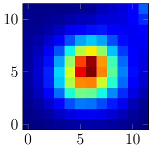
Faint and small galaxy

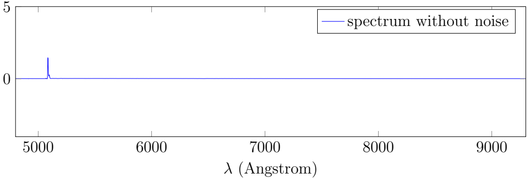
SNR ≃ -40dB



Main challenge

Objects of interest: small and faint galaxies with a spectrum composed of one emission line (Lyman-alpha emitters).





Approximation: Lyman-alpha emmiters $\simeq$ 3D point sources.

Observation: Low signal-to-noise ratio (SNR) .

Objective: Find possible positions of Lyman-alpha emitters in the 3D datacube.

Faint galaxies detection

Objective: detect and estimate objects whose position, number, shape, intensity and spectrum are unknown.

Challenges:

→ Detection of the smallest and faintest galaxies.

→ Large dynamics between galaxies intensity.

→ Control of the error

Proposed approach:

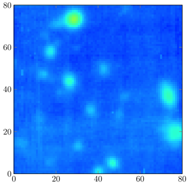
→ Galaxies configuration = a realization of a marked point process.

→ Observation model and Bayesian approach.

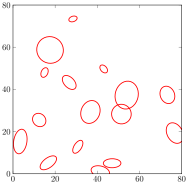
Outlines

Outlines

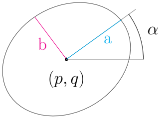
Galaxies configuration



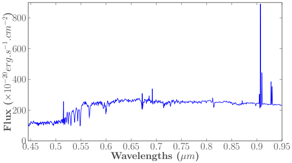
White image

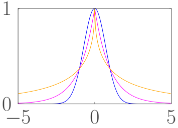


Marked point process



Spatial support





Sersic profiles

One object:

- geometrical marks
- spectral mark
- intensity mark

An object configuration = a realization of a marked point process (MPP)

Observation model

Global observation model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon} \qquad (1) \qquad \text{where} \qquad \begin{array}{lcl} \boldsymbol{Y} & = & [Y_1, \cdots, Y_\Lambda] \\ \boldsymbol{X} & = & [x_1, \cdots, x_n] \\ \boldsymbol{w} & = & [w_1, \cdots, w_n] \\ \boldsymbol{\epsilon} & = & [\epsilon_1, \cdots, \epsilon_\Lambda] \\ \Lambda & = & \text{wavelengths number} \\ n & = & \text{number of detected objects} \end{array}$$

and for all λ :

$$\boldsymbol{Y}_\lambda = \boldsymbol{X}\boldsymbol{w}_\lambda + \epsilon_\lambda$$

with:

(H1) $\epsilon_\lambda =$ vector of spatially independent Gaussian variables $\sim \mathcal{N}(m_\lambda, \sigma_\lambda^2)$.

(H2) The ϵ_λ are spectrally independent.

(H3) $\boldsymbol{X}$ includes FSF information (averaged on $\lambda \rightarrow \boldsymbol{X}$ is λ-invariant).

(H4) LSF is not directly included in the observation model.

Outlines

Bayesian approach

→ From (H2) and eq. (1) the global likelihood $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{\sigma}^2)$ of the data can be computed.

→ Priors on $\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{m}$, and $\boldsymbol{\sigma}^2$ can be added[1].

→ From Bayes approach, the joint posterior density can be written:

$$p(\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{\sigma}^2|\boldsymbol{Y}) \propto \underbrace{f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{\sigma}^2)}_{\text{data fidelity term}} \underbrace{p(\boldsymbol{m}, \boldsymbol{\sigma}^2)p(\boldsymbol{w}|\boldsymbol{X})p(\boldsymbol{X})}_{\text{priors}}$$

→ Estimation of $\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{m}$, and $\boldsymbol{\sigma}^2 \rightarrow$ maximization of the posterior density.

[1] C. Meillier et al. (2014). "Non-parametric Bayesian framework for detection of object configurations with large intensity dynamics in highly noisy hyperspectral data". In: *2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*

Detection and estimation algorithm

→ Posterior density too complex to be analytically used.

→ Reversible Jump Markov Chain Monte Carlo:

- Gibbs sampler[2] for parameters $\boldsymbol{m}$ and $\boldsymbol{\sigma^2}$.
- Metropolis-Hastings-Green sampler[3] for configuration $\boldsymbol{X}$.

→ RJMCMC : method that generates samples whose density is close to the posterior.

DATACUBE ⟶ Detection ⟶ OBJECT CONFIGURATION

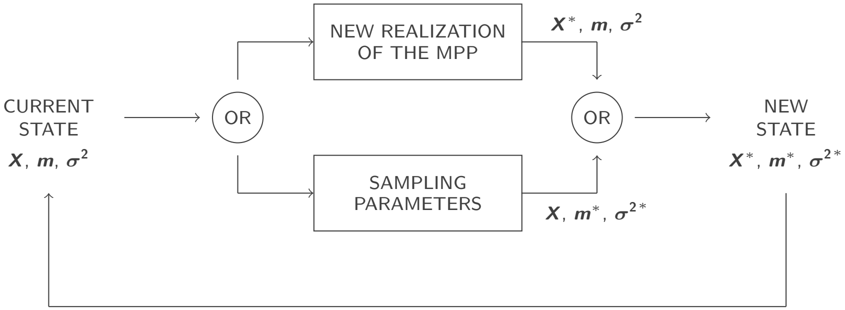
[2] S. Geman and D. Geman (1984). "Stochastic Relaxation, Gibbs Distribution and Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*

[3] P.J. Green (1995). "Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 52

Detection and estimation algorithm

→ Initialization: empty configuration, empirical mean and variance of the data.

→ At each iteration:



→ Maximum a posteriori estimation

Advantages and limitations of the methods

DATACUBE $\longrightarrow$ **Detection** $\longrightarrow$ OBJECT CONFIGURATION

+ Nonparametric method : detection and estimation.
+ Both the configuration and the background parameters are estimated.
+ The estimation is fully data-driven.

– Computational time increases in $\mathcal{O}(n^2)$.
– Errors control ?

Outlines

Problem formulation

Limitations of the detection procedure:

– Computational time increases in $\mathcal{O}(n^2)$.

– Errors control ?

Proposed solution:

→ Preprocess the data

→ Multiple hypotheses testing procedures

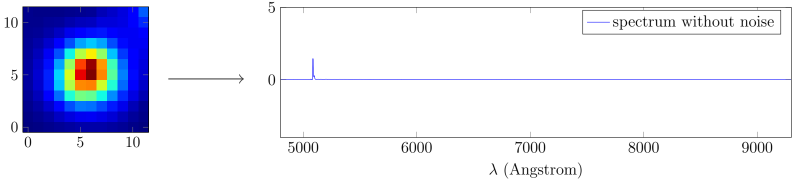
\+ Reduce the exploration space by the MPP.

\+ Introduce an error control criterion in the algorithm.

– Number of tests

Problem formulation

Objects of interest: small and faint galaxies with a spectrum composed of one emission line (Lyman-alpha emitters).



Approximation: Lyman-alpha emmiters $\simeq$ 3D point sources.

Lyman-alpha response $\rightarrow$ close to the 3D PSF.

Objective: Find possible positions of Lyman-alpha emitters in the 3D datacube.

$\rightarrow$ Multiple hypotheses testing

Multiple hypotheses testing procedures

→ $N = N_0 + N_1$ tests, with N_0 true $\mathcal{H}_0$ and N_1 true $\mathcal{H}_1$

Decision / Truth	$\widehat{\mathcal{H}_0}$	$\widehat{\mathcal{H}_1}$
$\mathcal{H}_0$	N_0 - a	a (Type I errors)
$\mathcal{H}_1$	N_1 - b	b

→ False alarms control :

$$\Pr(\widehat{\mathcal{H}_1}|\mathcal{H}_0) \leqslant \alpha \;\rightarrow\; a \simeq N \times \alpha$$

→ False detections control :

$$\frac{a}{a+b} \leqslant \alpha$$

→ Family-Wise Error Rate (**FWER**) → the probability of at least one type I error:

$$FWER = \Pr(a \geqslant 1)$$

→ False Discovery Rate (**FDR**) → expected proportion of Type I errors among the rejected hypotheses:

$$FDR = E\left(\frac{a}{a+b}\middle| a+b>0\right)\Pr\left(a+b>0\right)$$

Controlling the FDR via knockoffs filter [4]

Objective: Find relevant variables and control FDR

Linear Gaussian model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{where} \qquad \begin{array}{ccc} \boldsymbol{Y} & \in & \mathbb{R}^{n} \\ \boldsymbol{X} & \in & \mathbb{R}^{n \times p} \\ \boldsymbol{\beta} & \in & \mathbb{R}^{p} \\ \boldsymbol{\epsilon} & \sim & \mathcal{N}(0, \mathbf{I}_n) \end{array}$$

Knockoffs filter:

1. Construct the knockoff $\widetilde{X}_j$ for each feature X_j such as:
   - $\widetilde{X}^T \widetilde{X} = X^T X$
   - $\widetilde{X}_j^T X_k = X_j^T X_k$ for all $j \neq k$.
2. Calculate statistics for each pair $(X_j, \widetilde{X}_j)$
3. Calculate data-dependant threshold for the statistics

Control of the FDR

[4] Rina Foygel Barber and Emmanuel Candes (2014). "Controlling the False Discovery Rate via Knockoffs". In: *arXiv preprint arXiv:1404.5609*

Controlling the FDR via knockoffs filter - Application to the galaxies detection

Objective: Find possible positions of Lyman-alpha emitters in the 3D datacube.

Variables: Each position (x, y, λ) should be tested, $X_j = PSF_{x,y,\lambda}$.

Limitations of the knockoffs filter on the MUSE data:

- Dimensions $p = n = 360 \times 10^6$.
- Building the knockoffs (respecting the 3D correlations).

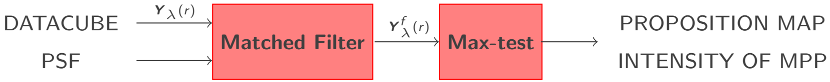
Max-test

→ **Highlight the small galaxies** with a spectrum composed of a few distinct emission lines.

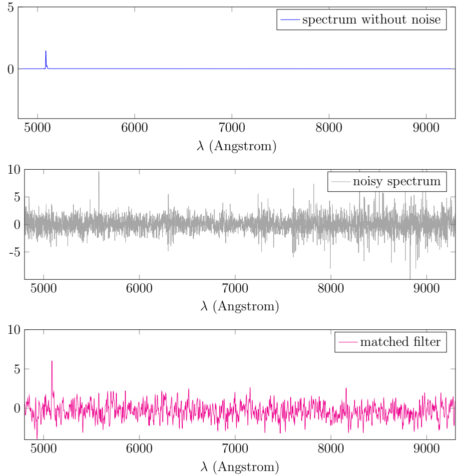
- ↪ Their response should be close to the PSF.
- ↪ Matched filter with the 3D PSF.

→ Define the intensity of the marked point process

Structure of the preprocessing step:



Max-test



Binary hypothesis test:

$$\begin{cases} \mathcal{H}_0: \text{ noise only} \\ \mathcal{H}_1: \text{ presence of an object} \end{cases}$$

Max-test:

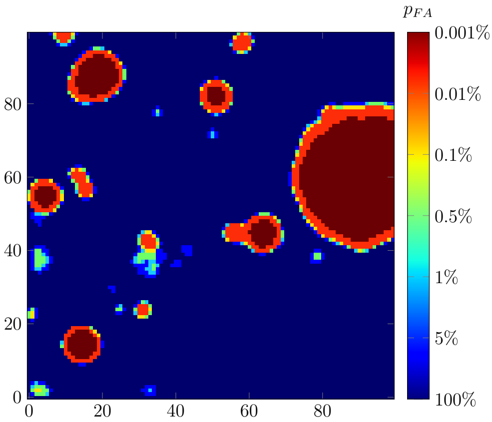
$$\max_{\lambda}(\boldsymbol{Y}^f_{\lambda}(r)) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \eta(p_{FA}),$$

Max-test statistics known under $\mathcal{H}_0$ (Monte Carlo simulations).

Note: this test can also be obtained by writing the GLRT[5].

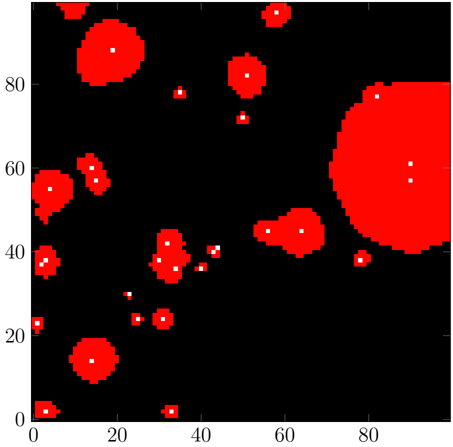
[5] Silvia Paris et al. (2013). "Constrained likelihood ratios for detecting sparse signals in highly noisy 3D data". In: *International Conference on Acoustics, Speech and Signal Processing*

Max-test



Segmentation of the 2D image for different false alarm probabilities (p_{FA})

Max-test



Segmentation of the 2D image for a given false alarm probabilities (p_{FA})

→ Proposition map

Higher Criticism[5]

→ For $i = 1, \ldots, n$ consider n independent tests:

$$\begin{cases} \mathcal{H}_{0,i}: & X_i \sim \mathcal{N}(0,1) \\ \mathcal{H}_{1,i}: & X_i \sim \mathcal{N}(\mu_i, 1) \qquad \text{with} \quad \mu_i > 0 \end{cases}$$

with a small proportion ϵ of the X_i such as $\mu_i > 0$.

↪ Asymptotically optimal

↪ **MUSE application:** model adapted to the Lyman alpha emitters.

→ Let $p_{(1)} \leqslant \ldots \leqslant p_N$ be the n sorted p-values and compute the HC^* statistic:

$$HC^* = \max_{0 < i \leqslant \alpha_0 \times n} \frac{\sqrt{n}\left(\frac{i}{n} - p_{(i)}\right)}{\sqrt{p_{(i)}(1 - p_{(i)})}}$$

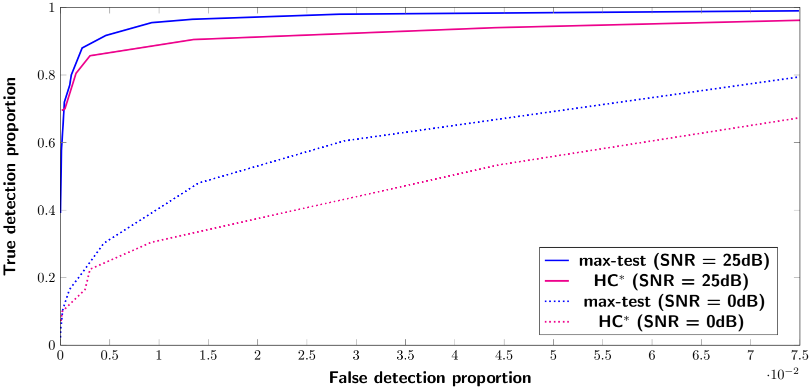
→ Reject $\mathcal{H}_{0,(1)}$, ..., $\mathcal{H}_{0,(i_{max}-1)}$

Application on the matched filtered result → **How to control false alarms for dependent tests ?**

[5] David Donoho and Jiashun Jin (2004). "Higher Criticism for detecting sparse heterogeneous mixtures". In: *Annals of Statistics*

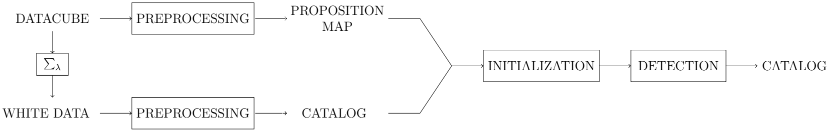
Performances

Detection performances on synthetic images:



Outlines

Application to DATACUBE-HDFS-v031c



Outlines

Conclusion and perspective

→ Nonparametric method for galaxy detection

→ Preprocessing step for error control:

- Pixel-wise false alarm control
- FDR under propoerty of positive regression dependancy on a subset I_0[6] → matched filtered data PRDS.

→ Good results on the real data

Future work:

→ Analyze the objects which are not in the HDFS-catalog to identify potential new discoveries.

→ Empiric FDR control in the catalog produced by the method.

[6] Yoav Benjamini and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency". In: *Annals of statistics*