Radio Surveys Data Analysis in the Visibility Domain

Marzia Rivi

in collaboration with F. Abdalla, S. Balan, I. Harrison, M. Lochner, S. Makhathini, A. Malyali, J. McEwen, L. Miller, I. Prandoni

SKA Data Challenges Workshop
Bologna, Sept. 30 - Oct. 2, 2019

INAF
ISTITUTO NAZIONALE DI ASTROFISICA
NATIONAL INSTITUTE FOR ASTROPHYSICS

Outline

- Galaxy catalogs from radio data analysis: image vs visibilities
- Visibility model
- Bayesian methods in the visibility domain
  - Galaxy shape measurement (for radio weak lensing)
  - Galaxy detection
- Conclusions

Data analysis: image vs visibilities

SKA extragalactic continuum surveys will allow **new scientific measurements** that will require more and more **accuracy in galaxy catalogs**.

Radio image is obtained from processing of the original data originated in the Fourier space:

- noise is highly correlated
- systematics introduced by the imaging process (due to iterative deconvolution procedure) may become too large

e.g. Radio Weak Lensing (RWL) shear measurement gives poor results w.r.t. requirements:

Patel+ 2015

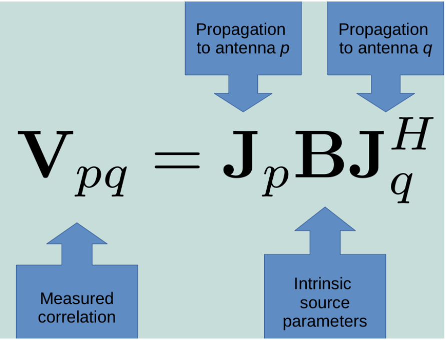
	Multiplicative Bias	Additive Bias
SKA1 requirement	0.0067	0.00082
CLEAN images	-0.265 ± 0.02	0.001 ± 0.005

Visibilities analysis is a more natural approach BUT:

- Computationally demanding (big data)
- Sources cannot be easily isolated
- Model fitting for source characterization

Radio Interferometry Measurement Equation (RIME)

Description of what happens to the radio signal from the source to the interferometer (Hamaker+ 1996, Hamaker 2000, Smirnov 2011):



Multiple propagation effects can be added by chaining up Jones matrices (e.g. antenna gains, antenna primary beam, ...):

$$\mathbf{V}_{pq} = \mathbf{J}_{p,n}(\dots(\mathbf{J}_{p,1}\mathbf{B}\mathbf{J}_{q,1}^H)\dots)\mathbf{J}_{q,n}^H$$

Galaxy Radio Surface Brightness Model

star-forming galaxies

- synchrotron radiation emitted by the interstellar medium in the *disc alone*
- Fourier Transform of optical disc model (*Sérsic profile index* 1)
- can be computed analitically! (Rivi+ 2016)

$$V(u,v) = \mathcal{F}(I \circ A)(\mathbf{k}), \quad I(r) = I_0 \exp(-r/\alpha), \quad A\mathbf{x} = \begin{pmatrix} 1-e_1 & -e_2 \\ -e_2 & 1+e_1 \end{pmatrix} \begin{pmatrix} l \\ m \end{pmatrix}$$

radio-quiet AGN

- compact radio emission within host galaxy (Guidetti+ 2017)
- 2-components model?

radio-loud AGN

- jets + lobes components
- *shapelets* model (invariant by Fourier Transform) for lobes? (Chang+ 2004)

SF Galaxy Shape Measurement for RWL

Two Bayesian approaches given the *sky catalog* (source position and integrated flux) and *calibrated visibilities*:

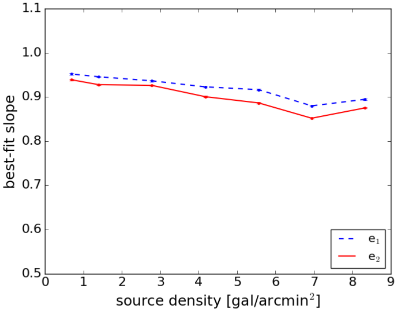
- **Single-source model**: *RadioLensfit* (Rivi & Miller 2018)
  - source extraction (sky model + faceting)
  - chi-square *fitting of a single source at a time* marginalising over position, flux and size
    $\rightarrow L = L(e_1, e_2)$, i.e. parameter space of dimension 2
  - likelihood sampling: ML + adaptive grid around the maximum

- **Multi-source model**: *BIRO - Hamiltonian Monte Carlo* (Rivi+ 2019)
  - *Joint fitting* for $e_{1,s}, e_{2,s}, \alpha_s$ parameters of all N sources in the FoV
    $\rightarrow$ parameter space of dimension $3N$
  - likelihood sampling: HMC with step size dependent on source flux and analytic likelihood gradient

SF Galaxy Shape Measurement for RWL

Simulation SKA1-MID 8 hrs, $t_{acc} = 60$ s, 1 channel at 1.07 GHz → 9,266,880 visibilities
Realistic distribution of SF galaxies with SNR $\geqslant 10$

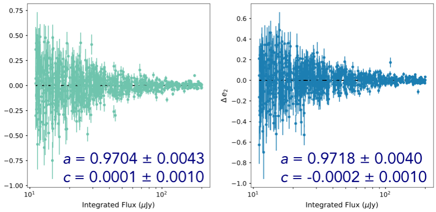
RadioLensfit (10^4 sources)



At SKA1-MID expected source density (2.7 gal/arcmin^2):

$a_1 = 0.9365 \pm 0.0017$
$a_2 = 0.9262 \pm 0.0017$

BIRO-HMC (1000 sources, 2.7 gal/arcmin^2)



Improved shape measurement **accuracy** but **computationally much more demanding**

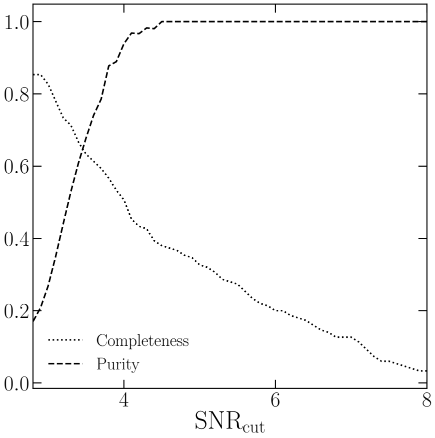
size best-fit line:
$a = 1.0048 \pm 0.0030$
$c = -0.0090 \pm 0.0051$

This two approaches can be combined to accelerate HMC convergence

Galaxy Detection in the Visibility Domain

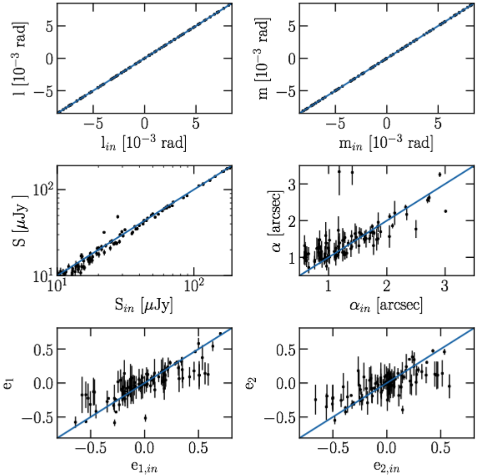
GalNest (Malyali, Rivi, Abdalla, McEwen, 2019)

- **single source model** + multimodal posterior sampler (*MultiNest*, Feroz+ 2009)
- **clustering** algorithm (SCIKIT-LEARN *mean shift*) to identify the source from clustered fake modes
- **SNR threshold** to remove remaining fake modes
- from SKA1-MID simulations of SF galaxies observation, **reliable source detection down to SNR $\sim$ 5**



Galaxy Detection: SKA1-MID simulation at 1.07 GHz

8 hrs integration time, $t_{\rm acc} = 60$ s, 1 channel → about 10^7 visibilities
realistic distribution of SF galaxies with **SNR $\geqslant$ 10**

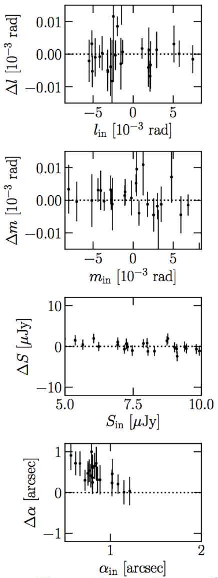


98/100 galaxy detections

Galaxy Detection: SKA1-MID simulation at 1.07 GHz

8 hrs integration time, $t_{\mathrm{acc}} = 60$ s,
single channel
realistic distribution of 50 SF galaxies with
SNR ranging 3 - 13

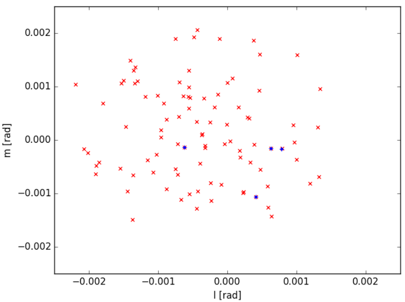
- NO fake modes at SNR $\geqslant 4.5$
- 58% detections of the population with SNR $\geqslant 5$
- still accurate position and flux measurements

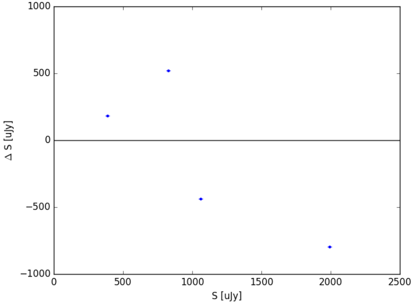


Galaxy Detection: JVLA Observations

7 pointings of the GOODS-N Field, VLA A-configuration 14 hrs, B-configuration 2.5 hrs
$t_{acc} = 1$ s, 16 adjiacent 64×2 MHz channels at a central frequency 5.5 GHz

Image Catalog: 94 sources with SNR $\geqslant 5$ (80% are AGN) Guidetti+ 2017





Using a **reduced dataset** (single pointing and single spectral window):

- all the 4 visible (brightest) sources are detected!
- flux discrepancy: source model and signal contamination, primary beam

work in progress!

SKA Data Challenge

Phase 1 SKA-MID Medium-Deep Band 2 Survey:
5000 $\deg^2$ to a depth of 2μJy RMS (10,000 hrs, $z < 0.4$)

SKA Cosmology SWG, Red Book 2018

Continuum **radio weak lensing survey requirements**:
$\sim 10^4$ pointings of ~ 1 hour each ($\Delta t = 0.5$ s sampling),
~ 6000 frequency channels at a resolution of $\Delta\nu = 50$ kHz,
necessary resolution for smearing-induced ellipticity to be acceptable.
Very large data volume for a continuum survey (**order of PBytes** per pointing)
but directly **comparable to that expected by HI line galaxy surveys**.

SKA ECP150007 v2, Brown & Harrison 2015

Possible solutions:
- *work on gridded visibilities*
- *dedicated RWL pipeline at the SKA SDP?*

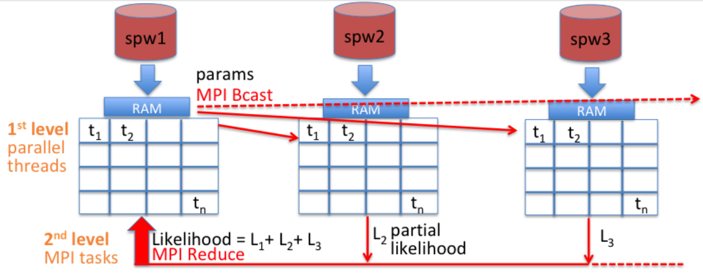
Parallelization strategy for analysis and simulation tools

The most computing intensive part is the **visibility model computation:** visibility function must be evaluated

- at each point (baseline, frequency, time) of the uv coverage
- at each iteration of the parameter space sampler

Massively embarassing parallelization along with uv points
⇒ exploitation of both multi-core CPUs and GPUs.

Large datasets may be split by spectral windows to be distributed on different computing nodes: **hierarchical parallelization**



Conclusions

- SKA will allow **new scientific measurements in the radio band**, e.g. *radio weak lensing*, for which analysis from radio images may not be accurate enough.
- **Methods in the visibility domain** may be more accurate but are computationally very challenging because of the big data.
- **Bayesian methods** in the visibility domain for **SF galaxy shape measurements** allow to reduce noise bias.
  - *RadioLensfit* working well for SKA1 source density and it is very fast
  - *HMC* more accurate but much slower for large number of sources
  - The two approaches may be combined for higher source density regions
- **Radio galaxy detection** in the visibility domain is possible, e.g. with a *Multimodal Nested Sampling* approach.
- All these approaches require **code parallelization** to exploit clusters for SKA data processing (and/or simulation).