Radio Surveys Data Analysis in the Visibility Domain

Marzia Rivi
marzia.rivi@inaf.it

INAF Institute of Radioastronomy

INAF Science Archives & the Big Data Challenge
Rome, June 18, 2019

Outline

- Radio data
- Use case: Radio Weak Lensing Surveys
- Image vs Visibilities
- HPC analysis methods in the Visibility Domain
- Data Challenge → RWL pipeline
- Conclusions

Radio Data

Radio interferometers produce **complex visibilities** sampled at discrete locations in the Fourier domain (*uv* plane)
sampling function (uv coverage): $S(u,v) = \sum_i \delta(u-u_i, v-v_i)$

complex raw visibilities

$$V(u,v) \sim \int\int I(l,m)e^{-2\pi i(ul+vm)}dldm$$

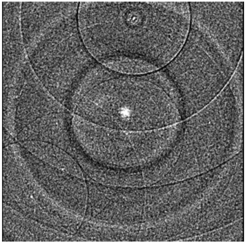
↓ calibration, gridding, FFT

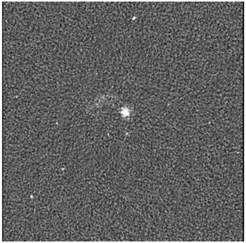
dirty image

$$I^D = \int\int S(u,v)V(u,v)e^{2\pi i(ul+vm)}dudv = I * PSF$$

↓ iterative PSF deconvolution

clean image





Radio Data

New generation of radio interferometers, such as SKA, provide:

- high resolution (antennas distributed over a very large area)
- high sensitivity

for radio continuum surveys

MEANING

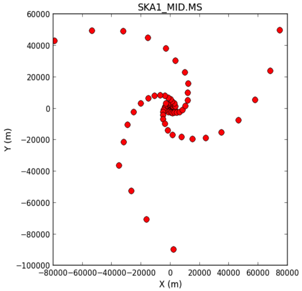
- observation of **large number density** of faint **extended sources**
- **big data volume** not supported by current tools
- traditional imaging algorithms cannot be applied anymore

Square Kilometre Array (SKA) Mid-Frequency

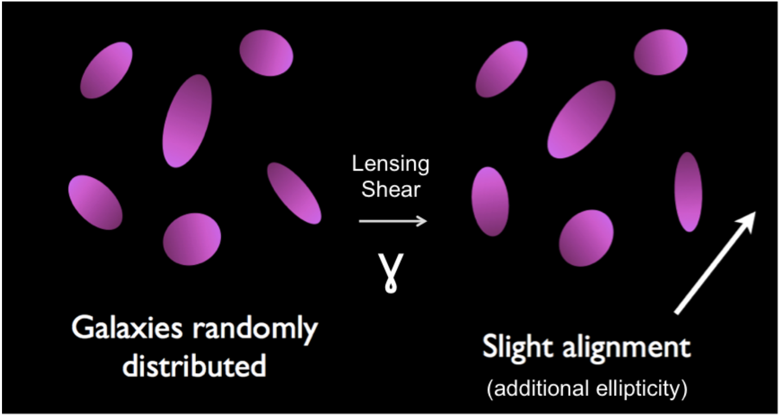
SKA-MID South Africa 350 MHz - 14 GHz

Phase 1: 64 MeerKAT 13.5m + 133 SKA 15m dishes
max baseline 150 km, resolution 0.3 arcsec
$\sim$ 3 gal/arcmin^2 $\rightarrow$ 10^4 sources in 1 deg^2 FoV!

Phase 2: $\sim$ 2000 dishes, $\sim$ 10 gal/arcmin^2



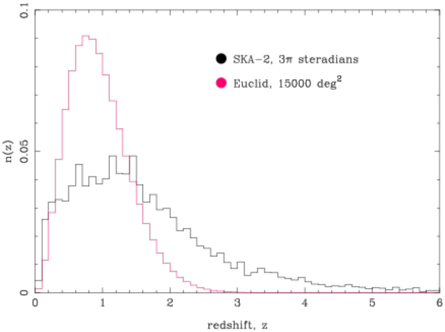
Use Case: Radio Weak Lensing



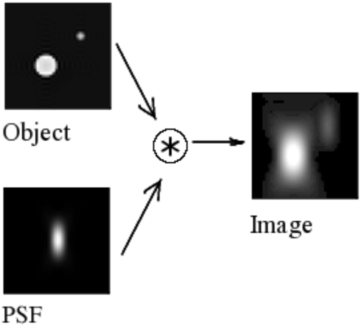
For a large sample of background galaxies: $\gamma \sim \langle \mathbf{e} \rangle$.

Use Case: Radio Weak Lensing

- **Higher redshift** source distribution (beyond LSST and Euclid).
- Redshift measurement from detection of HI 21 cm lines.



M. Brown et al., SKA Science Book Vol.2, 2015

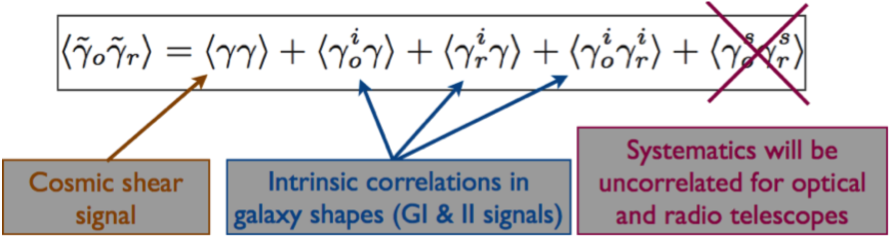


- **Well-known** and **deterministic knowledge of PSF** (from antennae positions) solves one of the biggest systematic errors.

Use Case: Radio Weak Lensing

- Other measurements may provide estimate of galaxy's intrinsic orientation allowing **mitigation of intrinsic alignments**:
  - *radio polarization* (Brown & Battye 2011, Thomas et al 2017)
  - *HI rotational velocity* (Morales 2006, Huff et al 2013)
- **Cross-correlation** of shear estimators of optical and radio surveys drops out wavelength dependent systematics. (Demetroullas & Brown, 2016, 2018)

$$\tilde{\gamma} = \gamma + \gamma^i + \gamma^s$$



Galaxy Shape Measurement in the Radio Band

At $\sim$ 1GHz faint SF galaxies flux densities should be dominated by synchrotron radiation emitted by the ISM in the disc alone.
Measurement approach: **galaxy disc model fitting**

- Image: use optical methods corrected for galaxy and PSF models
  - Noise correlated.
  - Iterative deconvolution imaging procedure gives poor results w.r.t. requirements. Patel et al. (2015)

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

- Visibility: adapt optical methods to the Fourier space or development of new ones.
  - Natural approach, data not yet affected by systematics introduced by the imaging process.
  - So far the only ones to successfully detect radio weak lensing signal. Chang et al. (2004)
  - Computationally demanding and sources not localised.

HPC Bayesian Methods in the Visibility Domain

Galaxy visibility model: analytical FT of the exponential profile

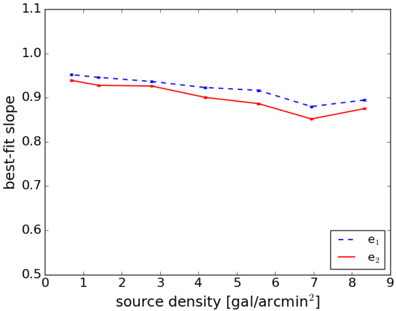
Two Bayesian approaches given the sky *source catalog* (position and integrated flux) and *calibrated visibilities*:

- **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
  - likelihood sampling (ML + adaptive grid around the maximum)
  - Multi-threaded code (C+OpenMP), working on MPI extension
- **Multi-source model**: *BIRO - Hamiltonian Monte Carlo* (Rivi et al. 2019)
  - Joint fitting of all sources in the field of view (*ellipticity and size*!)
  - Analytical likelihood gradient computation for HMC sampler
  - GPU-accelerated tool for model and likelihood computation: *Montblanc* (Perkins et al. 2015)

Performance

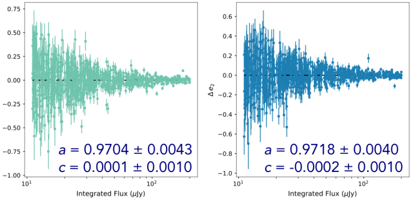
Simulation SKA1-MID 8 hrs, $t_{\rm acc} = 60$ s, 1 channel at 1.07 GHz → 9,266,880 visibilities
Realistic distribution of sources with SNR $\geqslant 10$

RadioLensfit



$a_1 = 0.9365 \pm 0.0017$
$a_2 = 0.9262 \pm 0.0017$ (at 2.7 gal/arcmin^2)
10^4 sources
8 cores Intel Xeon E5-2650
computing time per source: 2.4 sec/gal
(independent of number of sources)

BIRO-HMC



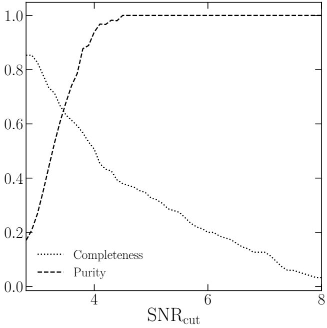
2.7 gal/arcmin^2
1,000 sources
1 core Intel Xeon E5-2650
+ 2 NVIDIA Tesla K40
computing time per source: 4.64 min/gal
(dependent on number of sources,
the size is also fitted)

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

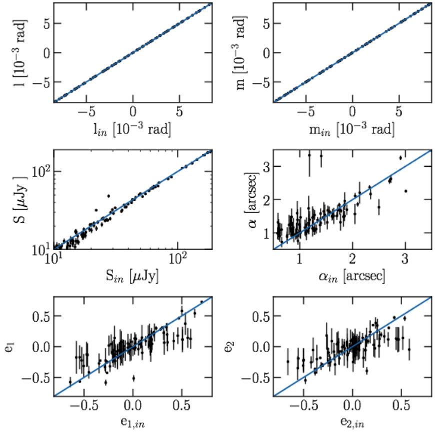
Source Detection in the Visibility Domain

GalNest (Malyali, Rivi, Abdalla, McEwen 2019)

- single model + multimodal posterior sampler (MultiNest)
- clustering algorithm (*mean shift*) to identify the source from clustered fake modes
- MPI + GPU-accelerated model and likelihood computation (Montblanc)



SKA1-MID simulation at 1.07 GHz



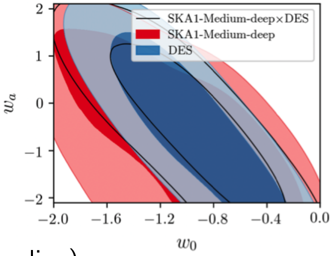
98/100 galaxy detections at SNR $\geqslant 10$
reliable source detection down to SNR ~ 5

Data Challenge

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

Very small signal at cosmological scales
Cosmic shear requirements:
$\gamma^{obs} - \gamma^{true} = m\gamma^{true} + c$

- *multiplicative* bias: $m < 6.4 \times 10^{-3}$
- *additive* bias: $c < 8.0 \times 10^{-4}$



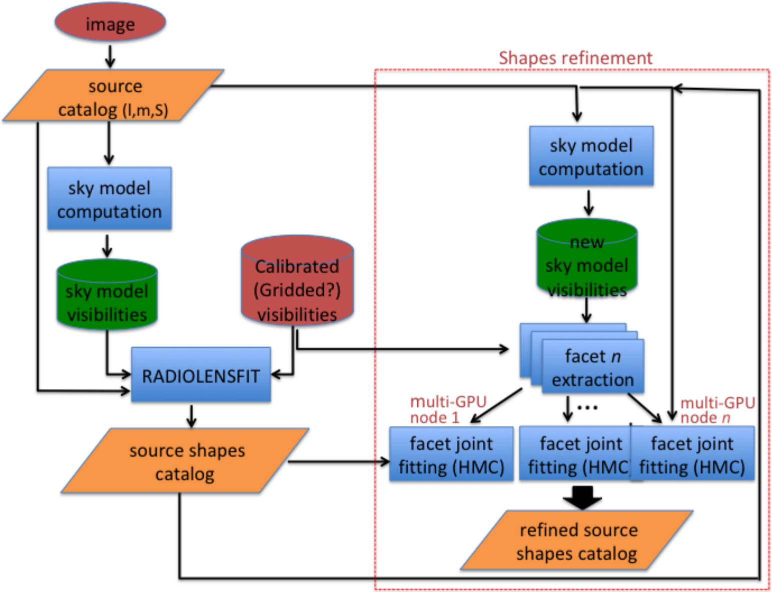
$\sim 10^4$ pointings of $\sim$ 1 hour each ($\Delta t = 0.5$ s sampling),
$\sim$ 6000 frequency channels at a resolution of $\Delta\nu = 50$ kHz,
necessary resolution for smearing-induced ellipticity to be acceptable.
Very large **data volume** for a continuum survey (**order of PBytes** per pointing)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

OR

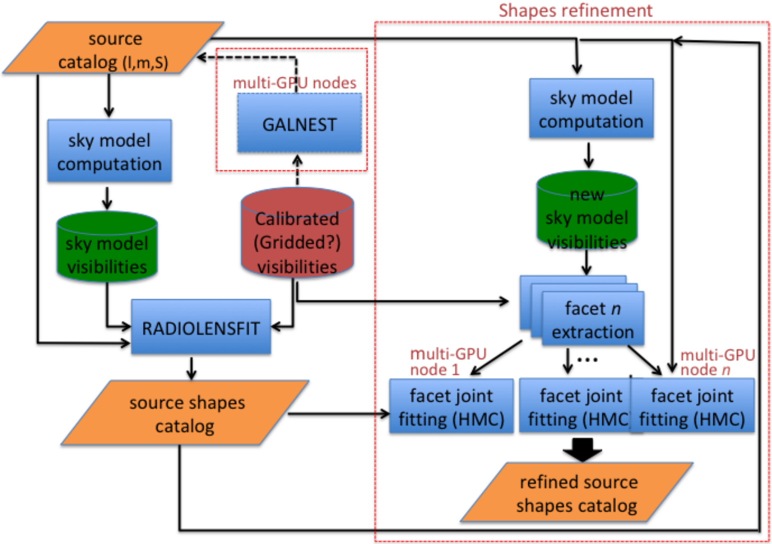
Dedicated WL data reduction pipeline to **gridded visibilities** at the SDP so that operations effects on the source morphology are known.

(SKA ECP150007 v2, Brown & Harrison 2018)

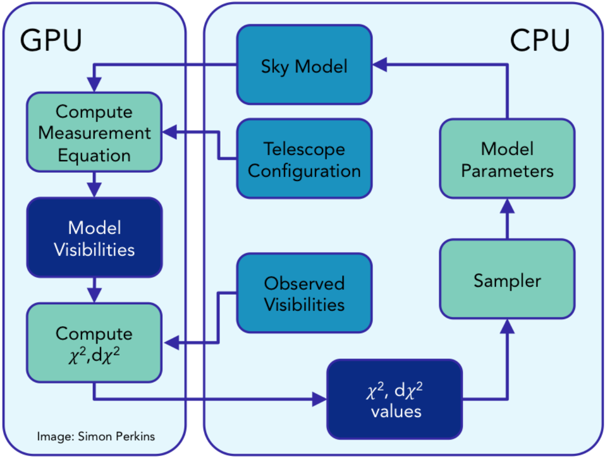
Radio Weak Lensing Pipeline (visibility+image)



Radio Weak Lensing Pipeline (visibility only)



Bayesian Inference of Radio Observations (BIRO)



MONTBLANC simulation tool.
https://github.com/ska-sa/montblanc

Fast data transfer: IBM NVLink + NVIDIA Volta GPUs

Conclusions

- High sensitivity and resolution of the new generation of radio telescopes allow **measurement of cosmological weak lensing signal in the radio band**
- **Radio images may not be accurate enough** for galaxy shape measurements
- **Methods in the visibility domain** are so far the only ones to successfully detect radio weak lensing but **are computationally very challenging because of the big data.**
- **HPC Bayesian methods** in the visibility domain can use more accurate SF galaxy models and reduce shear bias
  - *RadioLensfit* working well for SKA1 source density and is very fast
  - *HMC* more accurate but much slower for large number of sources, even using GPUs
  - The two approaches may be combined for higher source density regions.
- RWL may need of a dedicated **pipeline to run where the data is stored**