Cross-correlation Forecast of CSST Spectroscopic Galaxy and MeerKAT Neutral Hydrogen Intensity Mapping Surveys

Authors: Yan Gong, Yan Gong

Date: 2023-07-28T00:00:00+00:00

Abstract

Cross-correlating the data on neutral hydrogen (H I) 21 cm intensity mapping with galaxy surveys is an effective method to extract astrophysical and cosmological information. In this work, we investigate the cross-correlation of MeerKAT single-dish mode H I intensity mapping and China Space Station Telescope (CSST) spectroscopic galaxy surveys. We simulate a survey area of 300 deg2 of MeerKAT and CSST surveys at z = 0.5 using Multi-Dark N-body simulation. The PCA algorithm is applied to remove the foregrounds of H I intensity mapping, and signal compensation is considered to solve the signal loss problem in H I-galaxy cross power spectrum caused by the foreground removal process. We find that from CSST galaxy auto and MeerKAT-CSST cross power spectra, the constraint accuracy of the parameter product ΩH IbH IrH I,g can reach 1%, which is about one order of magnitude higher than the current results. After performing the full MeerKAT H I intensity mapping survey with 5000 deg2 survey area, the accuracy can be enhanced to <0.3%. This implies that the MeerKAT-CSST crosscorrelation can be a powerful tool to probe the cosmic H I property and the evolution of galaxies and the Universe.

Full Text

Preamble

Cross-correlation Forecast of CSST Spectroscopic Galaxy and MeerKAT Neutral Hydrogen Intensity Mapping Surveys

Yu-Er Jiang[12], Yan Gong[134], Meng Zhang[12], Qi Xiong[12], Xingchen Zhou[12], Furen Deng[12], Xuelei Chen[1256], Yin-Zhe Ma[478], and Bin Yue[1]

[1] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; gongyan@bao.ac.cn

[2] School of Astronomy and Space Sciences, University of Chinese Academy of Science (UCAS), Beijing 100049, China
[3] Science Center for China Space Station Telescope, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[4] NAOC-UKZN Computational Astrophysics Centre (NUCAC), University of KwaZulu-Natal, Durban, 4000, South Africa
[5] Center for High Energy Physics, Peking University, Beijing 100871, China
[6] School of Chemistry and Physics, University of KwaZulu-Natal, Westville Campus, Private Bag X54001, Durban 4000, South Africa
[7] Department of Physics, Stellenbosch University, Matieland 7602, South Africa

Received 2023 March 20; revised 2023 April 4; accepted 2023 April 11; published 2023 June 9

Abstract

Cross-correlating neutral hydrogen (H I) 21 cm intensity mapping data with galaxy surveys is an effective method for extracting astrophysical and cosmological information. In this work, we investigate the cross-correlation of MeerKAT single-dish mode H I intensity mapping and China Space Station Telescope (CSST) spectroscopic galaxy surveys. We simulate a survey area of 300 deg^2 for MeerKAT and CSST surveys at z = 0.5 using the MultiDark N-body simulation. The PCA algorithm is applied to remove foregrounds from the H I intensity mapping data, and signal compensation is implemented to address the signal loss problem in the H I-galaxy cross power spectrum caused by the foreground removal process. We find that from the CSST galaxy auto and MeerKAT-CSST cross power spectra, the constraint accuracy of the parameter product Ω_H I b_H I r_H I,g can reach 1%, which is about one order of magnitude higher than current results. After performing the full MeerKAT H I intensity mapping survey with a 5000 deg^2 survey area, the accuracy can be enhanced to <0.3%. This implies that the MeerKAT-CSST cross-correlation can be a powerful tool for probing cosmic H I properties and the evolution of galaxies and the Universe.

Key words: (cosmology:) large-scale structure of universe – (cosmology:) cosmological parameters – Cosmology

1. Introduction

Probing the large-scale structure (LSS) of the Universe has always been one of the main missions of cosmological observations. Constraining the properties of dark matter and dark energy, recovering primordial fluctuations, and testing gravity theories all require cosmological surveys with large survey areas and wide redshift coverage. To achieve this goal, line intensity mapping (LIM) has been proposed and proven to be an efficient technique. LIM makes use of emission lines from energy level transitions of atoms or molecules, such as H I 21 cm, C II, CO, Lyα, Hα, [O III], etc. (see, e.g., Visbal & Loeb 2010; Carilli 2011; Gong et al. 2011; Lidz et al. 2011; Gong et al. 2012, 2013; Silva et al. 2013; Gong et al. 2014; Pullen et al. 2014; Uzgil et al. 2014; Silva et al. 2015; Gong et al. 2017; Fonseca et al. 2017; Gong et al. 2020). These lines can reflect different

properties and processes of galaxy evolution and serve as good tracers of the LSS.

Unlike traditional observations targeting resolvable sources, intensity mapping probes the cumulative intensity of all sources in a spatial volume (voxel) defined by survey spatial and frequency resolutions. Even though some sources are too faint to be detected in traditional sky surveys, their signals can be probed in intensity mapping. In addition, the frequency shifts of emission lines provide a natural probe of redshift, so intensity mapping is expected to be a powerful tool for obtaining cosmic 3D matter structure information traced by emission lines from galaxies with high efficiency and relatively low cost. Among various emission lines, the H I 21 cm line from atomic hydrogen is the most widely studied in intensity mapping research (see e.g., Chen 2011, 2012; Battye et al. 2013; Dickinson 2014; Newburgh et al. 2014; Bandura et al. 2014; Santos et al. 2015; Smoot & Debono 2017; Wang et al. 2021; Cunnington et al. 2023; Deng et al. 2022; Zhang et al. 2022; Spinelli et al. 2022; Perdereau et al. 2022). Besides being a main probe of the epoch of reionization, the neutral hydrogen 21 cm line has a tight connection with star formation and galaxy evolution, and it can trace galaxy and hence dark matter distribution at low and high redshifts.

While many experiments on H I intensity mapping have been proposed or are already running, the foreground contamination problem remains one of the biggest challenges, as foregrounds can be as large as five orders of magnitude higher than the signal. The high brightness temperature of Galactic emission and other sources makes the H I signal difficult to detect from auto-correlations. In principle, cross-correlating 21 cm observations with an optical galaxy survey in the same survey area is a good method to reduce foreground contamination and instrumental noise and extract the signal (e.g., Chang et al. 2010). The signal-to-noise ratio (SNR) can be significantly improved since the foregrounds and instrumental noise of different wave bands in different surveys are barely correlated. However, in practice, the cross-correlation results are not fully satisfactory due to the complex components of the foreground, so foreground removal algorithms are still needed in cross-correlations. Various algorithms have been applied, including blind foreground removal like principal component analysis (PCA) (Davis et al. 1985) and independent component analysis (ICA) (Wolz et al. 2014) which exploit different frequency smoothness of foreground and signal, polynomial/parametric-fitting methods which fit the physical properties of the foreground (Bigot-Sazy et al. 2015), machine learning (ML) methods (Li & Wang 2022), etc. Although signal loss and foreground residual are usually inevitable, foreground removal techniques do make progress and are necessary for cross-correlation detection.

Currently, positive results on H I abundance and H I-galaxy correlation have been obtained by several experiments. The Green Bank Telescope (GBT) implemented H I intensity mapping correlation detection with the Deep2 optical redshift survey (Chang et al. 2010), WiggleZ Dark Energy Survey (Masui et al. 2013), and eBOSS survey (Wolz et al. 2022). In addition, the Parkes radio

telescope also presented work on correlating H I intensity mapping with the 2dF galaxy survey (Anderson et al. 2018). Recently, MeerKAT accomplished H I intensity mapping correlation detection with the WiggleZ survey (Cunnington et al. 2023). They all constrain the H I-galaxy correlation parameter product Ω_H I b_H I r_H I,g at different redshifts, where Ω_H I, b_H I, and r_H I,g are the H I energy density parameter, H I bias, and correlation coefficient of H I and galaxy, respectively. In this work, we determine the constraint power on neutral hydrogen parameters by observations of MeerKAT and the next-generation galaxy survey of China Space Station Telescope (CSST).

MeerKAT is a pathfinder project of the Square Kilometre Array (SKA) and will become part of SKA-mid in the future (Santos et al. 2017; Bacon et al. 2020). It is a state-of-the-art intensity mapping instrument capable of complementing and extending cosmological measurements across a wide range of wavelengths. While MeerKAT is a large interferometric array that can access small scales of cosmic structure, single-dish mode is preferred in intensity mapping experiments. We plan to perform MeerKAT H I intensity mapping cross-correlation with the China Space Station Optical Survey (CSS-OS) (Zhan 2011; Cao et al. 2018; Gong et al. 2019; Zhan 2021). CSS-OS is the major observation project of CSST, covering 17,500 deg^2 of sky area over a 10-year working period. In addition, the spectroscopic survey of CSS-OS will provide a large amount of data in the form of a galaxy catalog with verified redshift using slitless gratings. CSST is planned to start observations around 2024, while MeerKAT will continue full-time operations before SKA begins full operations in 2028, and these two surveys will have a large overlapping survey area. Thus we believe MeerKAT H I intensity mapping and CSST galaxy survey would make promising cross-correlation detection in the coming future.

This paper is organized as follows: in Section 2, we introduce our method for creating mock data of MeerKAT H I intensity mapping and CSST spectroscopic galaxy surveys; in Section 3, we apply the PCA algorithm to remove foregrounds in H I intensity maps; in Section 4, we calculate the galaxy auto and H I-galaxy cross power spectra and discuss the signal compensation method for the cross power spectrum; in Section 5 we forecast constraints on relevant cosmological parameters; we conclude our work and provide discussion in Section 6.

2. Mock Data

We generate MeerKAT intensity maps and CSST spectroscopic galaxy survey data using MultiDark cosmological simulations (Klypin et al. 2016). MultiDark is a suite of N-body cosmological simulations carried out with the L-GADGET-2 code. Most simulations in this suite have 3840^3 particles, with box sizes ranging from 250 Mpc/h to 2500 Mpc/h. Based on the survey area and redshift of the MeerKAT observation plan, the Small MultiDark Planck simulation (SMDPL) has been chosen for this work. The box size of SMDPL is 400 Mpc/h, and halos in SMDPL boxes are identified through the halo finding code Friends-of-Friends (FOF) with a relative linking length of 0.2 (Davis et al. 1985).

The relevant simulation and cosmological parameters adopted for SMDPL are listed in Table 1, and its halo catalog can be acquired from the CosmoSim database. In our work, we focus on cosmology at z = 0.5, which is one of the main observational target redshifts for both CSST and the MeerKAT L-band. Our mock data are generated from snapshot70 of SMDPL, whose redshift z 0.5. We find that the 400 Mpc/h box size of snapshot70 corresponds to a survey of 297 deg^2 at z = 0.5. Note that the non-flat sky effect may need to be considered for 300 deg^2 sky coverage, but for simplicity, we still use the flat sky approximation in our mock data analysis.

2.1. H I Intensity Mapping with MeerKAT

Since H I can only survive ultraviolet (UV) radiation in dense clumps in galaxies after the epoch of reionization, we assume that H I can only exist in halos hosting galaxies at z = 0.5. We place the H I mass at the center of a halo, as has been proven reasonable in previous studies (see, e.g., Villaescusa-Navarro et al. 2018). Under this assumption, we construct a catalog applying the halo H I mass function given by Villaescusa-Navarro et al. (2018), which takes the form:

$$\frac{dM_{\rm HI}}{dM} = \alpha \left(\frac{M}{M_0}\right)^{-\beta} \exp\left[-\left(\frac{M_{\rm min}}{M}\right)^{\gamma}\right]$$

Here M is the halo mass, and we have three free parameters—α, M_0, and M_{min}—that determine the shape of the fitting curve at different redshifts. To obtain the values of these three parameters at z = 0.5, we perform interpolation on the fitting values at z = 0 and 1 given in Villaescusa-Navarro et al. (2018). We find that $\alpha = 0.42$, $M_0 = 2.50 \times 10^{10}$ h^{-1} M , and M_{min} = 1.5×10^9 h^{-1} M_ at z = 0.5. In Figure 1 [Figure 1: see original paper], we plot the M_H I–M relation at z = 0.5 (green solid curve), and the relations at z = 0 (blue dashed curve) and 1 (orange dashed curve) from Villaescusa-Navarro et al. (2018) are also shown for comparison.

We can then calculate the H I energy density parameter Ω_H I, expressed as:

$$\Omega_{\rm HI} = \frac{1}{\rho_{c,0}} \int_{M_{\rm min}}^{M_{\rm max}} dM\, n(M,z) \frac{dM_{\rm HI}}{dM}$$

where _{c,0} is the critical density of the present Universe, and n(M,z) is the halo mass function (Sheth & Tormen 1999), which can be derived from our simulation. We find that Ω_H I = 6.73×10^{-4} in our simulation, which agrees with the estimation of the Ω_H I–z relation given in literature (see, e.g., Villaescusa-Navarro et al. 2018).

Next, we create the map of H I brightness temperature. The brightness temperature field δT_b traces the underlying matter fluctuations δ_m as:

$$\delta T_b(\mathbf{r}, z) = \bar{T}_b(z) b_{\rm HI} \delta_m(\mathbf{r}, z)$$

where $\bar{T}$_b(z) is the mean H I brightness temperature at z, and b_H I is the H I bias, which can be estimated by:

$$b_{\rm HI} = \frac{\int_{M_{\rm min}}^{M_{\rm max}} dM\, n(M,z) b(M,z) \frac{dM_{\rm HI}}{dM}}{\int_{M_{\rm min}}^{M_{\rm max}} dM\, n(M,z) \frac{dM_{\rm HI}}{dM}}$$

Here b(M,z) is the halo bias. For a voxel with position on the sky r and redshift z, its H I brightness temperature can be derived as:

$$T_b(\mathbf{r}, z) = \frac{3}{32\pi} \frac{hc^3 A_{10}}{k_B m_H \nu_{21}} \frac{(1+z)^2}{H(z)} \rho_{\rm HI}(\mathbf{r}, z)$$

where E(z) = H(z)/H_0 represents the evolution of the Hubble parameter, and T_0 is a redshift-dependent parameter defined as $T_0 = (3/32\pi)(h\ c^3\ A_{10})/(k\_B\ m\_H\ _{21})$. Then $\bar{T}$_b(z) can be estimated by averaging T_b(r,z) at different positions in the simulation box. At z = 0.5, we find that the corresponding mean H I brightness temperature is $\bar{T}$_b = 0.23 mK in our simulation. The brightness temperature of the H I distribution (right panel) and the corresponding dark matter distribution (left panel) in the simulation are shown in Figure 2 [Figure 2: see original paper].

After obtaining the H I brightness temperature in the simulation box, our next step is to create H I intensity maps with MeerKAT instrumental parameters and observational effects. Since the observable of H I intensity mapping is the H I brightness temperature of each voxel in the survey volume, we divide the survey volume (here meaning our simulation box) into voxels that MeerKAT can observe. The details are as follows:

1. To divide frequency bins along the line of sight (LOS), we place the center of the box at z = 0.5. As the box length of 400 Mpc/h is known, the redshift range of the survey volume can be calculated. We find that the redshift range of snapshot70 is 0.415 0.590, corresponding to the observed H I frequency of 1004.14 MHz 893.30 MHz. This frequency range can be observed by the MeerKAT L-band with frequency resolution of 0.2 MHz, allowing us to divide the survey volume into 554 bins, such that each bin width is about 0.72 Mpc/h. For simplicity, we assume there is no redshift evolution in this range.

2. For pixels perpendicular to the LOS, since we plan to use single-dish mode observation, resolution is defined by the full width at half maximum (FWHM) of the beam of an individual dish. The beam size or spatial resolution is given by:

$$\theta = 1.22 \frac{\lambda_{\rm obs}}{D_{\rm dish}}$$

where λ{obs} *is the observed wavelength, and* D{dish} is the dish aperture diameter. We find that the spatial resolution of MeerKAT at z = 0.5 is 1.36 arcmin. Since the size of a simulation box is 400×400 $(\mathrm{Mpc/h})^2$, corresponding to a 297 deg^2 survey area, the number of pixels in an H I map is found to be 12×12 for MeerKAT single-dish mode observation. We note that the current spatial resolution given by the FWHM of the beam is a choice of simplicity, and more realistic resolution will be considered in future work. Moreover, since the beam size actually changes with frequency, it can introduce more complexity and challenges into foreground subtraction. However, because our simulation snapshot has no redshift evolution, for simplicity we do not consider the frequency dependence of the beam size and set the pixel size of all maps to be the same.

The H I signal intensity map obtained by MeerKAT at z = 0.5 is displayed in the upper left panel of Figure 3 [Figure 3: see original paper]. In real observations, H I intensity mapping will be contaminated by different components, such as system thermal noise, foreground emission from the Milky Way, radio frequency interference (RFI), etc., which can lower the SNR. Here we model the system thermal noise of a single-dish as Gaussian noise. Its root mean square (rms) noise temperature can be calculated as (Bull et al. 2015):

$$\sigma_T = \frac{T_{\rm sys}}{\sqrt{\delta\nu\, t_{\rm tot}}} \frac{\sqrt{A_S}}{A_e}$$

where δ is the frequency interval, t_{tot} is the total observation time of the survey, A_e is the effective collecting area of a dish, A_S is the survey area, and T_{sys} is the system temperature, which is usually described as a combination of four components:

$$T_{\rm sys} = T_{\rm sky} + T_{\rm spill} + T_{\rm atm} + T_{\rm rec}$$

The mean sky temperature can be approximated by T_{sky} = 2.725 + 1.6(/GHz)$^{-2 \cdot 75}$, and T_{spill}, T_{atm}, and T_{rec} represent spillover temperature, atmosphere temperature, and receiver temperature, respectively. The values of these parameters we adopt are listed in Table 2 , and we obtain σ_T = 0.102 mK at = 946.7 MHz (z = 0.5). The corresponding map of Gaussian system noise is shown in the upper right panel of Figure 3.

Foreground emission from the Milky Way is the main challenge for H I intensity mapping. The brightness temperature of foregrounds can be more than four orders of magnitude brighter than the H I signal, so its effect must be seriously considered in our forecast of MeerKAT observations. Here we generate the

foreground emission using the GSM2016 model (Zheng et al. 2017). GSM2016 is an improved model of the original GSM that uses an extended PCA algorithm to identify different components in the diffuse Galactic emission. Six components of Galactic emission matching known physical emission mechanisms are obtained: synchrotron emission, free-free emission, cold and warm dust thermal emission, CMB anisotropy, and Galactic H I emission. This algorithm allows it to use 29 sky maps from 10 MHz to 5 THz and perform interpolation to produce a full-sky map at any frequency in this range.

To apply the foreground model to our H I map, coordinates of the survey area must be set. According to previous work (Wang et al. 2021), we set our survey area at 153°.38 < R.A. < 170°.62 and −5°.62 < decl. < 11°.62, mostly intersecting with the WiggleZ 11 hr field (Drinkwater et al. 2010, 2018). Note that this choice of survey area is only for discussion here, as this region has relatively low Galactic emission in the full-sky map. For future MeerKAT-CSST cross-correlation observations, the target survey area can be chosen from anywhere in the overlapping region of the MeerKAT and CSST survey areas with relatively low foreground emission. We generate foreground maps at each frequency bin and then interpolate them to the center of each voxel.

The foreground map for the survey area at = 946.7 MHz (z = 0.5) is shown in the lower left panel of Figure 3 [Figure 3: see original paper]. We find that the foreground contamination we consider is about four orders of magnitude brighter than the H I signal. After combining the H I signal, foreground emission, and system noise maps, we obtain the total sky map observed by MeerKAT. The mock total observational map is displayed in the lower right panel of Figure 3. Since the H I signal has been completely drowned in the contamination, foreground subtraction algorithms must be applied to extract the H I signal. We discuss the foreground removal method in the next section.

2.2. Galaxy Survey with CSST

We use the same simulation data SMDPL snapshot70 to create mock data for the CSST spectroscopic galaxy survey. We utilize the Python package Halotools to generate a galaxy distribution for each dark matter halo in the simulation. First, the structure of a cold dark matter (CDM) halo can be described by an NFW profile (Navarro et al. 1996). The halo concentration-mass relation under the NFW profile is fitted by Dutton & Macciò (2014):

$$c_{\rm vir}(M_{\rm vir}) = a\left(\frac{M_{\rm vir}}{10^{12}h^{-1}M_{\odot}}\right)^{-b}$$

where M_{vir} is the halo virial mass, c_{vir} is the concentration of the corresponding halo, and a and b are fitting parameters expressed as:

$$a = 0.537\exp(-0.138z)$$

$$b = 0.097 \exp(-0.071z)$$

After obtaining the halo concentration, the halo occupation distribution (HOD) model can be applied to obtain the galaxy distribution. The HOD model determines the population of central and satellite galaxies in a given halo. The central galaxy occupation statistics are given by Zheng et al. (2007):

$$\langle N_{\rm cen}(M_{\rm halo})\rangle = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log_{10} M_{\rm halo} - \log_{10} M_{\rm min}}{\sigma_{\log M}}\right)\right]$$

and central galaxies are assumed to reside at the centers of their host halos. On the other hand, the distribution of satellite galaxies is written as:

$$\langle N_{\rm sat}(M_{\rm halo})\rangle = \langle N_{\rm cen}(M_{\rm halo})\rangle \left(\frac{M_{\rm halo} - M_0}{M_1}\right)^{\alpha}$$

When redshift, cosmological model, and the threshold of galaxy absolute magnitude are set, the values of the parameters in Equations (12) and (13) are calculated by the Halotools package based on the model published in Zheng et al. (2007).

At z = 0.5, we set the threshold of galaxy absolute magnitude to be −19.5, and these parameters are expressed as:

$$\log_{10} M_{\rm min} = 11.88$$
$$\log_{10} M_1 = 13.08$$
$$\log_{10} M_0 = 11.28$$
$$\sigma_{\log M} = 0.25$$
$$\alpha = 1.06$$

The final step in generating a galaxy catalog is to determine which galaxies can be observed by CSST. We assign luminosity to galaxies using the relation between host halo mass and galaxy luminosity given by Vale & Ostriker (2008):

$$\frac{L_{\rm group}}{L_0} = \left(\frac{M_{\rm halo}}{M_\star}\right)^c \left[\frac{1}{2}\left(\frac{M_{\rm halo}}{M_\star}\right)^{a-c} + \frac{1}{2}\left(\frac{M_{\rm halo}}{M_\star}\right)^{b-c}\right]^{-1}$$

For simplicity, assuming all satellite galaxies have the same luminosity, we have:

$$L_{\rm group} = L_{\rm cen} + N_{\rm sat} L_{\rm sat}$$

Here L_{group}, L_{cen}, and L_{sat} are the luminosities of a galaxy group, central galaxy, and satellite galaxy, respectively. N_{sat} is the number of satellite galaxies in a galaxy group. The parameter values are chosen to be L_A = 0.3 L_ , L_0 = 2.8 × 10^9 L_ (Zheng et al. 2007), M_ = $10^{11 \cdot 9}$ h^{-1} M_ , a = 29.78, b = 29.5, and c = 0.0255 (Vale & Ostriker 2008). Then the galaxy luminosity can be converted to magnitudes in CSST spectroscopic bands. Since the magnitude limit of the CSST spectroscopic survey is 23 mag (Gong et al. 2019; Zhan 2021), galaxies brighter than this limit can be selected to form the CSST spectroscopic galaxy survey catalog. After selection, we find that the galaxy number density in the simulation box is 9.07 × 10^{-3} $(Mpc/h)^{-3}$, which is in good agreement with results from previous works (e.g., Gong et al. 2019). The mock map of the CSST spectroscopic galaxy survey at z = 0.5 is depicted in Figure 4 [Figure 4: see original paper].

3. Foreground Removal

Signal extraction from H I intensity mapping highly relies on foreground removal efficiency. Theoretically, cross-correlation with other tracers (e.g., galaxies and other emission lines) could be a good way to extract the H I signal and reduce the effects of foregrounds and system noise, since the foregrounds and instrumental noise of different wave bands in different surveys should be uncorrelated. However, it is problematic to directly cross-correlate the raw intensity map with other surveys. In Figure 5 [Figure 5: see original paper], we show the results of our mock MeerKAT H I raw (blue data points) and signal (red data points) maps cross-correlated with the mock CSST spectroscopic galaxy map. We see that foreground contamination is still too severe to extract correct cosmological information, as there is large deviation at all scales between the two curves. This indicates that extra foreground removal methods should be performed before cross-correlation.

Many foreground removal methods have been discussed in previous works. These include blind foreground subtraction algorithms such as PCA and Singular Value Decomposition (SVD) (Davis et al. 1985; Paciga et al. 2011; Villaescusa-Navarro et al. 2017; Yohana et al. 2021), ICA (Wolz et al. 2014), correlated component analysis (CCA) (Bonaldi et al. 2006), extended ICA (Zhang et al. 2016) and FASTICA (Chapman et al. 2012), non-parametric Bayesian methods like Gaussian Process Regression (GPR) (Mertens et al. 2018; Ghosh et al. 2020), and methods assuming some physical properties of the foregrounds such as polynomial/parametric-fitting (Bigot-Sazy et al. 2015; Alonso et al. 2015). Here we use the PCA/SVD algorithm to perform foreground removal. Since the PCA method is based on identifying different correlations of corresponding components in the frequency domain, it can distinguish foregrounds from the H I signal by their different frequency smoothness. Additionally, PCA does not require much knowledge about the models of data components, which is suitable for our case. SVD is a similar method that can be applied to a data matrix, obtaining similar results as PCA but with fewer calculation steps.

To apply our foreground removal procedure, we first transform the simulation result into data matrix X with dimensions N_ × N_p. Here N_ is the number of frequency channels, and N_p is the number of pixels in a frequency channel of the intensity map. Then SVD can decompose the data matrix X in the form:

$$X = W^T \Sigma R$$

where W^T and R are called left and right singular vectors, respectively, and Σ is a rectangular diagonal matrix of singular values. W^T and R are unitary matrices, defined as WW* = 1 and RR* = 1, where asterisk denotes conjugate transpose. Generally, when dealing with a complex valued matrix X, Equation (17) takes the form X = WΣR. *But since our data matrix X is real, we use transposed matrix W^T to substitute W.*

Singular vectors W^T and singular values Σ are equivalent to the eigenvectors and eigenvalues in PCA, respectively. So we rank the singular vectors corresponding to singular values in decreasing order to identify the principal components of the data matrix—i.e., the foregrounds. We then compose an N_ × m projection matrix W' with the first m columns of W, where m is the number of components thought to be foregrounds. In Figure 6 [Figure 6: see original paper], we show the first five principal components decomposed from the data matrix X. We notice that the first two components are relatively smooth in frequency, so they are identified as foreground (m = 2). The dominant principal components are obtained when the data matrix X is projected onto the projection matrix W' by:

$$U = W' \cdot W'^T X$$

Here U is the foreground information constructed from the data matrix. Then the H I signal can be recovered as:

$$S = X - U$$

Finally, the recovered signal is projected back to the original map position to obtain the foreground-removed map. In principle, the foreground-removed map is composed of H I signal and system noise. The effect of the PCA procedure can be indicated more clearly in a line intensity power spectrum as discussed in the next section.

4. Power Spectrum

We introduce the process of line intensity power spectrum estimation. We consider cross-correlation using the method based on Wolz et al. (2017). The galaxy survey and intensity mapping data are converted into galaxy over-density and brightness over-temperature contrasts respectively by:

$$\delta_g(\mathbf{x}_i) = \frac{N(\mathbf{x}_i)}{\langle N \rangle} - 1$$

$$\delta_T(\mathbf{x}_i) = \frac{T_{\rm HI}(\mathbf{x}_i)}{\langle T_{\rm HI} \rangle} - 1$$

where angled brackets denote mean values. The Fast Fourier Transforms of N(x_i) and T_H I(x_i) are given by:

$$\tilde{\delta}_g(\mathbf{k}) = \int d^{3x}\, \delta_g(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}$$

$$\tilde{\delta}_T(\mathbf{k}) = \int d^{3x}\, \delta_T(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}$$

Then our estimators for the galaxy auto-correlation power spectrum P_g and the cross power spectrum between the gridded galaxy distribution and intensity map P_× at wavevector k are:

$$\hat{P}_g(k) = \frac{1}{V} \left\langle |\tilde{\delta}_g(\mathbf{k})|^2 \right\rangle - P_{\rm SN}$$

$$\hat{P}_\times(k) = \frac{1}{V} \left\langle \tilde{\delta}_g(\mathbf{k}) \tilde{\delta}_T^*(\mathbf{k}) \right\rangle$$

Here V is the survey volume and P_{SN} is the shot noise term for the galaxy survey, which can be estimated as P_{SN} = 1/ N . The error of the cross power spectrum is given by Feldman et al. (1994); Wolz et al. (2017):

$$\sigma^2_{P_\times}(k) = \frac{1}{N_k} \left[P^2_\times(k) + P_g(k) P_T(k) \right]$$

where N_k is the number of modes in the k-bin, Δk is the k-bin width, P_T is the H I brightness temperature power spectrum, and P_N(k) is the power of system noise.

After performing PCA/SVD foreground subtraction, we estimate and show the cross power spectra of CSST galaxy with foreground-free (blue data points) and foreground-subtracted (red data points) maps from the MeerKAT H I intensity mapping survey in the left panel of Figure 7 [Figure 7: see original paper]. We find that signal loss can be caused by the PCA procedure, becoming severe especially at large scales of interest. Therefore, the over-eliminated signal must be compensated.

We compensate the cross power spectrum based on the method given in Cunnington et al. (2022). The procedure is as follows:

1. First, we generate mock data of halos. The mock halo catalogs include information on mass and position, following the same matter power spectrum and halo mass function as the SMDPL simulation.

2. After that, we calculate the H I brightness temperature of mock data using the H I model and MeerKAT observational effects. The mock H I intensity map data are obtained and further transformed into mock data matrix Y with the same dimensions as data matrix X.

3. Then the mock data matrix Y is injected into the data matrix X. We apply PCA cleaning to this data combination with the same projection matrix W' used in previous PCA. The foreground-removed mock data can be written as:

$$Y_c = X + Y - W'W'^T(X + Y)$$

So we can determine the signal loss of the cross power spectrum between Y and Y_c.

4. Finally, the transfer function is constructed as:

$$\mathcal{T}(k) = \frac{\Pi_{Y_c,Y_g}(k)}{\Pi_{Y,Y_g}(k)}$$

where Π denotes the cross power spectrum and Y_g is the corresponding mock galaxy data. To compensate the signal loss, we construct the transfer function T(k) by generating 100 H I intensity mapping mock data sets and corresponding galaxy survey data. The result of the transfer function is shown in Figure 8 [Figure 8: see original paper].

The cross-correlation compensated by the transfer function is displayed in the right panel of Figure 7. We find that the signal compensation method we use is efficient, and the compensated power spectrum is very consistent with the foreground-free power spectrum at 1σ. Although over-compensation may happen due to large variance in the low k range, the transfer function is reliable enough that the effect of signal loss can be effectively reduced.

Theoretically, the power spectra of different tracers have a similar relation to the matter power spectrum. The galaxy auto power spectrum P_g(k) is related to the matter power spectrum P_m(k) as:

$$P_g(k) = b_g^2 P_m(k)$$

where b_g is the galaxy bias. On the other hand, the relation of the H I intensity auto power spectrum P_T(k) and the matter power spectrum can be written as:

$$P_T(k) = \bar{T}_b^2 b_{\rm HI}^2 P_m(k)$$

Here b_H I is the H I bias, and $\bar{T}b$ *is the mean H I brightness temperature in the Universe. Then the cross power spectrum* $P\times$(k) is given by:

$$P_\times(k) = \bar{T}_b b_{\rm HI} b_g r_{\rm HI,g} P_m(k)$$

where r_H I,g is the H I-galaxy correlation coefficient indicating the correlation strength.

5. Cosmological Constraint

After obtaining the cross power spectrum of MeerKAT H I intensity mapping and CSST spectroscopic galaxy surveys, we can explore the constraint power on cosmological parameters. Note that the parameter T_0 can be absorbed into the H I bias b_H I (Wolz et al. 2017). Hence, in real observations, we can actually constrain the parameter product Ω_H I b_H I b_g r_H I,g using the cross power spectrum P_×(k), if the matter power spectrum P_m(k) is known. Furthermore, the constraint on Ω_H I b_H I r_H I,g can be achieved if b_g can be properly estimated. In the ideal case, if the auto-correlation of H I intensity mapping can be simultaneously detected, the correlation coefficient r_H I,g can be constrained by P_×/(P_g P_T). Unfortunately, due to inevitable foreground residuals, experiments aiming at the auto-correlation of H I intensity mapping have not achieved any convincing results so far. Here we assume the H I auto-correlation is unreachable, so the cosmological parameters we can constrain by cross-correlation power spectrum are basically limited to Ω_H I b_H I b_g r_H I,g and Ω_H I b_H I r_H I,g if b_g can be derived from a galaxy survey.

Since the matter power spectrum can be accurately calculated by a cosmological model (e.g., ΛCDM), and assuming its uncertainty is small enough to be neglected compared to the H I parameters, the theoretical matter power spectrum is a proper choice for P_m(k). We use CAMB (Lewis et al. 2000) to generate a non-linear matter power spectrum P_m(k) at z = 0.5. The cosmological parameters in CAMB are set to be the same as those in the MultiDark simulation (Klypin et al. 2016), though we can adopt values obtained from real cosmological observations in actual H I intensity mapping surveys.

Additionally, b_g can be estimated from the galaxy auto power spectrum as indicated in Equation (31). The galaxy auto power spectrum derived from the mock data of the CSST galaxy survey is displayed in Figure 9 [Figure 9: see original paper]. With the help of the theoretical matter power spectrum, the corresponding b_g can be estimated as a function of wavenumber k, as shown in Figure 10 [Figure 10: see original paper]. As can be seen, the value of b_g presents an increasing trend as the scale gets smaller. This is reasonable since baryonic matter has more complicated physical mechanisms at smaller scales,

like outflow feedback from galaxies and star formation processes, which can significantly affect density fluctuations. On the other hand, at linear scales, b_g is shown to be constant. We set the scale range to be k $<$ 0.3 h Mpc^{-1} (Cunnington et al. 2023; Deng et al. 2022) and calculate the average value of b_g from the power spectrum. The average value and error of b_g are shown in Table 3 .

In Figure 11 [Figure 11: see original paper], we show the constraint results of Ω_H I b_H I b_g r_H I,g and Ω_H I b_H I r_H I,g as functions of wavenumber k. Similar to b_g, we derive the average values of these two parameter products in the linear scales with k $<$ 0.3 h^{-1} Mpc, and the results are listed in Table 3 . We find that the errors are 1% of the average values, meaning the precision of our future MeerKAT-CSST survey can be one order of magnitude more accurate than present experimental results (Chang et al. 2010; Masui et al. 2013; Wolz et al. 2022). Furthermore, if we assume r_H I,g is close to 1 at linear scales (supported by simulation results; e.g., see Deng et al. 2022, and our simulation gives r_H I,g 0.94 at k $<$ 0.3 h Mpc^{-1}), the constraint on Ω_H I b_H I can be accurately obtained. These results indicate that the cross-correlation method is powerful for LIM surveys in studies of cosmic neutral hydrogen, galaxy evolution, and cosmology.

6. Summary and Discussion

H I intensity mapping is a promising technique for cosmological detection. Although strong foreground contamination means H I intensity mapping autocorrelation detection still faces great challenges at present, H I-galaxy cross-correlation could be easier to achieve by extracting H I and cosmological information with the help of galaxy surveys. In this work, we have investigated the cross-correlation of MeerKAT H I intensity mapping and CSST spectroscopic galaxy surveys at z 0.5.

We first generate mock H I intensity maps for MeerKAT and a galaxy catalog for the CSST spectroscopic galaxy survey using SMDPL N-body simulations at z = 0.5. System noise and foreground emission are also taken into account when making the sky map. The voxels of the simulation box are divided according to the frequency resolution of the L-band and beam size of MeerKAT single-dish mode. We construct the H I model to transform the dark matter distribution in the simulation snapshot into H I distribution. Then we calculate the H I brightness temperature of each voxel to obtain H I intensity maps. The galaxy survey catalog is generated based on the NFW profile and HOD model, and galaxy luminosity is derived from the galaxy luminosity-halo mass relation. We then filter the galaxies with the magnitude limit of the CSST spectroscopic survey to produce the mock galaxy data.

We apply the PCA/SVD algorithm to remove foregrounds in MeerKAT H I intensity mapping and cross-correlate the residual intensity map with the corresponding CSST galaxy map. Signal compensation is employed to solve the signal loss caused by the foreground removal process, and we construct a transfer func-

tion to compensate the cross power spectrum. After compensation, we derive the H I-galaxy cross power spectrum for constraining cosmological parameters.

We constrain the parameter products Ω_H I b_H I b_g r_H I,g and Ω_H I b_H I r_H I,g using the cross power spectrum and find that the constraint accuracy can achieve 1%, which is one order of magnitude higher than current results. Note that our simulation covers a 300 deg^2 survey area, and in the future, MeerKAT-CSST detection of H I-galaxy correlation can be performed on a much larger survey area of 5000 deg^2. Then the constraint accuracy can be reduced to <0.3% level. This indicates that the cross-correlation of MeerKAT H I intensity mapping and CSST galaxy survey is powerful for exploring the properties of cosmic neutral hydrogen and the evolution of galaxies and our Universe.

We also note that some assumptions in the current work may be too simple and could affect the prediction of results. For example, the frequency dependency of the beam size may contaminate the data, which will be hard to remove by the PCA/SVD foreground removal process. Moreover, the non-flat sky effect should also be seriously included, especially for the 5000 deg^2 MeerKAT-CSST joint analysis in the future. Other issues, such as the simple HOD model and systematics used in this work, probably need to be considered more carefully with more powerful simulations and precise H I models in future work.

Acknowledgments

We acknowledge support from the National Key R&D Program of China No. 2020SKA0110402, MOST-2018YFE0120800, No. 2022YFF0503404, and the National Natural Science Foundation of China (NSFC, Grant Nos. 11822305, 11773031, and 11633004). X.L.C. acknowledges support from the National Natural Science Foundation of China (NSFC, Grant Nos. 11473044 and 11973047), and Chinese Academy of Sciences grants QYZDJ-SSW-SLH017, XDB23040100, XDA15020200. Y.Z.M. is supported by the National Research Foundation of South Africa under Grant Nos. 150580, 120385, and 120378, and the NITheCS program "New Insights into Astrophysics and Cosmology with Theoretical Models confronting Observational Data". This work is also supported by science research grants from the China Manned Space Project with NO. CMS-CSST-2021-B01 and CMS-CSST-2021-A01.

References

Alonso, D., Bull, P., Ferreira, P. G., & Santos, M. G. 2015, MNRAS, 447, 400
Anderson, C. J., Luciw, N. J., Li, Y. C., et al. 2018, MNRAS, 476, 3382
Bacon, D. J., Battye, R. A., Bull, P., et al. 2020, PASP, 37, e007
Bandura, K., Addison, G. E., Amiri, M., et al. 2014, Proc. SPIE, Vol. 9145, 914522
Battye, R. A., Browne, I. W. A., Dickinson, C., et al. 2013, MNRAS, 434, 1239
Bigot-Sazy, M. A., Dickinson, C., Battye, R. A., et al. 2015, MNRAS, 454, 3240
Bonaldi, A., Bedini, L., Salerno, E., Baccigalupi, C., & De Zotti, G. 2006, MNRAS, 373, 271

Bull, P., Ferreira, P. G., Patel, P., & Santos, M. G. 2015, ApJ, 803, 21
Cao, Y., Gong, Y., Meng, X.-M., et al. 2018, MNRAS, 480, 2178
Carilli, C. L. 2011, ApJL, 730, L30
Chang, T.-C., Pen, U.-L., Bandura, K., & Peterson, J. B. 2010, Natur, 466, 463
Chapman, E., Abdalla, F. B., Harker, G., et al. 2012, MNRAS, 423, 2518
Chen, X. 2011, SSPMA, 41, 1358
Chen, X. 2012, International Journal of Modern Physics Conference Series, 12, 256
Cunnington, S., Li, Y., Santos, M. G., et al. 2023, MNRAS, 518, 6262
Cunnington, S., Li, Y., Santos, M. G., et al. 2022, arXiv:2206.01579
Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371
Deng, F., Gong, Y., Wang, Y., et al. 2022, MNRAS, 515, 5894
Dickinson, C. 2014, arXiv:1405.7936
Drinkwater, M. J., Byrne, Z. J., Blake, C., et al. 2018, MNRAS, 474, 4151
Drinkwater, M. J., Jurek, R. J., Blake, C., et al. 2010, MNRAS, 401, 1429
Dutton, A. A., & Macciò, A. V. 2014, MNRAS, 441, 3359
Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23
Fonseca, J., Silva, M. B., Santos, M. G., & Cooray, A. 2017, MNRAS, 464, 1948
Ghosh, A., Mertens, F., Bernardi, G., et al. 2020, MNRAS, 495, 2813
Gong, Y., Chen, X., & Cooray, A. 2020, ApJ, 894, 152
Gong, Y., Cooray, A., & Santos, M. G. 2013, ApJ, 768, 130
Gong, Y., Cooray, A., Silva, M., et al. 2012, ApJ, 745, 49
Gong, Y., Cooray, A., Silva, M. B., et al. 2017, ApJ, 835, 273
Gong, Y., Cooray, A., Silva, M. B., Santos, M. G., & Lubin, P. 2011, ApJL, 504, L57
Gong, Y., Liu, X., Cao, Y., et al. 2019, ApJ, 883, 203
Gong, Y., Silva, M., Cooray, A., & Santos, M. G. 2014, ApJ, 785, 72
Klypin, A., Yepes, G., Gottlöber, S., Prada, F., & Heß, S. 2016, MNRAS, 457, 4340
Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473
Li, L.-C., & Wang, Y.-G. 2022, RAA, 22, 115005
Lidz, A., Furlanetto, S. R., Oh, S. P., et al. 2011, ApJ, 741, 70
Masui, K. W., Switzer, E. R., Banavar, N., et al. 2013, ApJL, 763, L20
Mertens, F. G., Ghosh, A., & Koopmans, L. V. E. 2018, MNRAS, 478, 3640
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563
Newburgh, L. B., Addison, G. E., Amiri, M., et al. 2014, Proc. SPIE, 9145, 91454V
Paciga, G., Chang, T.-C., Gupta, Y., et al. 2011, MNRAS, 413, 1174
Perdereau, O., Ansari, R., Stebbins, A., et al. 2022, MNRAS, 517, 4637
Pullen, A. R., Doré, O., & Bock, J. 2014, ApJ, 786, 111
Santos, M., Bull, P., Alonso, D., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 19
Santos, M. G., Cluver, M., Hilton, M., et al. 2017, MeerKLASS: MeerKAT Large Area Synoptic Survey
Sheth, R. K., & Tormen, G. 1999, MNRAS, 308, 119
Silva, M., Santos, M. G., Cooray, A., & Gong, Y. 2015, ApJ, 806, 209

Silva, M. B., Santos, M. G., Gong, Y., Cooray, A., & Bock, J. 2013, ApJ, 763, 132
Smoot, G. F., & Debono, I. 2017, A&A, 597, A136
Spinelli, M., Carucci, I. P., Cunnington, S., et al. 2022, MNRAS, 509, 2048
Uzgil, B. D., Aguirre, J. E., Bradford, C. M., & Lidz, A. 2014, ApJ, 793, 116
Vale, A., & Ostriker, J. P. 2008, MNRAS, 383, 355
Villaescusa-Navarro, F., Alonso, D., & Viel, M. 2017, MNRAS, 466, 2736
Villaescusa-Navarro, F., Genel, S., Castorina, E., et al. 2018, ApJ, 866, 135
Visbal, E., & Loeb, A. 2010, JCAP, 2010, 016
Wang, J., Santos, M. G., Bull, P., et al. 2021, MNRAS, 505, 3698
Wolz, L., Abdalla, F. B., Blake, C., et al. 2014, MNRAS, 441, 3271
Wolz, L., Blake, C., & Wyithe, J. S. B. 2017, MNRAS, 470, 3220
Wolz, L., Pourtsidou, A., Masui, K. W., et al. 2022, MNRAS, 510, 3495
Yohana, E., Ma, Y.-Z., Li, D., Chen, X., & Dai, W.-M. 2021, MNRAS, 728, L46
Zhan, H. 2011, SSPMA, 41, 1441
Zhan, H. 2021, ChSBu, 66, 1290
Zhang, J., Motta, P., Novaes, C. P., et al. 2022, A&A, 664, A19
Zhang, L., Bunn, E. F., Karakci, A., et al. 2016, ApJS, 222, 3
Zheng, H., Tegmark, M., Dillon, J. S., et al. 2017, MNRAS, 464, 3486
Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760

Note: Figure translations are in progress. See original paper for figures.