Recovery of 21 cm intensity maps with sparse component separation

Isabella P. Carucci, Melis O. Irfan, Jérôme Bobin

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **499,** 304–319 (2020)
Advance Access publication 2020 September 19

doi:10.1093/mnras/staa2854

Recovery of 21-cm intensity maps with sparse component separation

Isabella P. Carucci ,★ Melis O. Irfan and Jérôme Bobin

AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France

Accepted 2020 September 15. Received 2020 September 15; in original form 2020 June 17

ABSTRACT

21-cm intensity mapping has emerged as a promising technique to map the large-scale structure of the Universe. However, the presence of foregrounds with amplitudes orders of magnitude larger than the cosmological signal constitutes a critical challenge. Here, we test the sparsity-based algorithm generalized morphological component analysis (GMCA) as a blind component separation technique for this class of experiments. We test the GMCA performance against realistic full-sky mock temperature maps that include, besides astrophysical foregrounds, also a fraction of the polarized part of the signal leaked into the unpolarized one, a very troublesome foreground to subtract, usually referred to as polarization leakage. To our knowledge, this is the first time the removal of such component is performed with no prior assumption. We assess the success of the cleaning by comparing the true and recovered power spectra, in the angular and radial directions. In the best scenario looked at, GMCA is able to recover the input angular (radial) power spectrum with an average bias of ∼5 per cent for $\ell > 25$ (20–30 per cent for $k_\parallel \gtrsim 0.02\,h^{-1}$ Mpc), in the presence of polarization leakage. Our results are robust also when up to 40 per cent of channels are missing, mimicking a radio-frequency interference (RFI) flagging of the data. Having quantified the notable effect of polarization leakage on our results, in perspective we advocate the use of more realistic simulations when testing 21-cm intensity mapping capabilities.

Key words: methods: data analysis – methods: statistical – large-scale structure of Universe – cosmology: observations – radio lines: galaxies – radio lines: ISM.

1 INTRODUCTION

If we would ask a large-scale structure scientist about her ideal survey, she would request big cosmological volumes and great redshift resolution. Both things are hard to achieve at the same time. For instance, if we consider galaxy surveys, those are either photometric (big volumes but also big redshift errors) or spectroscopic (accurate redshifts but small volumes). This motivates the development of 21-cm intensity mapping experiments that can ensure both advantages.

Indeed, the 21-cm – alternatively, the frequency $\nu_{21\,\mathrm{cm}} = 1420$ MHz – line is emitted by the hyperfine transition of neutral hydrogen, H I. Being spectrally isolated, we are confident we are observing a H I cloud at redshift z when detecting a signal at frequency $\nu = \nu_{21\,\mathrm{cm}}/(1+z)$. Hydrogen is the most abundant baryonic component of the Universe, however, its 21-cm line is weak and long integration times are necessary to detect galaxies beyond $z \gtrsim 0.1$ (e.g. Fernández et al. 2016). To overcome this, we can use the intensity mapping technique: we drop the idea of resolving individual galaxies and instead collect all their integrated emission, scanning the sky fast and economically. This way, we tomographically assemble temperature maps in 21 cm of the Universe, effectively mapping the cosmic web in three dimensions (Battye, Davies & Weller 2004; Chang et al. 2008; Loeb & Wyithe 2008).

However, since its first application in cross-correlation with galaxies by Chang et al. (2010) with Green Bank Telescope (GBT) data, 21-cm intensity mapping has proven to be hard to be performed. There have been updates with GBT data (Masui et al. 2013; Switzer et al. 2013; Wolz et al. 2017) and more recently also with the Parkes radio telescope, although still in cross-correlation with galaxies (Anderson et al. 2018). We still miss a truly independent detection.

The main challenge for these experiments is constituted by contaminants: foregrounds of astrophysical origin – that are orders of magnitude more intense than the sought-after signal – and those originated by instrumental issues, as systematics and calibration-driven effects; the latter can also mix the modes (of foregrounds and signal), making even more challenging the component separation (Switzer et al. 2015). The discussion about more adapted and optimized cleaning methods motivates this paper.

Many of the foreground cleaning methods tested in the literature make use of the expected smoothness in frequency of the astrophysical foregrounds. Some of them parametrize the foregrounds in order to separate them (Ansari et al. 2012; Shaw et al. 2014), others do not assume a specific model for the foregrounds and are said to be blind – principal component analysis (Alonso et al. 2015; Bigot-Sazy et al. 2015); independent component analysis (Wolz et al. 2014; Zhang et al. 2016; Cunnington et al. 2019); inverse variance (Liu & Tegmark 2011); quadratic estimators (Switzer et al. 2015); and generalized needlet internal linear combination (Olivari, Remazeilles & Dickinson 2016). Up to now, in real data analysis, only blind methods have been employed and proven to be suitable for the foreground cleaning task (Masui et al. 2013; Switzer et al. 2013; Wolz et al. 2017; Anderson et al. 2018).

All the methods and works mentioned above succeed at cleaning the maps with different levels of accuracy. However, none of them include and search for components that are not smooth in frequency.

★ E-mail: ipcarucci@gmail.com

In this paper, we upgrade the degree of complexity of the simulated data we want to clean, including a non-smooth component that we expect to manifest in these observations: polarization leakage, a fraction of the polarized part of the signal that spills into the total intensity one. To our knowledge, this is the first time the removal of such kind of contaminant is attempted assuming no prior knowledge about it. We do so using the generalized morphological component analysis (GMCA) algorithm (Bobin et al. 2007). It is also the first time GMCA is adapted and tested as a blind source separation method for $z < 6$ H I intensity mapping data. Different versions of GMCA have already been applied to various observational data sets, as cosmic microwave background data (e.g. Bobin et al. 2014), 21-cm interferometric data in the Epoch of Reionization (EoR) context (Patil et al. 2017), and X-ray images of supernova remnants (Picquenot et al. 2019).

A 21-cm intensity mapping survey can be performed either in single-dish mode (one or more single-dish antennas used as a set of autocorrelators) or in standard interferometry (Bull et al. 2015); current and planned surveys exist for both regimes. Here we choose to focus on a survey like MeerKAT Large Area Synoptic Survey (MeerKLASS; Santos et al. 2017), a proposed single-dish survey with the MeerKAT radio telescope. Nevertheless, the results of this paper could be extended to other instrumental configurations [as e.g. the Baryon acoustic oscillations In Neutral Gas Observations (BINGO)[1] and Five-hundred-meter Aperture Spherical radio Telescope (FAST)[2]], also in interferometry as we will point out.

The paper is organized as follows. In Section 2, we formalize the problem we are tackling and we present the GMCA assumptions and method. In Section 3, we describe the simulation we use for testing GMCA. In Section 4, we present how we apply GMCA on the simulated data and how we evaluate the outcome of the foreground removal. In Section 5, we describe and discuss the obtained results. Finally, we summarize our work in Section 6.

2 SOURCE SEPARATION FORMALISM

2.1 The 21-cm intensity mapping context

An intensity mapping survey scans the sky and for each channel of frequency ν compiles a map of the total brightness temperature T. For each given position on the sky (each pixel p) T is the sum of the cosmological 21-cm signal from H I, of the foregrounds and of the instrumental noise:

$$T(\nu, p) = T_{\rm C}(\nu, p) + T_{\rm F}(\nu, p) + T_{\rm N}(\nu, p). \tag{1}$$

In the source separation process, we think of the foreground contribution $T_{\rm F}$ as a sum of $n_{\rm s}$ sources modulated by a frequency-dependent amplitude, i.e. for each map at ν:

$$T_{\rm F}(\nu, p) = \sum_{i=1}^{n_{\rm s}} A_i(\nu) S_i(p). \tag{2}$$

We compress all maps in a data cube $\mathbf{X}$, i.e. a $n_{\rm pix} \times n_{\rm ch}$ matrix with $n_{\rm pix}$ the number of pixels in each map and $n_{\rm ch}$ the number of maps (channels). We merge equations (1) and (2) and we can write in matrix form:

$$\mathbf{X} = \mathbf{AS} + \mathbf{C} + \mathbf{N}, \tag{3}$$

where $\mathbf{A}$ is the mixing matrix governing the contribution of the $n_{\rm s}$ sources $\mathbf{S}$ in the resulting signal, up to the cosmological signal $\mathbf{C}$ and the noise contribution $\mathbf{N}$. It follows that $\mathbf{A}$ has $n_{\rm s} \times n_{\rm ch}$, while $\mathbf{S}$ has $n_{\rm pix} \times n_{\rm s}$ dimensions.

We recall that the cosmological 21-cm signal is (i) highly outweighed by the foregrounds and (ii) uncorrelated in frequency. This implies that the cosmological signal $\mathbf{C}$ is inherently coupled to the instrumental noise component $\mathbf{N}$. The problem of foreground removal reduces to estimate the foreground-driven $\mathbf{AS}$ so that $\mathbf{X} - \mathbf{AS}$ is as accurate as possible at predicting the cosmological H I brightness temperature field, taking into account the instrumental noise contribution.

2.2 Generalized morphological component analysis

GMCA is a blind source separation algorithm that relies on the morphological features that compose the sought-after components. To that purpose, such components are assumed to admit a sparse distribution in an adapted signal representation (e.g. Fourier, wavelets, to only name two). A source is sparse when most of its coefficients are zero, thus sparse sources are easier to disentangle as their signatures are uncorrelated. A classic example is Fourier space for periodic signals: they can be described by few coefficients. The sparsity assumption is essential as it allows to dramatically improve the contrast between distinct components, which ease the separation process.

For the science case of this paper, we make use of the starlet wavelet dictionary (Starck, Fadili & Murtagh 2007) that has proven to be well adapted for an efficient sparse description of galactic diffuse emissions and astrophysical images in general (e.g. Flöer, Winkel & Kerp 2014; Joseph, Courbin & Starck 2016; Offringa & Smirnov 2017; Irfan & Bobin 2018).

Once we wavelet transform $\mathbf{X}$ to $\mathbf{X}^{\rm wt}$, GMCA promotes sparsity in the sources $\mathbf{S}^{\rm wt}$ in wavelet base by solving iteratively the following optimization problem:

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{S}}\} = \min_{\mathbf{A},\mathbf{S}^{\rm wt}} \sum_{i=1}^{n_{\rm s}} \lambda_i \left|\left|\mathbf{S}_i^{\rm wt}\right|\right|_1 + \left|\left|\mathbf{X}^{\rm wt} - \mathbf{AS}^{\rm wt}\right|\right|_{\rm F}^2, \tag{4}$$

where the first term is a sparsity constraint term and the second is a data-fidelity term. Indeed, $||\cdot||_1$ is the ℓ_1 norm[3] defined by $||\mathbf{Y}||_1 = \sum_{i,j} |\mathbf{Y}_{i,j}|$; and $||\cdot||_{\rm F}$ the Frobenius norm defined by $||\mathbf{Y}||_{\rm F}^2 = {\rm Trace}(\mathbf{YY}^{\rm T})$. In particular, λ_i are regularization coefficients – sparsity thresholds – essential to provide robustness with respect to the noise of the problem, i.e. in our case the difference in intensity between the foregrounds and the cosmological signal; we first estimate them with the median absolute deviation (MAD) method and progressively decrease towards a final noise-related level. We refer the reader to Bobin et al. (2015) for details about the thresholding strategy.

As neither for $\mathbf{A}$ nor for $\mathbf{S}$ we use a model, GMCA is said to be a blind method, where the only input needed is the number of components $n_{\rm s}$ it searches and its assumption is constituted by sparsity.

The algorithm we employ here is openly available at www.cosmostat.org and demonstration scripts for reproducing the results of this paper are available at https://github.com/isab3lla/gmca4im.

[1] http://www.bingotelescope.org/en/
[2] https://fast.bao.ac.cn

[3] For a pure sparse solution, we could substitute it with the ℓ_0 norm: $||\mathbf{Y}||_0$, the number of non-zero entries in $\mathbf{Y}$.

Table 1. Schematic descriptions of the components of the simulated maps.

Component	Method/template	Parameters
Cosmological 21-cm signal	Lognormal approximation (CRIME; Alonso, Ferreira & Santos 2014)	$b_{\rm HI}(z) = 0.3(1+z) + 0.6$, $\Omega_{\rm HI}(z) = 4(1+z)^{0.6} \times 10^{-4}$
Galactic synchrotron	Planck Legacy Archive – FFP10 + high-resolution padding as in equation (8)	Spatially varying spectral index ($\beta_{\rm s} \sim -3$)
Galactic free–free	Planck Legacy Archive – FFP10	Constant spectral index $\beta_{\rm s} = -2.13$
Extragalactic point sources	Empirical model by Battye et al. (2013), Poisson and clustering contributions as in Olivari et al. (2018)	Source flux threshold $S_0 = 1$ Jy, background $T \sim \nu^{\alpha}$ (α from a Gaussian distribution with $\alpha_0 = -2.7$, $\sigma = 0.2$)
Polarization leakage	Galactic synchrotron polarization + rotation measurement (CRIME; Alonso et al. 2014)	Fraction of leaked polarization $\epsilon = 0.5\%$, Faraday space corr. length $\xi_\Psi = 0.5$ rad m^{-2}
Telescope beam	Gaussian smoothing	Frequency-dependent $\theta_{\rm FWHM}$, see equation (9)
Instrumental noise	White, frequency-dependent $\sigma_{\rm N}$ as in equation (10)	See Table 2

3 SIMULATIONS

In this section, we describe the simulated data we test the foreground removal technique on. To reproduce a truthful sky at frequencies of 900–1400 MHz, we sum together different components: (i) the 21-cm cosmological signal, for which we use the lognormal approximation as proposed by Alonso et al. (2014); (ii) astrophysical foregrounds, of galactic origin – synchrotron and free–free diffuse emissions – that we estimate using the Planck Sky Model, and of extragalactic origin, for which we adopt the model by Battye et al. (2013); (iii) lastly, we consider polarization leakage: a systematic known to be critical in H I intensity mapping experiments (Santos et al. 2015) that we model as in Alonso et al. (2014). Polarization leakage is difficult to deal with because it is expected to be non-smooth in frequency – as we will later explicitly show – and common foreground removal techniques are not aimed at picking components *misbehaving* in frequency. Thus, very little has been done in the literature to attempt to remove its contribution from the signal (Shaw et al. 2015) and to our knowledge there have been no attempts to remove it blindly, i.e. assuming no prior on its characteristics.

For each frequency and for all components, we generate HEALPIX maps with NSIDE $= 256$ that correspond to $n_{\rm pix} = 12\,{\rm NSIDE}^2$ pixels per map (Górski et al. 2005). We make this *parent* simulation publicly available at http://doi.org/10.5281/zenodo.3991818 (Carucci, Irfan & Bobin 2020); all scenarios addressed in this paper can be derived from it.

We then merge all components into sky maps and we mimic survey-specific features: we smooth maps with a frequency-dependent Gaussian filter to mimic the effect of the telescope beam and we finally add white noise to each channel, following standard thermal noise calculations. In the next paragraphs, we describe with more detail each of the above steps. Main properties of the simulated components are summarized in Table 1.

3.1 Cosmological signal

After reionization ($z < 6$), most H I in the Universe is stored inside galaxies, where it is dense enough to self-shield against the ionizing power of the cosmic ultraviolet background (Noterdaeme et al. 2012; Zafar et al. 2013). Thus, we can associate H I to the densest regions of the underlying dark matter field: we approximate the latter by a lognormal realization (Coles & Jones 1991) and assume H I is its linear biased tracer. We make use of the CRIME[4] algorithm, described in Alonso et al. (2014). We minimally modify CRIME, as we choose to set a redshift-dependent H I bias $b_{\rm HI}(z) = 0.3(1+z) + 0.6$ in agreement with observations at redshift $z \lesssim 0.8$ (Martin et al. 2012; Switzer et al. 2013) and to set the overall H I cosmic abundance to $\Omega_{\rm HI}(z) = 4(1+z)^{0.6} \times 10^{-4}$, as compiled by Crighton et al. (2015).

The lognormal realization has cosmological parameters $\{\Omega_{\rm m}, \Omega_\Lambda, \Omega_{\rm b}, h\} = \{0.3, 0.7, 0.049, 0.67\}$, with an initial cube of side 3 Gpc h^{-1} divided in 2048^3 cells. Light-cone effects and redshift-space distortions are included by construction. The original simulation is composed of 400 channels of 1 MHz of thickness, covering a redshift range of $z \in [0.09\text{–}0.58]$, corresponding to frequencies $\nu \in [900\text{–}1300]$ MHz. We later rebin the simulation as described in Section 3.5 before performing the blind source separation.

The lognormal approximation is appropriate for this study, especially considering that we later smooth the maps with a typical beam of $\approx 1^\circ$, losing the small-scale information of the field. The large-scale properties displayed by the simulated H I field roughly match those seen in local Universe H I galaxy survey (Obuljen et al. 2019), in state-of-the-art hydrodynamical simulations (Villaescusa-Navarro et al. 2018) and in state-of-the-art galaxy evolution models coupled to *N*-body simulations (Spinelli et al. 2020).

3.2 Astrophysical foregrounds

The astrophysical foregrounds featured in these simulations can be divided into two groups: galactic and extragalactic. For the extragalactic radio sources we implement the empirical model of Battye et al. (2013), who obtain their differential source counts from an empirical fit to numerous 1.4 GHz source surveys. By integrating these source counts a mean offset temperature, representing the unresolved sources is calculated for 1.4 GHz. The point sources also contribute a clustering and Poisson component to the overall point source temperature per pixel; these are calculated in angular power space and then converted to pixel space using the HEALPIX SYNFAST routine. Finally, any point sources over 0.01 Jy are injected at random into the map as fully resolved sources using the map pixel area and the number of sources (N) per steradian with a flux of S (Olivari et al. 2016):

$$T_{\rm ps}(1.4\,{\rm GHz},\, p) = \left(\frac{\lambda^2}{2k_{\rm B}}\right) \Omega_{\rm pixel}^{-1} \sum_{i=1}^{N} S_i\,. \tag{5}$$

We assume that sources brighter than 1 Jy have been identified and removed from the data. In order to scale this 1.4 GHz estimate across our frequency range we use a power law where the spectral

[4] http://intensitymapping.physics.ox.ac.uk/CRIME.html

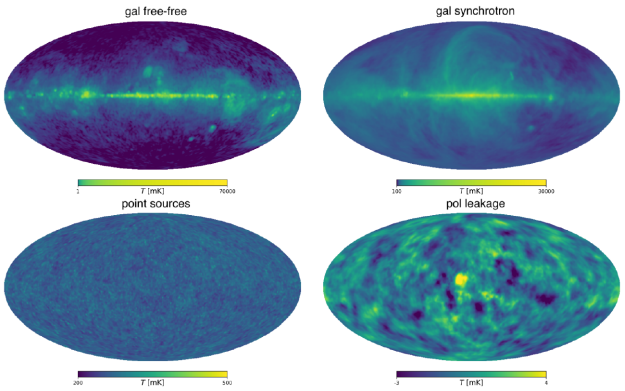

Figure 1. Mollweide projections of the temperature of the contaminant components in the simulation, from top left clockwise: galactic free–free, galactic synchrotron, polarization leakage, and point sources. Units are in mK and the colour bar is in logarithmic (linear) scale for the top (bottom) maps. Maps correspond to frequency $\nu = 1101$ MHz and have been convolved with the telescope beam as described in the text.

index varies, according to a Gaussian distribution, over the sky. For the Gaussian distribution we choose a mean value of -2.7 and a standard deviation of 0.2 (Bigot-Sazy et al. 2015).

The diffuse galactic foregrounds present in intensity at MHz frequencies are synchrotron and free–free emission. These emissions can both be modelled at each pixel p as power laws with an amplitude T_s and spectral index β_s:

$$T(\nu, p) = T_s(p) \left(\frac{\nu}{\nu_0}\right)^{\beta_s(\nu,p)} . \quad (6)$$

The Planck Legacy Archive[5] FFP10 simulations provide the synchrotron and free–free all-sky amplitudes and the synchrotron spectral index. We use the FFP10 simulations at 217 GHz for the free–free and synchrotron amplitudes, at NSIDE=2048, and degrade and smooth these maps to our desired NSIDE and resolution using the HEALPIX routines. The synchrotron spectral index map used (β) is that of Miville-Deschênes et al. (2008) and is at a resolution of around 5°. To provide spectral index information at angular scales smaller than 5° we combine the synchrotron spectral map with a map of small-scale structure:

$$\beta_{sy} = \beta + \beta_{ss}, \quad (7)$$

where the small-scale fluctuations (β_{ss}) are taken from Santos, Cooray & Knox (2005) and adapted to have a smaller amplitude:

$$C_\ell^{\beta_{ss}} = 7 \times 10^{-6} \left(\frac{1000}{\ell}\right)^{2.4} \left(\frac{\nu_r^2}{\nu_1 \nu_2}\right)^{2.8} \exp\left(\frac{-\log(\nu_1/\nu_2)^2}{2 \times 4^2}\right), \quad (8)$$

where ν_r is 130 MHz, ν_1 is 580 MHz, and ν_2 is 1000 MHz.

The synchrotron spectral index varies across pixels; this is not the case for the free–free spectral index. In alignment with the known range of -2.15 to -2.10 (Dickinson, Davies & Davis 2003), we set the value for the free–free spectral index to be -2.13 and keep this constant across the whole sky and over our full frequency range.

Fig. 1 shows the all-sky foreground maps that constitute the astrophysical components of our simulation.

[5] http://pla.esac.esa.int/pla

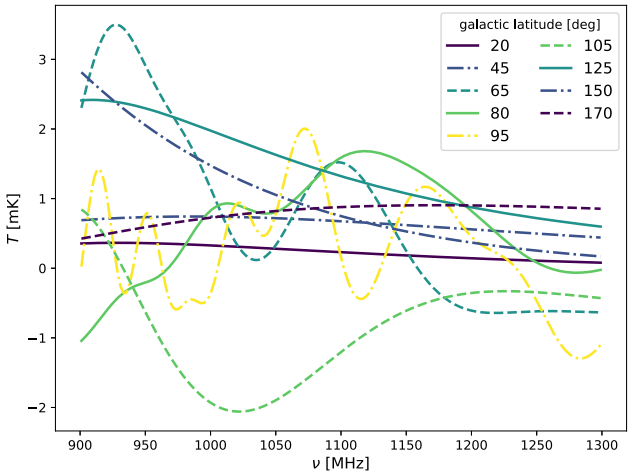

Figure 2. Temperature contribution of the polarization leakage as function of frequency, observed along different lines of sight at constant galactic longitude but different latitudes. The leakage has a smooth behaviour at the poles – although dependent on line of sight – and oscillates when closer to the galactic plane.

3.3 A non-smooth contaminant: polarization leakage

While the H I radiation is unpolarized, polarized foregrounds such as the galactic and extragalactic synchrotron emission – additionally altered by Faraday rotation in the interstellar medium – can spill into the unpolarized part of the received signal due to miscalibration issues (Moore et al. 2013).

In the community, we still lack a baseline on how to model this systematic, due to lack of knowledge of the galactic synchrotron polarization at the frequencies relevant for 21-cm intensity mapping and of the galactic magnetic field and ionized medium where Faraday rotation happens. Ongoing and future surveys in polarization will help us bridge this gap (e.g. Carretti et al. 2019).

Meanwhile, in the literature there are two polarization leakage models available for these frequencies, described in Alonso et al. (2014) and Shaw et al. (2015). Even if both models are built on the same data set (the galactic Faraday depth by Oppermann et al. 2012), their resulting polarization leakage maps are qualitatively different (because astrophysical assumptions ought to be made even without strong supporting observational evidence). Heuristically, the Alonso et al. (2014) model outputs a more dramatic[6] leakage contamination. Therefore, we take it as a conservative guess of the true systematic and use it to complement the simulated data of this work.

Although instrument dependent, the fraction ϵ of the spilled polarized signal is expected to be below 1 per cent (Santos et al. 2015); for instance, it has been estimated to be of the order 0.6–0.8 per cent for the GBT at 800 MHz (Liao et al. 2016). Foreseeing improvements in newer instruments and for updated calibration techniques, we set $\epsilon = 0.5$ per cent for this study; this ϵ yields to a temperature contribution one order of magnitude higher than the cosmological signal in 21 cm, as we will later see.

A snapshot of the polarization leakage contribution simulated with CRIME is in the bottom right of Fig. 1 for the $\nu \in [1100–1102]$ channel: it has a spotty angular dependence. To appreciate its line-of-sight behaviour, we plot in Fig. 2 its equivalent brightness temperature as function of frequency, with different colours corresponding to different galactic latitudes: especially when getting closer to the

[6] In terms of non-smooth frequency behaviour.

Table 2. Instrumental parameters used for computing the instrumental noise and the beam size.

Telescope specifics		
Dish diameter	D	13.5 m
Instrumental temperature	$T_{\rm instr}$	20.0 K
Observed fraction of the sky	$f_{\rm sky}$	0.1
Observation time	$t_{\rm obs}$	4000 h
Number of dishes	$N_{\rm dishes}$	64

galactic equator it fluctuates substantially, becoming highly pixel dependent.

We leave for subsequent work the study of the effect of other sources of systematics, as for instance satellites contribution (Harper & Dickinson 2018) and the so-called $1/f$ noise (Harper et al. 2018). Up to now, polarization leakage is the least explored of known systematics and has been proven to be hard to calibrate it out (Liao et al. 2016), in contrast with – keeping the same examples as before – satellite contamination that can be modelled and even avoided and the $1/f$ noise that can be mitigated with the scanning strategy and in the map-making process. This is why we prioritize the inclusion of the polarization leakage in the simulated data for this first GMCA study.

3.4 Instrumental effects: telescope beam and thermal noise

Once all the components are generated and combined, two instrumental effects are implemented to all maps: the smearing of a frequency-dependent beam and the addition of uncorrelated thermal noise.

We approximate the telescope beam with a symmetric Gaussian beam whose width depends on frequency as

$$\theta_{\rm FWHM} = \frac{c}{\nu D}, \tag{9}$$

with c the speed of light and D the telescope dish diameter. Considering the frequency range (900–1300 MHz) and the dish diameter chosen (see Table 2), the observed maps are smeared out to 1°–1.°4.

Approximating the telescope beam with a spherically symmetric Gaussian smoothing is what is usually done in the 21-cm intensity mapping foreground cleaning literature. However, it is a simplistic assumption as the presence of side lobes in the beam profile is responsible for additional mode mixing in the data. In this respect, the spectral complexity of the leakage contribution we include in this work can be seen as a first attempt to consider a component whose behaviour is close to what we would expect with a realistic beam too. Moreover, work is ongoing for adding the effect of proper beam side lobes in the simulated data (Asad et al. 2019), hence we leave the issue for a next study.

We assume the instrumental noise follows a uniform Gaussian distribution over the sky, with a frequency-dependent standard deviation of

$$\sigma_{\rm N}(\nu) = T_{\rm sys}(\nu)\sqrt{\frac{4\pi f_{\rm sky}}{\Delta\nu\, t_{\rm obs} N_{\rm dishes}\Omega_{\rm beam}}}, \tag{10}$$

where $T_{\rm sys}(\nu)$ is the system temperature, $f_{\rm sky}$ the observed fraction of the sky, $\Delta\nu$ the channel width, $t_{\rm obs}$ the total survey time, and $N_{\rm dishes}$ the number of dishes. The beam solid angle is related to its width as $\Omega_{\rm beam} = 1.133\theta^2_{\rm FWHM}$. The system temperature $T_{\rm sys}(\nu)$ is the sum of the receiver temperature and the sky temperature at a given frequency, which results in a combination of the instrument temperature $T_{\rm instr}$

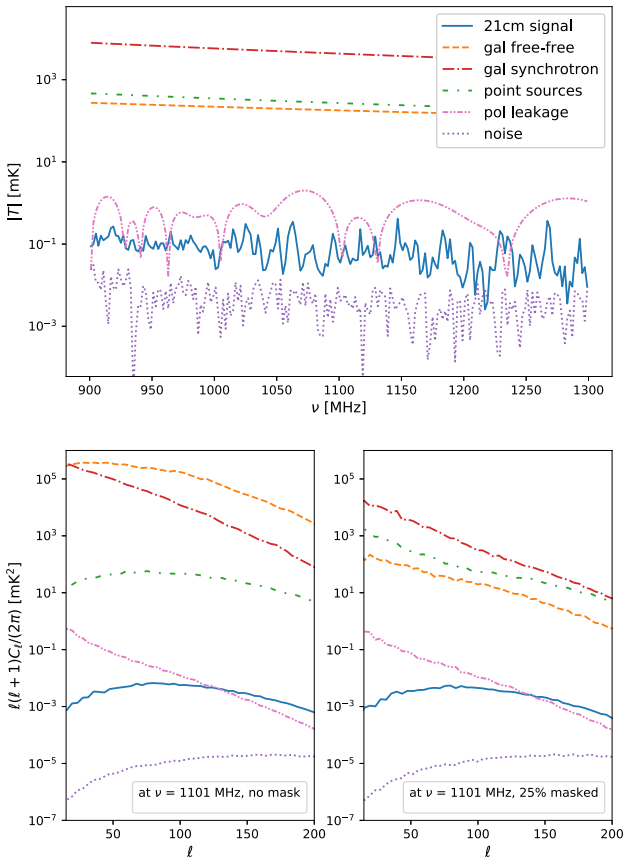

Figure 3. Components of the simulation. Top panel: brightness temperature as a function of frequency, observed along a random line of sight at galactic latitude of 95°. Bottom panel: angular power spectra of channel $\nu \in$ [1100–1102] MHz ($z \approx 0.3$) of the full-sky maps (left) and of the 75 per cent of the sky (right) when a mask has been applied. Removing the brightest pixels at the galactic equator, where most of the galactic free–free emission is concentrated, changes the contribution of the different contaminants. The data have been smoothed as discussed in the text, suppressing power at small scale.

and the observed frequency (O'Neil 2002):

$$T_{\rm sys}(\nu) = T_{\rm instr}\,[{\rm K}] + 66\left(\frac{300}{\nu\,[{\rm MHz}]}\right)^{2.55}. \tag{11}$$

The instrument and survey specifications used here are based on a MeerKLASS-like survey (Santos et al. 2017) and summarized in Table 2. We generate full-sky noise maps using equation (10) as variance per pixel.

3.5 Observed temperature cubes

We summarize the different simulated components in Fig. 3: in the top panel their temperature contributions are plotted as a function of frequency along a random line of sight; the polarization leakage (pink dash–dotted line) clearly sticks out among the foregrounds, as the others are indeed smooth and order of magnitudes above the cosmological signal (solid), that looks as noisy as the instrumental noise (dotted).

For each channel – or correspondingly frequency or redshift – we sum the maps of the 21 cm cosmological signal **C** with those

of all foregrounds **F**, we convolve the total temperature map with the frequency-dependent beam, we add the white noise **N**. This makes up the observed data cube. According to the mixture model in equation (3), the temperature maps are all assumed to be at the same resolution. For that purpose, we further reconvolve all maps appropriately to let them all share the same resolution, i.e. that of the lowest frequency channel where the beam $\theta_{\rm FWHM}$ is the largest. Schematically,

$$\mathbf{X} = [(\mathbf{C} + \mathbf{F}) * B + \mathbf{N}] * (B_{\rm low} - B). \tag{12}$$

Having all temperature maps at the same resolution is not essential. We also perform the source separation without the additional deconvolution, but we typically have to set a higher number of sources $n_{\rm s}$ for reaching a satisfactory foreground cleaning – compared to the case where all maps share the same resolution – thus risking to overclean and miss true signal **C** in the residuals. This is expected as the mixture model of the signal becomes more complex due to the frequency-dependent effect of the beam. So we take into account the latter with the deconvolution $(B_{\rm low} - B)$.

Ultimately, this further deconvolution is a loss of smallest angular information available in the observed maps. We do not discuss this issue here, since a version of GMCA that performs the beam deconvolution at the same time as the blind source separation has been tested on two-dimensional data (decGMCA; Jiang, Bobin & Starck 2017) and effort is ongoing for extending decGMCA on data sampled on the sphere (Carloni Gertosio 2020); thus, the results of this paper would generically hold for a decGMCA application, with the advantage of retaining the fully available small-scale information.

The simulation spans a frequency range of $\nu \in [900–1300]$ MHz, corresponding to redshift $z \in [0.09–0.58]$. We slice the data cubes in bins of $\Delta\nu = 2,\ 5,$ and 10 MHz, corresponding to numbers of channels $n_{\rm ch} = 200,\ 80,$ and 40: in this way we can test the dependence of GMCA performance on $\Delta\nu$ and on $n_{\rm ch}$.

4 THE PIPELINE

4.1 Recovering the input signal

As anticipated in Section 2, GMCA looks for components of the signal that are sparse in the wavelet domain. Thus, once the data cube **X** of equation (12) has been assembled and its mean removed channel-wise, we wavelet transform it, obtaining $\mathbf{X}^{\rm wt}$. By looking at the principal eigenvalues of the covariance matrix of the data cubes **X** and $\mathbf{X}^{\rm wt}$ (Fig. 4, symbols in blue and orange, respectively), we can already appreciate the advantage of running the blind source separation in wavelet space rather than pixel space: the transition between highly correlated modes in frequency (foregrounds) and uncorrelated modes (signal and noise) happens for smaller eigenvalue number for the wavelet case; the latter is especially true when we add a component like polarization leakage (dots versus squares in Fig. 4) that mixes the modes of the covariance matrix and makes the transition among them smoother.

GMCA promotes sparsity in the decomposition process of $\mathbf{X}^{\rm wt}$, as in equation (4), and estimates the mixing matrix $\tilde{\mathbf{A}}$. We determine the foreground components that GMCA identifies, $\mathbf{X}^{\rm GMCA}$, by projecting the input data **X** on $\tilde{\mathbf{A}}$; the cleaned maps $\mathbf{X}^{\rm cleaned}$ are the residuals of the GMCA source separation:

$$\mathbf{X}^{\rm cleaned} = \mathbf{X} - \mathbf{X}^{\rm GMCA} = \mathbf{X} - \tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{\rm T}\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{A}}^{\rm T}\mathbf{X}. \tag{13}$$

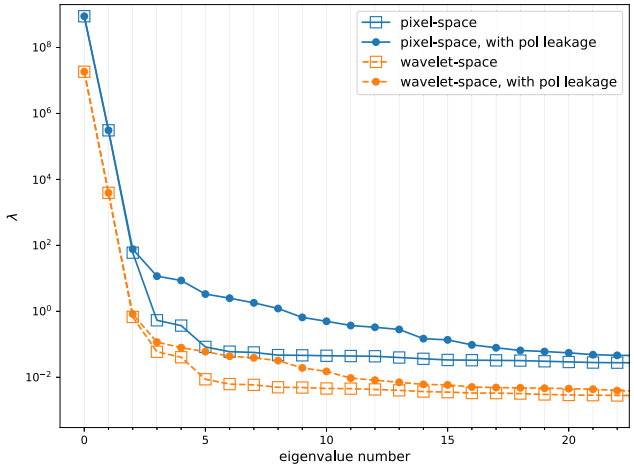

Figure 4. Principal eigenvalues of the frequency covariance matrix of the $n_{\rm ch} = 200$ simulation, using the standard pixel-space data cube **X** (blue lines) and its wavelet transformed counterpart $\mathbf{X}^{\rm wt}$ (orange). Empty squares (filled circles) correspond to the scenario without (including) the polarization leakage component. The contribution of a component like the polarization leakage spreads the foregrounds through a larger number of eigenvalues, making foreground cleaning more problematic, i.e. a larger number of degrees of freedom should be eliminated. On the other hand, working in wavelet space restricts the spreading in fewer degrees of freedom and makes the contamination more tractable.

4.2 GMCA performance

$\mathbf{X}^{\rm cleaned}$ of equation (13) are the maps that would be analysed for extracting science in a real context. In next paragraphs, we show how we evaluate the performance of the foreground cleaning by comparing $\mathbf{X}^{\rm cleaned}$ with the input data, i.e. the cosmological signal and the instrumental noise $\mathbf{C} + \mathbf{N}$.

4.2.1 Power spectrum estimation

The HI intensity two-point statistics carries a great deal of the cosmological information, as for any tracer of the underlying matter field. Hence, the performance of a given foreground cleaning method should be assessed at least in terms of its ability to recover the true power spectrum at different radial and angular scales.

Angular scales. Since the observed temperature data **X** is sampled on spheres of $n_{\rm pix}$ pixels, for a shell at fixed frequency ν, it is convenient to expand its distribution $\Delta T(\nu) = \mathbf{X}_\nu - \langle\mathbf{X}\rangle_\nu$ in spherical harmonic functions $Y_{\ell m}(p)$. We estimate the harmonic coefficients as a summation over the pixels p of the map:

$$a_{\ell m}(\nu) = \sum_{p=1}^{n_{\rm pix}} \Delta T(\nu, p) Y^*_{\ell m}(p). \tag{14}$$

All the above holds for any temperature data cube we assess, i.e. we can substitute **X** with foregrounds **F** or 21-cm cosmological signal **C** and so on. The angular power spectrum is defined as $C_\ell \equiv \langle|a_{\ell m}|^2\rangle$. For calculating the C_ℓ of each map, we make use of the software package NaMaster,[7] whose algorithm is described in Alonso et al. (2019). NaMaster is a pseudo-C_ℓ estimator that can also efficiently take into account incomplete sky coverage, as it will be the case in Section 5.4.

[7] https://github.com/LSSTDESC/NaMaster

For instance, in the bottom panels of Fig. 3 we plot the angular power spectra of all components of the simulation sampled at frequency $\nu = 1101$ MHz; in the left, the C_ℓ have been computed for the full-sky; in the right, we have first applied a mask covering the equatorial 25 per cent of map. The power amplitude of the astrophysical foregrounds is up to 7–8 orders of magnitude higher than the 21-cm signal in both panels; what change in the two cases are the individual foregrounds components spectra and their hierarchy, as consequence of the different spatial features of the foregrounds – morphological differences that will help GMCA to detect them. On the contrary, the C_ℓ of the cosmological signal and noise do not change among panels, showing that they are spatially isotropic.

For each channel map, we compare the angular power spectrum of $\mathbf{X}^{\rm cleaned}$ that we dub $C_\ell^{\rm cleaned}$, with that of the input $\mathbf{C}+\mathbf{N}$, i.e. $C_\ell^{\rm true}$. Averaging over all channels of the simulation, we build the quantity

$$R_\ell = \left\langle \frac{C_\ell^{\rm true} - C_\ell^{\rm cleaned}}{C_\ell^{\rm true}} \right\rangle_{\rm chs}. \tag{15}$$

We will use R_ℓ to assess the performance of GMCA in different scenarios.

Radial scales. Given the spectral nature of the 21-cm signal, the possibility of achieving unprecedented redshift resolution while sampling big volumes is one of the characteristics that makes intensity mapping highly appealing for cosmology. In this sense, it is crucial to investigate that also the radial direction information is retrieved after the cleaning process.

To estimate the two-point statistic in the radial direction, one could either rely on the angular cross-correlation of maps at different redshifts to avoid dealing with light-cone effects and curved-sky issues (Asorey et al. 2012; Montanari & Durrer 2012), or otherwise proceed with defining a one-dimensional $k_{\shortparallel}$ power spectrum estimator (Alonso et al. 2014; Villaescusa-Navarro, Alonso & Viel 2017; Blake 2019). Here we opt for a simpler approach: we compute the one-dimensional power spectrum directly in frequency space, $P(k_\nu)$ with $k_\nu = \frac{2\pi}{\nu}$. This choice makes difficult a direct comparison with cosmological observables, nevertheless it supplies a straightforward insight about the efficiency and deficiencies of foreground cleaning in the radial direction. In practice:

(i) for each line of sight (i.e. pixel), we Fourier transform the $\Delta T(\nu)$ field along ν,

$$\tilde{\Delta T}(k_\nu) = \int {\rm d}\nu\, \Delta T(\nu)\, {\rm e}^{-{\rm i}k_\nu \nu}; \tag{16}$$

(ii) we compute the power spectrum,

$$P(k_\nu) = \Delta\nu\, \langle |\tilde{\Delta T}(k_\nu, p)|^2 \rangle\,, \tag{17}$$

by averaging over all the lines of sight p.

In Fig. 5, we show the $P(k_\nu)$ for each component of the simulation. As for the C_ℓ, the amplitude of the foregrounds $P(k_\nu)$ is by far higher than that of the cosmological signal, the one of the noise is negligible. The high correlation in frequency of the foregrounds is also evident in their $P(k_\nu)$ that sharply increase towards higher frequency scales, in contrast with the 21-cm signal that – as the noise – displays a flat $P(k_\nu)$, with a slow decrease for high k_ν due to the effect of the beam smoothing (Villaescusa-Navarro et al. 2017).

In the same fashion of equation (15), we define the quantity R_ν for comparing the input and reconstructed radial power spectra:

$$R_\nu = \frac{P(k_\nu)^{\rm true} - P(k_\nu)^{\rm cleaned}}{P(k_\nu)^{\rm true}}\,. \tag{18}$$

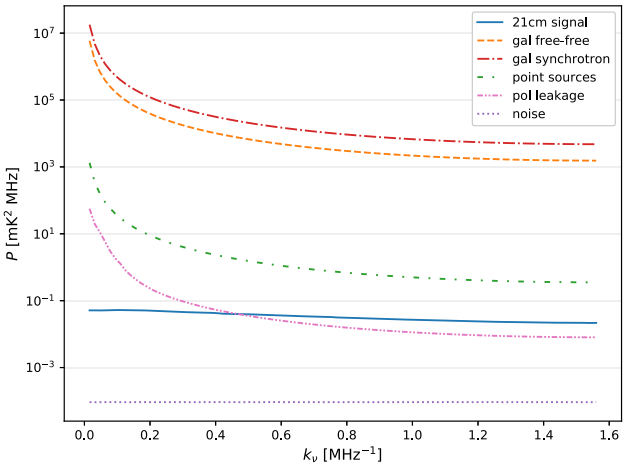

Figure 5. Radial power spectra in frequency space for all components of the $n_{\rm ch} = 200$ simulation, being smoothed by the telescope beam.

4.2.2 *Residual projection*

Two contributions make the cleaned maps go astray from the input $\mathbf{C}+\mathbf{N}$: (i) foregrounds are not fully captured in $\mathbf{X}^{\rm GMCA}$, contaminating $\mathbf{X}^{\rm cleaned}$; and (ii) true cosmological signal partly leaks into $\mathbf{X}^{\rm GMCA}$ and is lost. To quantify those effects individually, we define the residual projections.

The foreground residual that leaks into the recovered signal and noise is

$$\mathbf{X}_{\rm R}^{\rm F} = \mathbf{F} - \tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{\rm T}\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{A}}^{\rm T}\mathbf{F}, \tag{19}$$

where $\mathbf{F}$ is the input foregrounds data cube, from which we subtract the foreground maps projected on to the GMCA-estimated mixing matrix $\tilde{\mathbf{A}}$. Similar to equation (19), we define the signal plus noise, $\mathbf{C}+\mathbf{N}$, that leaks in the estimated foregrounds as

$$\mathbf{X}_{\rm F}^{\rm CN} = \tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{\rm T}\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{A}}^{\rm T}(\mathbf{C}+\mathbf{N}). \tag{20}$$

The foreground removal succeeds when the power spectra of both $\mathbf{X}_{\rm R}^{\rm F}$ and $\mathbf{X}_{\rm F}^{\rm CN}$ are negligible compared to that of $\mathbf{C}+\mathbf{N}$.

5 RESULTS

To better understand the foreground removal problem, we first run GMCA on data cubes with foreground contributions of galactic synchrotron and free–free diffuse emissions and extragalactic diffuse emission and point sources; we later increase the degree of complexity of the foregrounds by adding the polarization leakage. This first assessment makes also possible a more direct comparison of GMCA with other methods in the literature.

We study how the number and thickness of the channels affect the performance of the foreground cleaning; we further assess its performance when some channels are missing altogether, which is often the case in real surveys; we check whether masking the pixels with higher foreground contamination eases the cleaning task; we eventually add the polarization leakage in the game and, lastly, we try the same tasks with another source separation algorithm – FastICA – for comparison.

We start by visually inspecting the GMCA-reconstructed maps. We feed the 200-channel data cube (missing the polarization leakage contribution) to GMCA setting to $n_{\rm s} = 3$ the number of morphologically different sources to search. Fig. 6 shows the results for the $\bar{\nu} = 1101$ MHz channel: we show the sky Mollweide projections

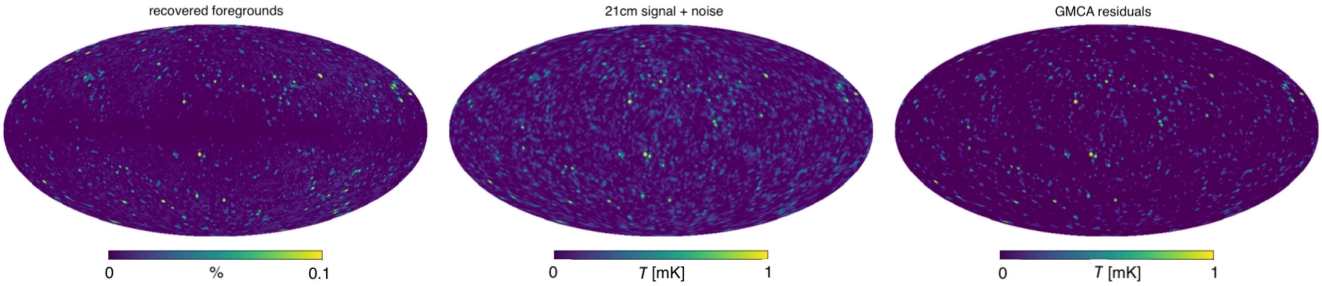

Figure 6. Foreground removal for the $\nu \in [1100–1102]$ MHz channel, GMCA has looked for $n_s = 3$ components. Left-hand map: relative difference in percentage between the input foregrounds, $\mathbf{F}$, and what found by GMCA, $\mathbf{X}^{\rm GMCA}$. Middle map: input 21-cm signal and instrumental noise, $\mathbf{C} + \mathbf{N}$. Right-hand map: foreground removal residuals $\mathbf{X}^{\rm cleaned}$, to be compared with middle map. Foregrounds total temperature is recovered at sub-per cent level, especially in the galactic plane, yet we miss some of the low-temperature features of the input signal map; the brightest 21-cm spots are recovered in the map in the right, and they also represent the regions at most off-set in the foreground recovery (left).

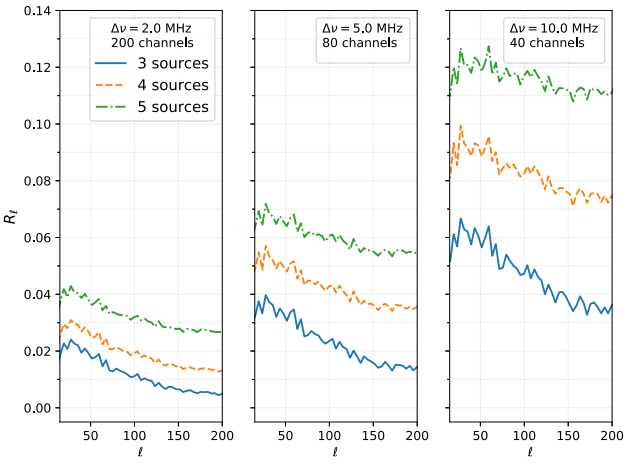

Figure 7. Relative difference in angular power spectrum between the GMCA recovered signal and the input cosmological signal and instrumental noise, averaged among all channels. The three different simulations (panels) are full-sky and GMCA has been run with number of sources $n_s = 3$, 4, and 5 (solid, dashed, and dash–dotted lines, respectively). These simulations lack the polarization leakage contribution. Overall, GMCA recovers the signal within a few per cent bias in angular power spectrum.

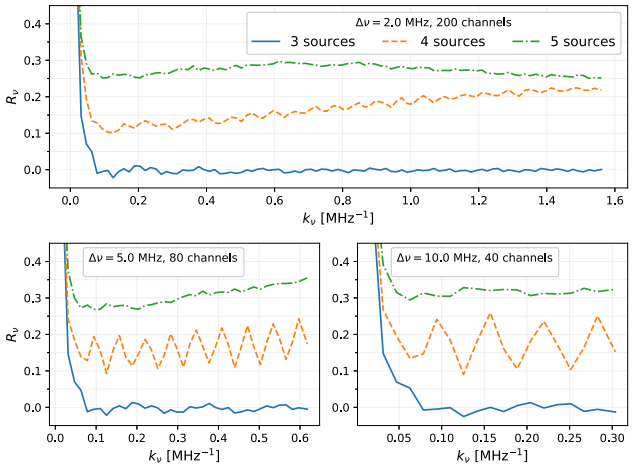

Figure 8. Relative difference in radial power spectrum between the GMCA recovered signal and the input cosmological signal and instrumental noise. The three different simulations (panels) are full-sky and GMCA has been run with number of sources n_s =3, 4, and 5 (solid, dashed, and dash–dotted lines, respectively). These simulations lack the polarization leakage contribution. GMCA recovers unbiased information in the radial direction for $k_\nu \gtrsim 0.05$ MHz^{-1} when $n_s = 3$ is set.

(i) left-hand panel: the difference in intensity between the input foregrounds $\mathbf{F}$ and what is identified by GMCA, i.e. $(1 - \mathbf{X}^{\rm GMCA}/\mathbf{F})_{\bar{\nu}}$; (ii) middle panel: the input signal and instrumental noise, $(\mathbf{C} + \mathbf{N})_{\bar{\nu}}$; (iii) right-hand panel: the cleaned map recovered with GMCA, $\mathbf{X}^{\rm cleaned}_{\bar{\nu}}$. Looking at the left-hand panel: GMCA has remarkably identified the true intensity of the foregrounds with sub-per cent level of accuracy. Is this achievement enough for the recovery of the feeble 21-cm signal? We compare the remaining panels: the input $\mathbf{C} + \mathbf{N}$ (middle) with the output $\mathbf{X}^{\rm cleaned}$ (right). The bright spots where emission is greatest are clearly present in both maps, however much of the fainter features present at all scales in the $\mathbf{C} + \mathbf{N}$ map are missing in $\mathbf{X}^{\rm cleaned}$. Next we will plot the power spectra of these maps to assess the information that we can still safely extract from $\mathbf{X}^{\rm cleaned}$. Inspecting further the maps of Fig. 6, we notice that the pixels where foregrounds are worse caught correspond to the bright spots of the true signal outside the galactic plane, whereas the galactic plane pixels, where foregrounds more strongly shine, correspond to those pixels that – counter-intuitively – experience the best recovery of the foreground emission. The latter remark will be further corroborated in Section 5.4 where we perform the foregrounds cleaning after applying masks to the maps.

5.1 The dependence on the number of channels

For each simulation set-up previously described (with different number and thickness of channels), we perform various foreground removals with GMCA varying the number of sources n_s. We compute angular and radial power spectra of all cleaned maps, and compare with the ground truth ones, as described in Section 4.2.1. We show the angular power spectra relative difference R_ℓ in Fig. 7 and the radial counterpart R_ν in Fig. 8. We plot these quantities for the simulation with $n_{\rm ch} = 200$ and $\Delta\nu = 2$ MHz, $n_{\rm ch} = 80$ and $\Delta\nu = 5$ MHz, and $n_{\rm ch} = 40$ and $\Delta\nu = 10$ MHz in panels from left to right (from top to bottom) in Fig. 7 (Fig. 8). Different lines correspond to $n_s = 3$, 4, and 5 (solid, dashed, dot–dashed, respectively).

Focusing on Fig. 7, in the best scenario – 200 channels and $n_s = 3$ – the angular power spectrum of $\mathbf{C} + \mathbf{N}$ is recovered on average with a 2 per cent bias on large scales down to 0.5 per cent for $\ell > 150$. Setting n_s to higher values leads R_ℓ to increase in amplitude, for instance doubling for the $n_s = 5$ case. Having a lower number of channels also impacts negatively R_ℓ, as its amplitudes in the middle and right-hand panels are higher and go up to 12 per cent for the $n_{\rm ch} = 40$, $n_s = 5$ case. We expect this as, even if the simulations cover the

same range in frequency, the higher number of channels/maps, the larger the data set GMCA can rely on for extracting components and mixing matrix.

Also assessing the information in the radial direction, results are promising: setting $n_s = 3$ we recover the power spectrum in frequency space within few per cent (solid lines in all panels of Fig. 8). This bias increases up to ≈35 per cent when increasing n_s, showing some mild scale dependence. In contrast to the angular R_ℓ, the results in R_ν happen to be quite n_{ch} independent; of course the smallest scale we can reach in k_ν is dictated by the frequency resolution $\Delta\nu$ of the simulations (the highest wavenumber to be trusted is $\pi n_{ch}/\Delta\nu$), none the less, R_ν is under control for $k_\nu \gtrsim 0.05$ MHz^{-1} for all $\Delta\nu$ scenarios. Ignoring light-cone effects, we can crudely relate k_ν to $k_{\shortparallel}$ – its comoving distance counterpart: $k_\parallel \approx \frac{\nu_{21\,\mathrm{cm}} H(z)}{c\,(1+z^2)}\, k_\nu$, with $H(z)$ the Hubble parameter; by using the cosmological parameters of the simulation and the middle redshift of the data cube, we can claim to recover the true radial power spectrum for $k_\parallel \gtrsim 0.02\,h$ Mpc^{-1}. A noticeable feature of the results in Fig. 8 is the oscillating behaviour of some of the R_ν displayed: it is due to *ringing* effects in computing the Fourier transforms because of the presence of numerical zeros in the ΔT data, originated when subtracting the map mean from pixels whose values were close to the mean; we explicitly checked that those effects disappear when we apply an additional and more aggressive smoothing on the liable maps, converging to a still R_ν.

Looking at Fig. 7, we confirm the expectation that the larger the number of channels available, i.e. the more the data, the better the GMCA performance at characterizing the foregrounds. However, since the three different simulations cover the same frequency range, a different number of channels lead to a different thickness $\Delta\nu$ of channels: could this latter parameter play a role in the way GMCA works? The angular power spectra of $\mathbf{C} + \mathbf{N}$ are higher for thinner channels, because of the higher instrumental noise but mainly because of purely geometrical considerations (e.g. the C_ℓ of $\mathbf{C} + \mathbf{N}$ for the $\Delta\nu = 2$ MHz case is roughly 40 times higher than in the $\Delta\nu = 10$ MHz case).

To clarify the role of both n_{ch} and $\Delta\nu$ in the foreground cleaning, we perform the following exercise. We run GMCA using only a sample of 40 consecutive channels of both the $n_{ch} = 200$ and $n_{ch} = 80$ channel simulations: the level of R_ℓ increases by five and three times, respectively, and independently of n_s, compared to the results in Fig. 7. It is thus clear that GMCA struggles more when it has access to less channels, independently of $\Delta\nu$. Moreover, remarking that (i) with 40 consecutive channels of the $\Delta\nu = 2$ MHz simulation the situation worsens more than with 40 consecutive channels of the $\Delta\nu = 5$ MHz simulation, and (ii) in both cases the performance of GMCA is worse than with the full 40-channel simulation with $\Delta\nu = 10$ MHz (right-hand panel of Fig. 7), points to the importance of the span in frequency of the data cube for a successful foreground removal. We will come again to the same conclusion when we will try GMCA on cropped data in the radio-frequency interference (RFI; Section 5.3): regardless of $\Delta\nu$, it is better to have GMCA working on the full frequency range available even when channels are missing. Instead, we find no strong arguments for aiming to a specific channel width, as far as it concerns the GMCA reconstruction.

To show how compelling are the span in frequency of the data cube and the number of channels we work with, we plot in Fig. 9 the results of the same $\Delta\nu = 2$ MHz channel (middle frequency $\nu = 1097$ MHz) when GMCA has run on the whole 200-channel data (left-hand column), or just on a 40-channel subset (right-hand column). All curves are angular power spectra: solid blue is the input $\mathbf{C} + \mathbf{N}$ and with orange plus signs we plot $\mathbf{X}^{\mathrm{cleaned}}$; other colours and line styles refer to the projections of the leaked signal $\mathbf{X}_{\mathrm{F}}^{\mathrm{CN}}$, of the total residual

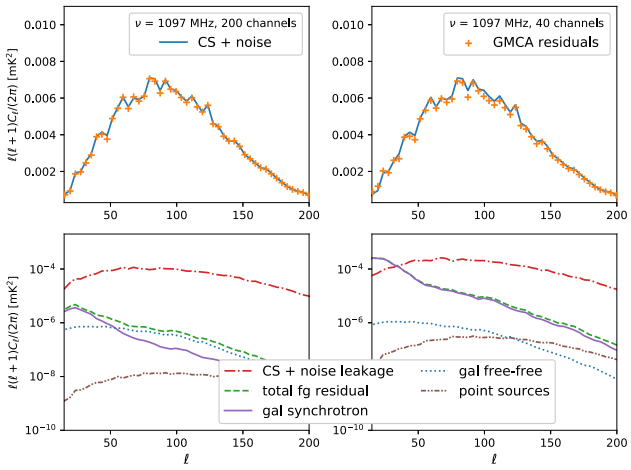

Figure 9. GMCA results for the $\nu \in [1096–1098]$ MHz channel of the full-sky simulation with no polarization leakage, using $n_s = 3$ number of components. In the top panels are the angular power spectra of the input cosmological signal and noise $\mathbf{C} + \mathbf{N}$ (solid line) and the GMCA recovered signal $\mathbf{X}^{\mathrm{cleaned}}$ (plus signs). In the bottom panels are the leakage of the input cosmological signal and noise into the GMCA found foregrounds $\mathbf{X}_{\mathrm{F}}^{\mathrm{CN}}$ (dash–dotted line), and the residuals of the input foregrounds left over in the GMCA recovered signal $\mathbf{X}_{\mathrm{R}}^{\mathrm{F}}$ (dashed line) and for the individual foreground component (see legend). The left-hand panel corresponds to a GMCA run on the full 200 channels available, in the right using only 40 consecutive channels of the simulation. With less channels, i.e. less frequency information to characterize the foregrounds, GMCA performs worse, especially for identifying the galactic synchrotron (featuring a spatially varying spectral index), whose residual leaks in the recovered signal at large scales.

foregrounds $\mathbf{X}_{\mathrm{R}}^{\mathrm{F}}$ and of residuals of single foreground components. The change in amplitude of the projection of the residual galactic synchrotron (solid violet lines in the lower panels) is evidence that, for the very same channel, GMCA characterizes synchrotron more poorly in the case on the right with the only difference being the smaller number of channels used and frequency span covered.

We choose $n_{ch} = 200$ to be the reference simulation in the rest of the analysis.

5.2 Selecting n_s

From Figs 7 and 8 it is clear that setting the sources GMCA looks for to $n_s = 3$ is optimal for the foregrounds contribution, sky coverage, and frequency range set-ups we are considering, leading to an unbiased recovery of the information in the radial direction and within few per cent in the perpendicular one. However, we will contradict this result later in the analysis, when masking the brightest pixels of the maps or adding a mode-mixing component as the polarization leakage. Practically, we cannot expect a specific values of n_s to hold in general because of the variety of realistic survey scenarios, which would be impossible to simulate perfectly, also taking into account the addition of unknown systematics or astrophysical contributions that could actually manifest in the observations and our ignorance of the 21-cm signal itself; moreover, specifically concerning how GMCA works, we cannot rely on the same level of sparsity when considering different regions of the sky and maps with different resolutions.

Indeed, when dealing with real data, it has been removed order ∼10 or more number of sources/independent components/principal modes (Chang et al. 2010; Masui et al. 2013; Switzer et al. 2013; Wolz et al. 2017; Anderson et al. 2018).

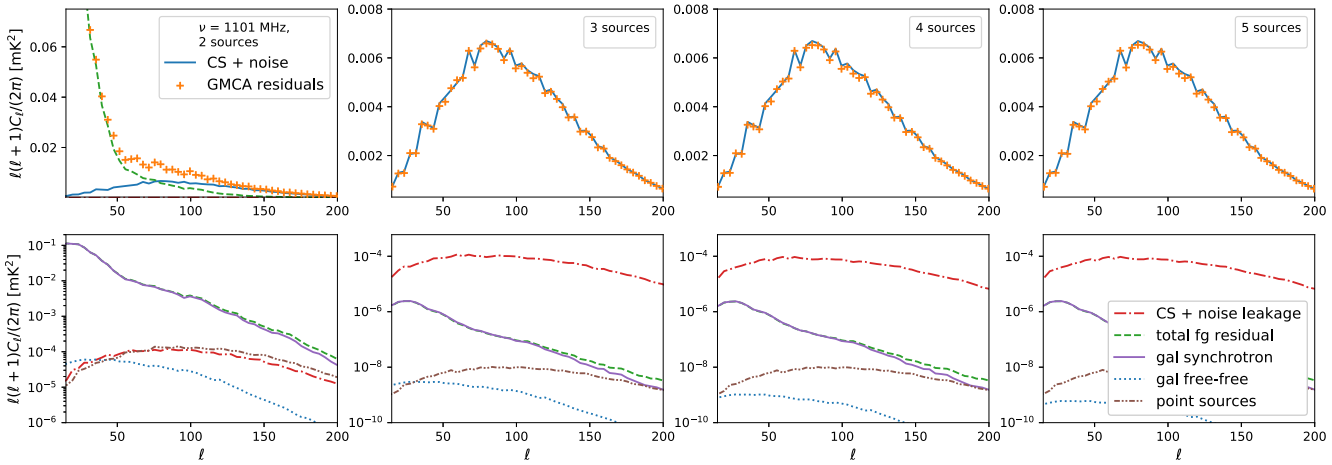

Figure 10. GMCA results for the $\nu \in [1100–1102]$ MHz channel of the full-sky simulation with no polarization leakage, using $n_s = 2$, 3, 4, and 5 number of components (panels from left to right). In the top panels are the angular power spectra of the input cosmological signal and noise $\mathbf{C} + \mathbf{N}$ (solid line) and the GMCA recovered signal $\mathbf{X}^{\rm cleaned}$ (plus signs). In the bottom panels are the leakage of the input cosmological signal and noise into the GMCA found foregrounds $\mathbf{X}_{\rm F}^{\rm CN}$ (dash–dotted line), the residuals of the input foregrounds left over in the GMCA recovered signal $\mathbf{X}_{\rm R}^{\rm F}$ (dashed line), and for the individual foreground components. There is a clear convergence in the foreground removal setting $n_s = 3$ and higher.

How to set the number of sources the blind source separation algorithm has to look for, having no ground truth to compare against? A good starting point is to look at the eigenvalues of the covariance matrix of the signal in the domain we work in, as we do in Fig. 4, although, especially with the inclusion of mode-mixing components (filled dots), it can be problematic to distinguish foregrounds modes from cosmological ones.

Here, we have a closer look at results of the simplified scenario (no polarization leakage) to check whether we could tell a priori $n_s = 3$ is optimal by looking at the recovered power spectra. We will later check if we will be able to apply what we learn in this simplified scenario in more complex ones.

We look more closely at how the GMCA performance changes when we vary n_s. In Fig. 10, we plot GMCA results for just one channel (of central frequency $\nu = 1101$ MHz); columns refer to different runs of GMCA where the number of sources has been set to $n_s = 2$, 3, 4, and 5 from left to right. All curves are angular power spectra: solid blue is the input $\mathbf{C} + \mathbf{N}$ and with orange plus signs we plot $\mathbf{X}^{\rm cleaned}$; other colours and line styles refer to the projections of the leaked signal $\mathbf{X}_{\rm F}^{\rm CN}$ (red dash–dotted), of the total residual foregrounds $\mathbf{X}_{\rm R}^{\rm F}$ (green dashed), and of residuals of single foreground components (these plots have same structure and colour coding of Fig. 9). The behaviour of the $\mathbf{X}^{\rm cleaned}$ spectrum changes abruptly from the $n_s = 2$–3 case (first two columns from left), whereas it stays stationary for the remaining $n_s = 4$ and 5 cases. In the $n_s = 2$ run, $\mathbf{X}^{\rm cleaned}$ is severely contaminated by foregrounds, up to fully overlap with the power spectrum of $\mathbf{X}_{\rm R}^{\rm F}$ for $\ell < 50$. Asking GMCA to look for $n_s = 2$ morphologically diverse components is not enough to pinpoint the foregrounds. The leap – in amplitude and behaviour – the spectra of $\mathbf{X}^{\rm cleaned}$ exhibits when passing to the $n_s = 3$ scenario is a hint for having reached an optimal n_s, further validated by the convergence the spectrum of $\mathbf{X}^{\rm cleaned}$ shows in the $n_s = 4$ and 5 plots. Looking at the power spectra of projections: increasing further number of components $n_s > 3$ helps (marginally) to better characterized the foregrounds (almost imperceptibly in these plots, with the exception of the galactic free–free component: the blue dotted line keeps decreasing in amplitude with increasing n_s), but it comes at the expense of having more leakage of the true signal (although imperceptible by eye as well). Setting $n_s = 3$ is optimal in this observational set-up, as already proven by Figs 7 and 8, and, noteworthy, we can reach this conclusion by examining the power spectrum of $\mathbf{X}^{\rm cleaned}$ alone, without comparing with the ground truth one.

5.3 Mimicking RFI

When performing radio observations, whole channels are discarded due to irreversible contamination by radio-frequency interference (RFI) generated for instance by FM radios and television stations, cellular network of mobile phones, satellites, and so on. Even in radio-quiet areas designated and protected for those experiments, RFI flagging is usually still necessary. For instance, for the ongoing MeerKLASS 21-cm intensity mapping *L*-band preliminary observations, roughly 40 per cent of the data in two separated chunks of flagged channels are typically discarded. As previously pointed out, the performance of the foreground removal depends on the number of channels and on the frequency range covered by the data cube. This motivates the question: how does having missing channels effect the foreground removal?

We mimic the RFI flagging effect by removing 40 per cent of the channels in the simulation and run GMCA on the 60 per cent that is left, i.e. on 120 channels in our case. We adopt three flagging scenarios, removing channels: (i) in one chunk at the centre of the frequency interval; (ii) in one chunk at the beginning of the frequency interval (remaining with the first 10 per cent of channels and the last 50 per cent); and (iii) in two chunks of different lengths (in the order: 20 per cent good, 30 per cent flagged, 20 per cent good, 10 per cent flagged, 20 per cent good). Results are shown in Fig. 11, in terms of recovery of angular scale information R_ℓ in the top panel and of parallel scale information R_ν in the bottom panel; the different line styles correspond to the three RFI scenarios. GMCA has been run setting $n_s = 3$. The overall bias level in the angular power spectra of the cleaned maps is analogous with what we measure for a non-RFI-contaminated data cube composed by 120 channels (i.e. a situation between the left-hand and middle panel of Fig. 7). Also the scale dependence of R_ℓ is not stronger than that of the continuous data cube case. Among the three different RFI scenarios, the last one with three frequency-discontinuous chunks of data is slightly better

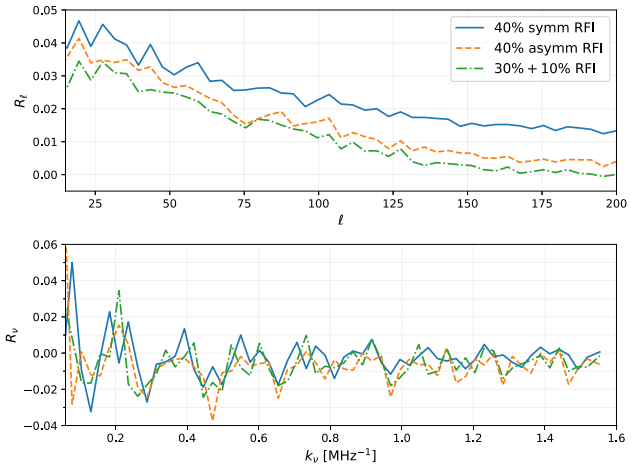

Figure 11. In the top (bottom) panel is the relative difference in angular (radial) power spectrum between the GMCA ($n_s = 3$) recovered signal and the input cosmological signal. Out of the 200 channels of the simulations, 80 have been discarded: in one chunk at the centre of the frequency interval (solid line), in one chunk in the first half of the frequency interval (dashed), and in two separated chunks (dot–dashed). Frequency-incomplete data do not compromise the GMCA foreground separation.

performing, probably due to the better frequency coverage of the data. Also the radial power spectrum results R_ν in the bottom panel of Fig. 11 are consistent with those obtained with continuous data cubes in Fig. 8, being below 5 per cent at large scales and going down to ≈1 per cent for small scales, with essentially no difference among the three RFI scenarios. When run on RFI-affected data cubes, GMCA yields to mixing matrices $\tilde{\mathbf{A}}$ with *jumps* in columns, thus recognizing the discontinuous nature of the data and being able to benefit from the whole data available without the need to partition and lose frequency information of the components.

Summarizing, it is reliable and still effective to use GMCA with flagged – i.e. discontinuous – data.

5.4 Masking

It has been reported that masking the angular regions where foregrounds are more intense benefits the foreground cleaning process (Wolz et al. 2014; Alonso et al. 2015; Bigot-Sazy et al. 2015; Olivari et al. 2016). We test if this is also the case for the cleaning performed with GMCA, masking out the pixels of the sky where the simulated observed temperature is brightest. We consider brightness thresholds that lead to masks covering the 10, 25, and 50 per cent of the full sky, inevitably hiding the galactic plane, as shown in Fig. 12.

The wider the mask, the less the pixels and the information GMCA relies on, making unfair a direct comparison of the exercise of this section with the previous ones. Nevertheless, it can tell us whether covering the most contaminated region helps the cleaning in the leftover area.

Our findings are summarized in Fig. 13: in the top row the angular power spectrum relative difference R_ℓ, in the bottom row the radial counterpart R_ν, for runs of GMCA looking for $n_s = 3$ (left-hand column) and 4 sources (right-hand column). In the $n_s = 3$ scenario, GMCA struggles more to identify the foregrounds in the masked data. In the case of masks of 25 and 50 per cent, R_ℓ is negative, thus the spectrum of the cleaned maps is higher than that of the ground truth: we can push the number of sources to look for, as we do in the right-hand panel. For $n_s = 4$, results for the masked scenarios are

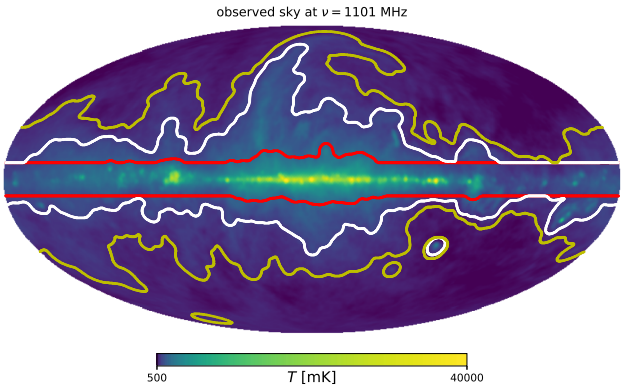

Figure 12. Total temperature map of channel $\nu \in [1100–1102]$ of the simulation. We overplot with solid coloured lines the different masks we use, covering the brightest pixels up to the 10, 25, and 50 per cent of the sky. As expected, it is the galactic plane to be masked out, up to the synchrotron North Polar Spur for larger masks.

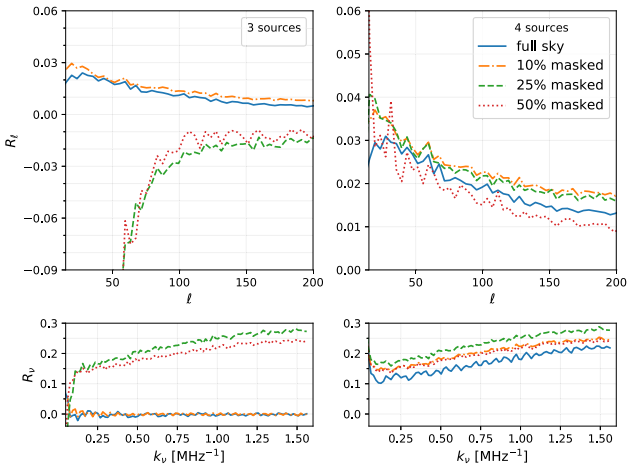

Figure 13. Relative difference in power spectrum between the GMCA recovered signal and the input cosmological signal and instrumental noise: angular R_ℓ (radial R_ν) in the top (bottom) row. GMCA has been run on the data cube without polarization leakage and using $n = 3$ (4) number of sources in the left (right) column, and masking out the 10 per cent (dotted line), 25 per cent (dash–dotted), 50 per cent (solid), or using the full-sky maps (dashed). Masking out the region where galactic synchrotron and free–free emissions are more intense, makes it harder for GMCA to reconstruct them. Increasing the number of sources can overcome this at the angular power spectrum level (top right), but the radial one is nevertheless compromised (bottom panels). We note that the masking has an effect on the power spectrum estimation, for the angular one it reduces the number of large modes available (but this affects mainly scales $\ell < 10$) for the radial it reduces the number of lines of sight available.

indeed closer to the full-sky reference, expect at the very large-scales where anyway the angular power spectrum estimation is affected by having less large modes at disposal due to the partial-sky maps. On the other hand, looking at R_ν in the lower panels, the 10 per cent mask does not compromise the recovery of information in the radial direction for $n_s = 3$, and increasing to $n_s = 4$ does not improve the R_ν level for the 25 and 50 per cent masked cases.

Masking the most contaminated pixels does not help the GMCA reconstruction. On the contrary, we suspect that the morphological detection part of the algorithm (sparsity in the wavelet domain) characterizes contaminants better when their features are strongly

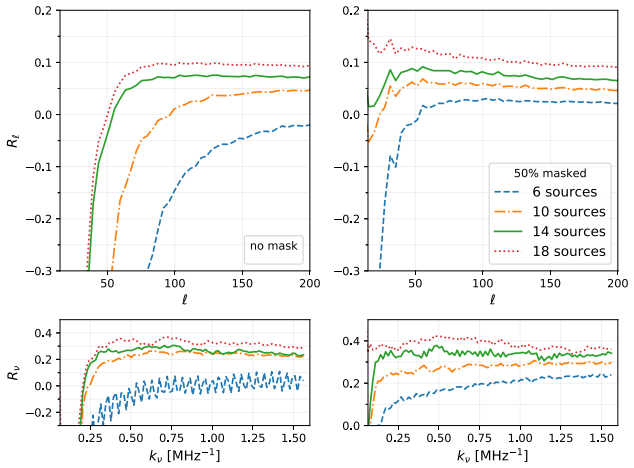

Figure 14. Relative difference in power spectrum between the GMCA recovered signal and the input cosmological signal and instrumental noise: angular R_ℓ (radial R_ν) in the top (bottom) row. GMCA has been run on the data cube that also contains a 0.5 per cent polarization leaked into the unpolarized signal. On the left-hand panel, we consider the full-sky maps, on the right 50 per cent of the maps have been masked out. Different line styles and colours correspond to different number of components $n_s = 6$, 10, 14, and 18 GMCA has been run with. When excluding the galactic plane region (right-hand panel), GMCA reconstructions improve. Anyway, also for the full-sky scenarios in left-hand panel, results are encouraging at small scales.

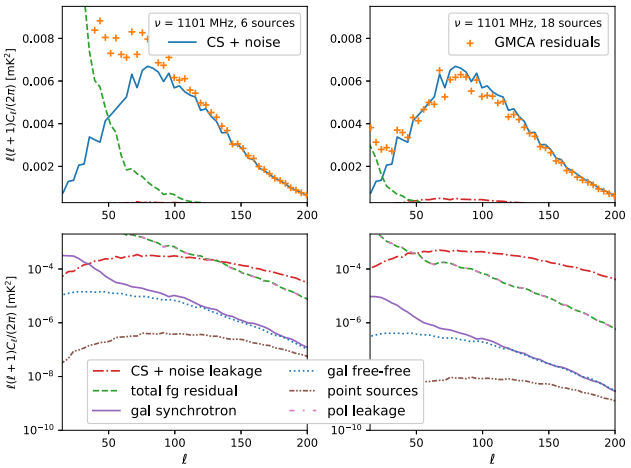

Figure 15. GMCA results for the $\nu \in [1096–1098]$ MHz channel of the simulation full sky and with polarization leakage. In the top panels are the angular power spectra of the input cosmological signal and noise $\mathbf{C}+\mathbf{N}$ (solid line) and the GMCA recovered signal $\mathbf{X}^{\rm cleaned}$ (plus signs). In the bottom panels are the leakage of the input cosmological signal and noise into the GMCA found foregrounds $\mathbf{X}_{\rm F}^{\rm CN}$ (dash–dotted line), the residuals of the input foregrounds left over in the GMCA recovered signal $\mathbf{X}_{\rm R}^{\rm F}$ (dashed line), and of the individual foreground component (see legend). The left (right) column corresponds to a GMCA run with $n_s = 6$ (18) number of components. Polarization leakage (pink dashed) is the worst-identified foreground as it dominates the total foreground residual budget $\mathbf{X}_{\rm R}^{\rm F}$. The latter is more under control in the right-hand panel, although there is an increase of the cosmological signal that gets lost: the increase of the $\mathbf{X}_{\rm F}^{\rm CN}$ power spectrum from the left to the right scenario is hardly visible in the bottom panels (with logarithmic scales), but it is evident as we start seeing it in the top right-hand panel too, less than an order of magnitude away from the recovered $\mathbf{X}^{\rm cleaned}$ power spectrum.

present. This is at odds with other foreground removal methods and yields to (i) the advantage of working with the full data set available and (ii) to more flexibility with the choice of the survey target sky area to begin with, allowing for survey designs with greater commensality with other science scopes (e.g. galactic astrophysics).

We stress again that the masking under study in this section refers to an a posteriori covering of bright pixels in the data available. Real surveys do have a mask – footprint – on their own, as it is highly improbable to observe the full sky. The study of the performance of GMCA in different regions of the sky – with different levels of sparsity of the foregrounds – is another issue that merits more detailed work.

5.5 Including the polarization leakage

Up to now, we have looked at the performance of GMCA on simulated data that do not include polarization leakage. In this section, we finally add the distressing component in the game.

Our findings are summarized in Fig. 14, where we plot R_ℓ and R_ν (top and bottom rows) of the results for full and 50 per cent masked sky scenarios (left- and right-hand columns, respectively) that GMCA yields when run with $n_s = 6$, 10, 14, and 18 number of components. The addition of polarization leakage undoubtedly makes source identification by GMCA more troublesome and the number of components to look for has to increase to reach satisfactory levels of cleaning, as it could already be expected by looking at the principal components of the data frequency covariance matrix in Fig. 4. Looking at the right-hand panels of Fig. 14, the situation remarkably improves when we hide the region of the sky where the polarization leakage has the most complex and uneven frequency behaviour (see Fig. 2). For the left-hand panels case, we can nevertheless make use of the GMCA reconstruction for $\ell > 80$ and $k_\nu > 0.3$ MHz^{-1} for the higher n_s considered, as the bias introduced in the recovered power spectrum is scale independent and, therefore, can be easily taken into account (Cunnington et al. 2020) and even marginalized over in cross-correlation analysis.

More about the scale independence of both R_ℓ and R_ν: we can push it to hold for lower scales by increasing n_s at the expense of increasing the amplitude of R_ℓ and R_ν, i.e. yielding to a $C_\ell^{\rm cleaned}$ and a $P(k_\nu)^{\rm cleaned}$ that underestimate the true spectra. We have a hint about this when looking at the principal eigenvalues of Fig. 4: when polarization leakage is included, there is not a clear discrepancy between foreground eigenvalues and cosmological ones as the transition between the two is smoother, the modes are mixed. Therefore, the risk of increasing n_s is to lose progressively more true cosmological signal that leaks in the identified foregrounds $\mathbf{X}^{\rm GMCA}$. We illustrate this last point in Fig. 15, where we plot results for a single channel for two GMCA runs: with $n_s = 6$ on the left-hand column and $n_s = 18$ on the right-hand column. The polarization leakage is the least identified of the foregrounds: it dominates the whole foreground residual (in the bottom panels its pink dashed line C_ℓ completely overlaps the green dashed of the total foreground residual). The recovered 21-cm signal of the $n_s = 6$ case (crosses in top left panel) has an angular power spectrum already off at $\ell \simeq 120$ because of the polarization leakage and (marginally) of the galactic synchrotron left in the residuals maps – further confirmation of the mode mixing. Results are more sound for the $n_s = 18$ case (right-hand panels), although the true 21-cm signal that leaks into the detected foregrounds starts becoming relevant: its corresponding red dashed–dotted line enters in the top panel too, where the input signal and GMCA residual live.

Clearly, in this more realistic scenario, we are not anymore able to identify the optimal n_s just by looking at the behaviour of the recovered power spectra, as we did in Section 5.2 for the simplified scenario with no leakage. Moreover, we are not assured that by arbitrarily looking for higher numbers of sources n_s we have converging results.

Nevertheless, even if the information retained is more compromised when including a polarization leakage component, the resulting bias both in R_ℓ and R_ν is tractable and can be modelled because of its flatness within a range of scales (Cunnington et al. 2020). Overall a compromise has to be looked for, aiming at maximizing the foreground identification and minimizing the loss of true signal. This choice should also depend on the scope of the experiment: it is better to overestimates the signal for detecting the 21-cm emission in cross-correlation with other cosmic tracers, whereas it is important to perform a more aggressive cleaning when aiming for an autocorrelation detection.

We have attempted improving the cleaning in the presence of the polarization leakage, for instance by imposing one column of the mixing matrix[8] or additionally whitening the data. We do not report any substantial improvements and therefore we choose to not present those results here. We postpone to another study a more in-depth and dedicated analysis aimed at identifying sources that are non-smooth in frequency as the polarization leakage, by using more sophisticated versions of GMCA (e.g. L-GMCA; Bobin et al. 2013, 2015) or abandoning the full-blind strategy and imposing extra priors, either on the signal or on the contaminants.

5.6 Comparison with independent component analysis

For comparison, in this section we test another foreground cleaning algorithm on the same simulated data. From the currently available and tested methods, we pick the independent component analysis – in particular the algorithm proposed by Hyvarinen (1999), FastICA – that has recently been used on 21-cm intensity mapping real data by Wolz et al. (2017). In contrast to the GMCA algorithm, which seeks sparse sources in the wavelet domain, the FastICA algorithm looks for statistically independent components by favouring the estimation of non-Gaussian components.

We run FastICA[9] on the reference $n_{ch} = 200$ simulation full sky, with and without the inclusion of polarization leakage. Results are in Fig. 16, where we plot the relative difference in angular and radial power spectrum, R_ℓ and R_ν, between the residuals of the FastICA analysis and the true $\mathbf{C} + \mathbf{N}$. In the scenario without polarization leakage (left-hand panels), the amplitude of the bias achieved in R_ℓ is overall in agreement with that obtained with GMCA (left-hand panel of Fig. 7), however, the striking difference is the behaviour of R_ℓ as function of the angular scale ℓ. For instance, by setting to four the number of independent components (orange dashed line), the resulting average bias in angular power spectrum is of order $\sim$3 per cent at large scales, rapidly falls off for increasing ℓ and reaches -12 per cent at $\ell \approx 170$, meaning that FastICA underestimates the true signal at large scales and greatly overestimates it at small scales. We can draw similar conclusions for the scenario with polarization leakage: comparing the right-hand panel of Fig. 16 with the GMCA results in the left-hand panel of Fig. 14: FastICA reaches similar levels of bias in C_ℓ as GMCA, but the relative difference R_ℓ has a more complicated angular scale dependence, which makes results harder to interpret and, eventually, foreground cleaning effects harder to model. Concerning the radial direction, in the scenario with no polarization leakage (bottom left-hand panel of Fig. 16) FastICA needs five independent components to reach a scale-independent R_ν, which has amplitude of 30 per cent, and with the inclusion of polarization leakage (bottom right), R_ν displays overall the same levels as for the GMCA reconstructed maps (bottom left of Fig. 14).

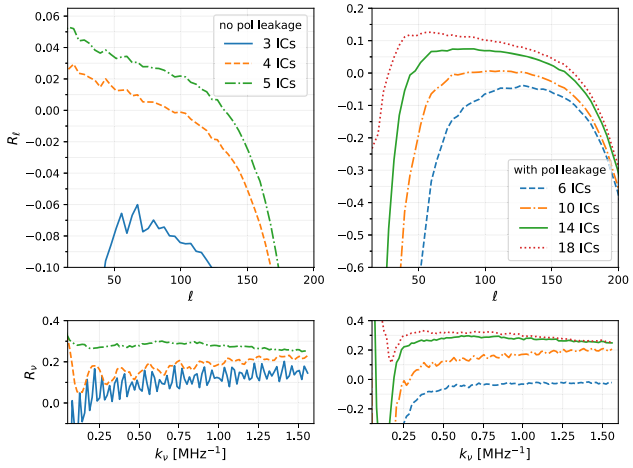

Figure 16. Relative difference in angular and radial power spectrum (top and bottom rows, respectively) between the FastICA recovered signal and the input cosmological signal and instrumental noise for a 200-channel full-sky simulated data that include (do not include) polarization leakage in the left-hand (right-hand) panels. Different lines correspond to a different number of independent components FastICA have identified. The left-hand (right-hand) panels correspond to its GMCA counterpart in the left-hand panels of Fig. 7 (right-hand panels of Fig. 14), same colour coding.

Interestingly, we find a salient difference with respect of GMCA in the RFI-affected scenario. We run FastICA on the same cropped data cube as described previously in Section 5.3; results are in Fig. 17, with the same colour coding of the GMCA counterpart in Fig. 11. Again, the R_ℓ quantity is much more scale dependent for the residuals obtained with FastICA. Setting the number of independent components to four – which has been proven optimal in the non-RFI-contaminated case – gives different R_ℓ curves for the different RFI scenarios; setting the components to five leads to more consistent results, however, the strong dependence on angular scale is still present. Concerning the radial direction, FastICA yields to R_ν that are higher ($\approx$35 per cent) than what obtained with GMCA (few per cent); moreover, for the symmetric RFI scenario, R_ν is scale dependent even when increasing the number of independent components to five.

6 CONCLUSIONS AND PERSPECTIVES

The purpose of this work is investigating the foreground cleaning of 21-cm intensity mapping data performed with the GMCA algorithm, assessing how much information we can recover in terms of the 21-cm field power spectrum. We use a full-sky simulation of the sky in the 900–1400 MHz frequency range composed of the 21-cm signal, the expected astrophysical foregrounds, a polarization

[8]Setting it equal to the galactic free–free spectral index, for which there is greater consensus in the community on its expected value at these frequencies (Bennett et al. 1992).

[9]scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html

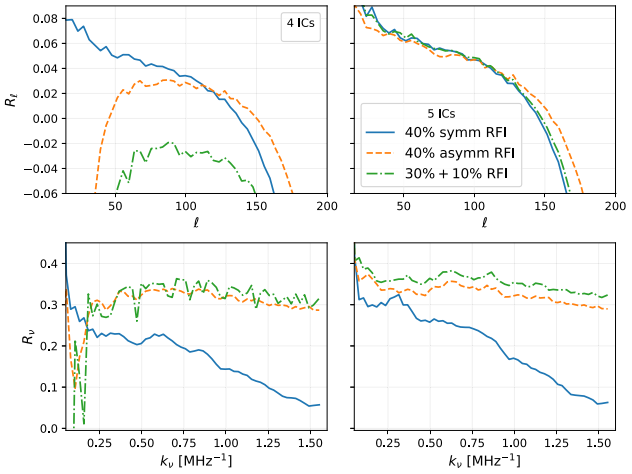

Figure 17. Relative difference in angular and radial power spectrum (top and bottom rows, respectively) between the FastICA recovered signal and the input 21-cm signal in RFI-compromised scenarios (different line styles). On the left-hand (right-hand) panels, results refer to a FastICA run set to find four (five) independent components. Frequency-incomplete data compromise the FastICA foreground separation. Concerning the angular scales (top panels), setting to five the number of independent components results in having cleaned maps independent of the RFI scenario, although the bias in the angular power spectrum is highly scale dependent. For the information in the radial direction, its recovery does not improve with the increase in number of independent components.

leakage component, the smoothing due to the telescope beam, and the thermal noise of the instrument. We hereby summarize our main findings.

(i) When polarization leakage is not included, we find $n_s = 3$ components appropriate for the GMCA cleaning, leading to residuals that underestimate the ground truth angular power spectrum by $\lesssim$2 per cent (channel average) and reproduce at sub-per cent level the radial power spectrum for $k_\parallel \gtrsim 0.02\,h$ Mpc^{-1}.

(ii) When we increase the complexity of the simulation, higher number of sources n_s are needed, and results convergence with increasing n_s is not assured.

(iii) Including polarization leakage and adopting $n_s = 14$ sources, the angular power spectrum is recovered with a scale-independent $\approx$7 per cent bias for scales $\ell > 75$ and the radial counterpart with a scale-independent 20–30 per cent bias for scales $k_\parallel \gtrsim 0.1\,h$ Mpc^{-1}.

(iv) The latter biases improve if we mask the sky region where the adopted polarization leakage component has the most fluctuating behaviour in frequency.

(v) The GMCA source separation benefits from using the highest number of channels available. That is to say, for a fix bandwidth of the experiment, it has to be privileged the thinnest binning possible.

(vi) The GMCA cleaning benefits when it runs on the available data for the full range in frequency, rather than partitioning the data in smaller chunks.

(vii) The latter still holds for incomplete data cubes, i.e. GMCA performance does not deteriorate for RFI-contaminated data.

(viii) The GMCA source separation does not benefit from masking the sky regions where foregrounds are stronger. For instance, the foreground removal is not less successful in the galactic plane region.

The latter point implies that no data are wasted a posteriori and that, at the planning stage, experiments do not need to take into account foreground avoidance for designing the survey footprint, letting focus be rather on issues as overlaps with other samples for cross-correlation and validation purposes, commensality, and so on.

As said, when dealing with polarization leakage, cleaning improves when knowing and masking the pixels where this component has a fluctuating temperature contribution as function of frequency. However, this result depends on the model we have adopted for the leakage. Work is needed for a more physical-motivated polarization leakage model, built upon more recent diffuse polarized emission data and galactic magnetic field structure data (e.g. extending the work by Spinelli, Bernardi & Santos 2018 to the frequencies of interest).

To our knowledge, this is the first work that studies the possibility of a blind removal for a troublesome foreground component as the polarization leakage. More is still to be done and many are the perspectives of this work. We plan to keep adapting GMCA for better dealing with the leakage and also with other sources of systematics that we did not tackle in this work, as for instance a more realistic telescope beam that generates mode mixing in the data and satellite contamination.

For comparison, we also run the FastICA algorithm on the same data cubes. We can appreciate that – with respect to FastICA – GMCA provides results overall more consistent in scale independence and handles RFI-contaminated data better. More exhaustive comparisons are beyond the scope of this paper. GMCA has already been compared with a Gaussian process regression method on EoR-like data by Mertens, Ghosh & Koopmans (2018), who applied GMCA in the Fourier domain. The key assumption underlying the GMCA algorithm is the sparsity of the components to extract in a given domain and, whether it is for EoR or for $z < 6$ science, the foregrounds are smoothly distributed in the Fourier domain and therefore not sparse at all. This is why in this work we prefer a wavelet-based representation to better model the sought-after foregrounds. In short, applying GMCA in a signal representation where the components to be extracted are not sparse is very likely to be less effective, leading to poor separation results. More comparisons between GMCA and other separation methods have been done in the cosmic microwave background context; Leach et al. (2008) offer an exhaustive review.

In this work, we have proven that the number of sources needed for the cleaning sharply increases with the complexity of the simulated data, and this holds for any blind foreground removal method that assumes data can be linearly decomposed in a fixed number of components as in equation (2), e.g. also for FastICA. It is thus important for the community to start testing cleaning algorithms on the most realistic simulations possible. This conclusion does not come unexpectedly as we are aware that in real data analysis, the number of components that is removed is usually higher than what suggested and quoted in simulation papers (e.g. in the most recent analysis Wolz et al. 2017 use 10 and 20 independent components with ICA on GBT data, Anderson et al. 2018 use 10 modes with the SVD method with Parkes data; moreover, in both analysis maps are first resmoothed to further lower resolutions to mitigate the polarization leakage). It is timely to assess the different systematics in simulations to understand what is at play in the data collected by radio telescopes and to prepare for next surveys.

In this paper, we consider full-sky maps. Ongoing work is dedicated to smaller patches and different sky regions, where we expect different foreground contributions and different levels of sparsity that GMCA can rely on. Also depending on the resolution one works with, the sparsity of foregrounds may not always be an appropriate assumption.

Concerning the beam and the noise choices, in this work we consider a single-dish experiment with characteristics of a radio telescope like the MeerKAT. Nevertheless our analysis is meaningful for other experimental set-ups, also including interferometry-driven 21-cm intensity mapping experiment as CHIME,[10] Tianlai,[11] HIRAX[12] or the proposed PUMA[13]: as the decGMCA version of the algorithm performs deconvolution at the same time as the source separation (Jiang et al. 2017; Carloni Gertosio 2020), it is possible to work directly with the visibility data. This constitutes another interesting line of work.

In this paper, we did not consider the effects a GMCA cleaning would have on cosmological analysis, as we mainly focused on a comparison at the maps/data cubes level; we leave this for future work.

For reproducing the results of this paper, we make available demonstration scripts and notebooks[14] together with the main simulated maps.[15]

ACKNOWLEDGEMENTS

IPC thanks Marta Spinelli for insightful discussions and Mario Santos and Jingying Wang for MeerKLASS RFI information. This work is supported by the European Union through the grant LENA (ERC StG no. 678282) within the H2020 Framework Program.

Software used: NumPy (Oliphant 2006), healpy (Zonca et al. 2019), scikit-learn (Pedregosa et al. 2011), hickle (Price et al. 2018), and matplotlib (Hunter 2007).

DATA AVAILABILITY

The simulated data underlying this paper are publicly available in Zenodo, at http://doi.org/10.5281/zenodo.3991818.

REFERENCES

Alonso D., Ferreira P. G., Santos M. G., 2014, MNRAS, 444, 3183
Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015, MNRAS, 447, 400
Alonso D., Sanchez J., Slosar A., LSST Dark Energy Science Collaboration, 2019, MNRAS, 484, 4127
Anderson C. J. et al., 2018, MNRAS, 476, 3382
Ansari R. et al., 2012, A&A, 540, A129
Asad K. M. B. et al., 2019, MNRAS, preprint (arXiv:1904.07155)
Asorey J., Crocce M., Gaztañaga E., Lewis A., 2012, MNRAS, 427, 1891
Battye R. A., Davies R. D., Weller J., 2004, MNRAS, 355, 1339
Battye R. A., Browne I. W. A., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, MNRAS, 434, 1239
Bennett C. L. et al., 1992, ApJ, 396, L7
Bigot-Sazy M.-A. et al., 2015, MNRAS, 454, 3240
Blake C., 2019, MNRAS, 489, 153
Bobin J., Starck J.-L., Fadili J., Moudden Y., 2007, IEEE Trans. Image Processing, 16, 2662
Bobin J., Starck J. L., Sureau F., Basak S., 2013, A&A, 550, A73
Bobin J., Sureau F., Starck J. L., Rassat A., Paykari P., 2014, A&A, 563, A105
Bobin J., Rapin J., Larue A., Starck J.-L., 2015, IEEE Trans. Signal Processing, 63, 1199
Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21
Carloni Gertosio R., 2020, preprint (arXiv:2009.03606)
Carretti E. et al., 2019, MNRAS, 489, 2330
Carucci I. P., Irfan M. O., Bobin J., 2020, 21 cm Intensity Mapping: A 900-1300 MHz Full-Sky Simulation. Available at: https://doi.org/10.5281/zenodo.3991818
Chang T.-C., Pen U.-L., Peterson J. B., McDonald P., 2008, Phys. Rev. Lett., 100, 091303
Chang T.-C., Pen U.-L., Bandura K., Peterson J. B., 2010, Nature, 466, 463
Coles P., Jones B., 1991, MNRAS, 248, 1
Crighton N. H. M. et al., 2015, MNRAS, 452, 217
Cunnington S., Wolz L., Pourtsidou A., Bacon D., 2019, MNRAS, 488, 5452
Cunnington S., Pourtsidou A., Soares P. S., Blake C., Bacon D., 2020, MNRAS, 496, 415
Dickinson C., Davies R. D., Davis R. J., 2003, MNRAS, 341, 369
Fernández X. et al., 2016, ApJ, 824, L1
Flöer L., Winkel B., Kerp J., 2014, A&A, 569, A101
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Harper S. E., Dickinson C., 2018, MNRAS, 479, 2024
Harper S. E., Dickinson C., Battye R. A., Roychowdhury S., Browne I. W. A., Ma Y. Z., Olivari L. C., Chen T., 2018, MNRAS, 478, 2416
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Hyvarinen A., 1999, IEEE Trans. Neural Networks, 10, 626
Irfan M. O., Bobin J., 2018, MNRAS, 474, 5560
Jiang M., Bobin J., Starck J.-L., 2017, SIAM J. Imaging Sci., 10, 1997
Joseph R., Courbin F., Starck J. L., 2016, A&A, 589, A2
Leach S. M. et al., 2008, A&A, 491, 597
Liao Y.-W., Chang T.-C., Kuo C.-Y., Masui K. W., Oppermann N., Pen U.-L., Peterson J. B., 2016, ApJ, 833, 289
Liu A., Tegmark M., 2011, Phys. Rev. D, 83, 103006
Loeb A., Wyithe J. S. B., 2008, Phys. Rev. Lett., 100, 161301
Martin A. M., Giovanelli R., Haynes M. P., Guzzo L., 2012, ApJ, 750, 38
Masui K. W. et al., 2013, ApJ, 763, L20
Mertens F. G., Ghosh A., Koopmans L. V. E., 2018, MNRAS, 478, 3640
Miville-Deschênes M.-A., Ysard N., Lavabre A., Ponthieu N., Macías-Pérez J. F., Aumont J., Bernard J. P., 2008, A&A, 490, 1093
Montanari F., Durrer R., 2012, Phys. Rev. D, 86, 063503
Moore D. F., Aguirre J. E., Parsons A. R., Jacobs D. C., Pober J. C., 2013, ApJ, 769, 154
Noterdaeme P. et al., 2012, A&A, 547, L1
Obuljen A., Alonso D., Villaescusa-Navarro F., Yoon I., Jones M., 2019, MNRAS, 486, 5124
Offringa A. R., Smirnov O., 2017, MNRAS, 471, 301
Oliphant T. E., 2006, A Guide to NumPy. Trelgol Publishing, USA
Olivari L. C., Remazeilles M., Dickinson C., 2016, MNRAS, 456, 2749
Olivari L. C., Dickinson C., Battye R. A., Ma Y.-Z., Costa A. A., Remazeilles M., Harper S., 2018, MNRAS, 473, 4242
O'Neil K., 2002, in Stanimirovic S., Altschuler D., Goldsmith P., Salter C., eds, ASP Conf. Ser. Vol. 278, Single-Dish Radio Astronomy: Techniques and Applications. Astron. Soc. Pac., San Francisco, p. 293
Oppermann N. et al., 2012, A&A, 542, A93
Patil A. H. et al., 2017, ApJ, 838, 65
Pedregosa F. et al., 2011, J. Machine Learning Res., 12, 2825
Picquenot A., Acero F., Bobin J., Maggi P., Ballet J., Pratt G. W., 2019, A&A, 627, A139
Price D. et al., 2018, J. Open Source Software, 3, 1115
Santos M. G., Cooray A., Knox L., 2005, ApJ, 625, 575
Santos M. et al., 2015, Proc. Sci., Cosmology with a SKA HI intensity Mapping Survey. Sissa, Trieste, PoS(AASKA14)019
Santos M. G. et al., 2017, preprint (arXiv:1709.06099)
Shaw J. R., Sigurdson K., Pen U.-L., Stebbins A., Sitwell M., 2014, ApJ, 781, 57
Shaw J. R., Sigurdson K., Sitwell M., Stebbins A., Pen U.-L., 2015, Phys. Rev. D, 91, 083514
Spinelli M., Bernardi G., Santos M. G., 2018, MNRAS, 479, 275

[10] https://chime-experiment.ca
[11] http://tianlai.bao.ac.cn
[12] https://hirax.ukzn.ac.za
[13] https://www.puma.bnl.gov
[14] https://github.com/isab3lla/gmca4im
[15] http://doi.org/10.5281/zenodo.3991818

Spinelli M., Zoldan A., De Lucia G., Xie L., Viel M., 2020, MNRAS, 493, 5434
Starck J.-L., Fadili J., Murtagh F., 2007, IEEE Trans. Image Processing, 16, 297
Switzer E. R. et al., 2013, MNRAS, 434, L46
Switzer E. R., Chang T. C., Masui K. W., Pen U. L., Voytek T. C., 2015, ApJ, 815, 51
Villaescusa-Navarro F., Alonso D., Viel M., 2017, MNRAS, 466, 2736
Villaescusa-Navarro F. et al., 2018, ApJ, 866, 135
Wolz L., Abdalla F. B., Blake C., Shaw J. R., Chapman E., Rawlings S., 2014, MNRAS, 441, 3271
Wolz L. et al., 2017, MNRAS, 464, 4938
Zafar T., Péroux C., Popping A., Milliard B., Deharveng J. M., Frank S., 2013, A&A, 556, A141
Zhang L., Bunn E. F., Karakci A., Korotkov A., Sutter P. M., Timbie P. T., Tucker G. S., Wandelt B. D., 2016, ApJS, 222, 3
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Software, 4, 1298

This paper has been typeset from a TEX/LATEX file prepared by the author.