GNILC

Lucas Collis Olivari

December 4, 2018

1 Using the Software

To use the software just type in a terminal: 'python gnilc.py'. The parameters of the code are defined in 'parameter.ini'. In it, you will need to define: the name of the fits file where you have the input maps (located in the input directory and a channels x Healpix pixels matrix), the name of the file where you have the prior maps (located in the input directory and a channels x Healpix pixels matrix), name of the file where you have the mask used (located in the input directory and a Healpix pixels vector), the needlet bands you want to use, their bandcenters, the ILC bias, and a few extra things about the output.

The prior can be the cosmological signal or the cosmological signal plus the thermal noise, with the latter being preferred over the former.

The output files will be saved in the output directory.

Finally, a few **rules** should be followed by the users:

i) This software should be used only for members of the BINGO collaboration.

ii) If you ever use this software in your BINGO research, please cite Remazeilles M., Delabrouille J., Cardoso J.-F., 2011, MNRAS, 418, 467 and Olivari L. C., Remazeilles M., Dickinson C., 2016, MNRAS, 456, 2749. Also, as a matter of courtesy, before publishing any work in which you have used GNILC, please send an email (you can CC me) to Mathieu Remazeilles saying that you are using GNILC for a BINGO related work and asking whether he has any objection about it.

iii) Never make this code public.

1.1 GNILC Weights

If you have chosen to save the GNILC weights, you can apply them to an ensemble of maps (see Section 4 below for an explanation of why you would like to do this). To do so, choose your parameters in 'parameters_weights.ini' and type in a terminal: 'python gnilc_weights.py'.

2 Pedagogical Description of the Method

The GNILC method has been extensively discussed within the IM context in Olivari et al. 2016 and in my thesis (my thesis can be found here: https://www.research.manchester.ac.uk/portal/en/theses/intensity-mapping-a-new-approach-to-probe-the-largescale-structure-of-the-universe(cd5b7586-7210-441e-838f-545d397893e5).html). I this and the next section, I only provide a summary of the method.

The component separation problem is the same for IM and CMB. The only difference is that we have a 'prior knowledge' about the spectral properties of the CMB and we can use this knowledge as an external constraint to the problem. The same does not happen for IM since we observe spectral lines in this case. The power spectrum of the cosmological signal, however, is easier to predict (it is just a tracer of the dark matter power spectrum), and is distinct from the power spectrum of the foregrounds. The use of spatial information (power spectrum) instead of spectral information (frequency spectrum) is our proposition to separate the HI signal with unknown spectral signature from the foregrounds.

In the GNILC method this is done in the following way. First, using an estimate of the HI power spectra, we simulate HI maps and use them to determine the ratio between the cosmological and the total signal (for simplicity, we will ignore the thermal noise in this brief description of the method). This information is then used to 'divide' the observed data covariance matrix into two sub-matrices: a foreground-dominated sub-matrix and an HI-dominated sub-matrix. This allows us to remove from the resulting analysis the eigenvectors (i.e., the 'effective vector basis') that are dominated by the foregrounds, which is equivalent of saying that we have found the HI subspace of the covariance matrix. We note that the covariance matrix is calculated with the frequency maps, so it corresponds to an $n_{\rm ch} \times n_{\rm ch}$ matrix, where $n_{\rm ch}$ is the number of frequency channels of the experiment.

In this step of the method, the HI subspace is estimated both in space (i.e., in the pixel space) and in angular scale (i.e., in the multipole space). This means that, before calculating the covariance matrix, we apply some 'multipole filters' (these filters are called needlets) to our frequency maps. This gives us access to the way that the ratio between the HI and total signal changes with multipoles. Moreover, when calculating the observation covariance matrices (we have one for each of our 'multipole filters'), we do it several times, always choosing different sets of pixels in the process.

After we have found the HI sub-matrices for each multipole range and group of pixels, we use them to create filters (more precisely, we create Internal Linear Combination filters) that we apply to the observed maps. For construction, these filters minimize the foreground signal while maintaining the HI signal unchanged. Of course, this process is not perfect and some HI signal may be lost together with the foregrounds and/or some foregrounds may remain in the reconstructed HI signal.

3 Technical Description of the Method

In what follows, we describe the method assuming the user wants to recover the HI plus noise signal, which is preferred over the choice of recovering the HI signal alone.

The GNILC method can be divided into two main steps. First, using an estimate of the HI plus noise power spectra, we determine the local ratio between the HI plus noise signal and the total signal and perform a 'constrained' PCA of the observed data to determine the effective dimension of the HI plus noise subspace. In this step, the number of principal components (which corresponds to the foregrounds) of the observation covariance matrix is estimated locally both in space and in angular scale by using a wavelet (needlet) decomposition of the observations. We also use a statistical information criterion (AIC) to select the principal components of the observation covariance matrix (i.e., the subset of eigenvectors spanning the foreground subspace). Second, using only the eigenspace (i.e., the effective dimensions) that refer to the HI plus noise subspace, we apply multidimensional ILC filters to the observed maps and recover the HI plus noise signal for each of them.

The GNILC method can be summarized as follows:

1) To isolate the different ranges of (angular) scales, we first define a set of windows in harmonic space, which we call needlets. These needlets work as band-pass filters, isolating certain scales of interest. The spherical harmonic coefficients, $a_{\ell m}$, of the observed frequency maps are then band-pass filtered by the previously defined needlets. This means that for each frequency map, we have several needlet maps.

2) For each range of scales (or needlets) j, we compute the data covariance matrix, at pixel p, of a pair of frequencies a and b as

$$\widehat{\mathbf{R}}_{ab}(p) = \sum_{p' \in \mathcal{D}(p)} \boldsymbol{x}_a(p')\boldsymbol{x}_b^T(p'), \tag{1}$$

where $\mathcal{D}$ is a domain of pixels centred around the pixel p.

3) For each range of scales j, we also compute the HI plus noise covariance matrix by using HI plus noise maps $\boldsymbol{y}$ simulated from some estimator of the HI plus noise angular power spectra,

$$\widehat{\mathbf{R}}_{s\,ab}(p) = \sum_{p' \in \mathcal{D}(p)} \boldsymbol{y}_a(p')\boldsymbol{y}_b^T(p'). \tag{2}$$

Note that this estimator (or prior) is blind about the particular realization of the HI plus noise signal that is found in the observed (here, simulated) sky.

4) We diagonalize the transformed data covariance matrix, $\widehat{\mathbf{R}}_s^{-1/2}\widehat{\mathbf{R}}\widehat{\mathbf{R}}_s^{-1/2}$ as

$$\widehat{\mathbf{R}}_s^{-1/2}\widehat{\mathbf{R}}(m)\widehat{\mathbf{R}}_s^{-1/2} = \mathbf{U}_N\mathbf{D}_N\mathbf{U}_N^T + \mathbf{U}_S\mathbf{U}_S^T, \tag{3}$$

where $\mathbf{D}_N$ collects the m largest eigenvalues, $\mathbf{U}_N$ the corresponding eigenvectors, and $\mathbf{U}_S$ the $(n_{\rm ch} - m)$ eigenvectors related to the HI plus noise emission subspace. This is the PCA step of the algorithm.

The effective dimension m of the foreground subspace (number of principal components), however, differently from traditional PCA, is estimated by minimizing the AIC,

$$\min\left(2m + \sum_{i=m+1}^{n_{\rm ch}} [\mu_i - \log\mu_i - 1]\right) \quad \text{with} \quad m \in [1, n_{\rm ch}],$$

where μ_i are the eigenvalues of $\widehat{\mathbf{R}}_s^{-1/2}\widehat{\mathbf{R}}\widehat{\mathbf{R}}_s^{-1/2}$.

5) For each range of scales j, we apply an $(n_{\rm ch} - m)$-dimensional ILC filter to the observed data,

$$\begin{aligned}\widehat{\boldsymbol{s}}^{(j)} &= \mathbf{W}\mathbf{x}^{(j)} \\ &= \widehat{\mathbf{S}}(\widehat{\mathbf{S}}^T\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{S}})^{-1}\widehat{\mathbf{S}}^T\widehat{\mathbf{R}}^{-1}\boldsymbol{x}^{(j)},\end{aligned} \tag{4}$$

where the estimated HI plus noise mixing matrix is given by

$$\widehat{\mathbf{S}} = \widehat{\mathbf{R}}_s^{1/2}\mathbf{U}_S. \tag{5}$$

6) Finally, we synthesize the reconstructed needlet HI plus noise maps $\widehat{\boldsymbol{s}}^{(j)}$ into one map. First, the needlet maps are transformed to spherical harmonic space, their harmonic coefficients are again band-pass filtered by the respective needlet window (this step is necessary to put the reconstructed signal in the same unit of the observed data), and the filtered harmonic coefficients are transformed back to real space maps. This operation gives one reconstructed HI plus noise map per needlet scale and per frequency channel. Second, for each frequency channel, the needlet maps are added to give the complete GNILC reconstructed HI plus noise map.

4 Correcting for the Component Separation Bias

The component separation bias, although relatively small, can be seen as a hindrance for precision cosmology. In the case of GNILC, however, this bias can be estimated and corrected for in the following way. One of the outputs of the method is the ILC filters (weight matrices, see Eq. 4) that are used to reconstruct the target signal. If we apply the ILC filters to an ensemble of realisations of the HI plus noise signal generated from our prior HI plus noise power spectrum and calculate the mean power spectrum, the result will be an estimate of the GNILC reconstructed power spectrum. The difference between the true GNILC reconstructed power spectrum and its estimate will be the fluctuations due to the cosmic variance (we cannot predict the particular realisation of the HI plus noise signal that is found in the data) and the foreground residuals.

The procedure (for simplicity, I use the $P(k)$ in the following, but the logic is the same for C_ℓ) for correcting for the component separation bias is the following. If P_S is the true (input) HI plus noise power spectrum, P_R is the reconstructed HI plus noise power spectrum, P_E is the mean power spectrum of an ensemble of realisations of the HI plus noise signal based on the prior power spectrum, and P_{WE} is the mean power spectrum of the GNILC filters applied to this ensemble of realisations, then the following is true:

$$P_S \simeq P_R + (P_E - P_{WE}) \tag{6}$$

4.1 Marginalising over the Component Separation Bias

The outputs of the GNILC method are two: the reconstructed HI plus noise maps and the GNILC filters that are used for their reconstruction. To properly fit for the astrophysical or cosmological parameters, both should be used. For reasonable priors (for example, a prior that either over or underestimate the true signal by less than 25%), they will be both driven by the data and not by the prior, i.e. they will be largely independent of the assumed prior. Consequently, when fitting for the parameters in a Monte-Carlo analysis, P_{WT} should be compared to the reconstructed (observed) HI plus noise power spectrum, P_R, instead of comparing directly P_R with P_T, where P_{WT} here refers to the power spectrum of a theoretical set of maps applied to the GNILC filters and P_T to the theoretical HI plus noise power spectrum.