A Semiblind PCA-based Foreground Subtraction Method for 21 cm Intensity Mapping

Shifan Zuo1,2, Xuelei Chen2,3,4,5, and Yi Mao1

1 Department of Astronomy, Tsinghua University, Beijing 100084, People’s Republic of China; sfzuo@bao.ac.cn, xuielei@cosmology.bao.ac.cn, ymao@tsinghua.edu.cn
2 Key Laboratory of Computational Astrophysics, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, People’s Republic of China
3 School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049, People’s Republic of China
4 Department of Physics, College of Sciences, Northeastern University, Shenyang 110819, People’s Republic of China
5 Center of High Energy Physics, Peking University, Beijing 100871, People’s Republic of China

Received 2022 August 31; revised 2023 January 18; accepted 2023 January 31; published 2023 March 6

Abstract

The principal component analysis (PCA) method and the singular value decomposition (SVD) method are widely used for foreground subtraction in 21 cm intensity mapping experiments. We show their equivalence, and point out that the condition for completely clean separation of foregrounds and cosmic 21 cm signal using the PCA/SVD is unrealistic. We propose a PCA-based foreground subtraction method, dubbed the “singular vector projection (SVP)” method, which exploits a priori information of the left and/or right singular vectors of the foregrounds. We demonstrate with simulation tests that this new, semiblind method can reduce the error of the recovered 21 cm signal by orders of magnitude, even if only the left and/or right singular vectors in the largest few modes are exploited. The SVP estimators provide a new, effective approach for 21 cm observations to remove foregrounds and uncover the physics in the cosmic 21 cm signal.

Unified Astronomy Thesaurus concepts: Cosmology (343); Astronomical methods (1043); H I line emission (690); Principal component analysis (1944)

1. Introduction

The 21 cm radiation due to the hyperfine transition of atomic hydrogen (Furlanetto et al. 2006; Morales & Wyithe 2010; Pritchard & Loeb 2012) has emerged as a promising cosmological probe that can be used to reconstruct the history from cosmic dawn to the epoch of reionization, as well as the large-scale structure of the universe after reionization. A number of completed and ongoing radio interferometric array experiments are targeting the 21 cm intensity mapping and/or its power spectrum measurement, including the Precision Array for Probing the Epoch of Reionization (Parsons et al. 2010), the Giant Meterwave Radio Telescope (Paciga et al. 2013; Interna et al. 2017), the Murchison Widefield Array (Bowman et al. 2013), the LOw Frequency Array (LOFAR; Patil et al. 2013; Intema et al. 2015; Bigot-Sazy et al. 2015), the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017), the Green Bank Telescope (GBT; Chang et al. 2010; Masui et al. 2013; Switzer et al. 2013), the Parkes Radio Telescope (Anderson et al. 2018), the Owens Valley Long Wavelength Array (Eastwood et al. 2018, 2019), the MeerKAT telescope (Wang et al. 2021), the Canadian Hydrogen Intensity Mapping Experiment (CHIME; Newburgh et al. 2014; CHIME Collaboration et al. 2022), the Hydrogen Intensity and Real-time Analysis eXperiment (Chrichton et al. 2022), the Tianlai experiment (CHEN 2012; Xu et al. 2015; Li et al. 2020; Wu et al. 2021), and the BAO from Integrated Neutral Gas Observations (Battye et al. 2012). Next-generation experiments with higher sensitivities are also under construction, including the Canadian Hydrogen Observatory and Radio-transient Detector (Vanderlinde et al. 2019), and the Square Kilometre Array (SKA; Koopmans et al. 2015; Maartens et al. 2015).

Detection of the cosmic 21 cm signal is technically very challenging, even for next-generation interferometric arrays with large receiving area, because the astrophysical foregrounds from the galactic and extragalactic sources are 4–5 orders of magnitude larger than the cosmic 21 cm signal. Advanced techniques for foreground subtraction or avoidance have been developed (see, e.g., the review of Liu & Shaw 2020). However, residual foreground still remains a main source of systematics for 21 cm intensity mapping. These methods include the low-order polynomial fitting (Wang et al. 2006; Jelić et al. 2008; Liu et al. 2009a, 2009b), the principal component analysis (PCA) method (Masui et al. 2013; Alonso et al. 2015; Bigot-Sazy et al. 2015), the singular value decomposition (SVD) method (Switzer et al. 2013, 2015), the robust PCA method (Zuo et al. 2019), the independent component analysis method (Chapman et al. 2012; Wolz et al. 2014; Alonso et al. 2015), the generalized morphological component analysis method (Chapman et al. 2013), and the Gaussian process regression (GPR) method (Mertens et al. 2018).

In particular, the PCA method has been shown to work well in simulation tests (Alonso et al. 2015; Bigot-Sazy et al. 2015) and is applied extensively to the analysis of observational data (Masui et al. 2013; Cunnington et al. 2023). It has been demonstrated using the mock data from simulations that the residual foreground can be effectively suppressed to the level comparable to, or even below, that of the cosmic 21 cm signal, after removing four to five PCA modes (see, e.g., Alonso et al. 2015; Zuo et al. 2019). Those demonstrations often only assume that the spectrum of foreground is smooth in the frequency direction. Furthermore, it is shown that satisfactory performance can be obtained by removing six to seven PCA modes, if we take into account complex instrumental effects such as frequency-dependent beam, $1/f$ noise and other...
imperfections (Alonso et al. 2015; Bigot-Sazy et al. 2015). Masui et al. (2013) made a measurement of the 21 cm brightness fluctuation power spectrum at \( z \sim 0.8 \) by the cross-correlation of 21 cm signal using MeerKAT with the large-scale structure traced by galaxies that are optically detected in the WiggleZ Dark Energy Survey. Nevertheless, it is necessary to subtract out 20 PCA modes in the MeerKAT observations in order to suppress the residual foreground to an acceptably low level (Masui et al. 2013). Moreover, in a recent, similar detection of the cross-correlation of 21 cm signal using MeerKAT with galaxies from the WiggleZ data, 30 PCA modes are removed (Cunnington et al. 2023), due to the modulation of instrumental responses such as calibration imperfections, residual radio frequency interference (RFI), and other systematics.

Removing such a large number of PCA modes results in severe signal loss and foreground mixing. In fact, even after removal of 30 PCA modes, the resulting 21 cm signal in MeerKAT has not reached the estimated auto-power spectrum level (Cunnington et al. 2023), which implies that there are still residual RFI, foregrounds, and other systematics in the 21 cm data. To avoid these additive biases, Masui et al. (2013) and Cunnington et al. (2023) employed the cross-correlation technique of the 21 cm signal with galaxy surveys for the detection of the 21 cm brightness fluctuations. This difficulty suggests that the PCA method needs to be further improved in order to remove the foregrounds to such a cleaned level that we can realize the detection of the cosmic 21 cm signal through the auto-power spectrum.

The standard PCA method is a blind method, in the sense that no prior information about the signal or foreground is assumed other than the assumption of a smooth spectrum of the foregrounds in frequency, which is too generic. Some semiblind methods have been proposed to make assumptions about the foregrounds and/or signal that are not so strong as assuming an ansatz of foregrounds in the polynomial fitting method, and demonstrated to alleviate the problem of signal loss in foreground subtraction. These methods include the Karhunen–Loève (KL) transform method (Shaw et al. 2014, 2015) that employs the modeled foregrounds and signal covariance matrices to form a set of modes that are ordered by their signal-to-contaminant ratios, and the generalized needlet internal linear combination method (Olivari et al. 2016) that uses the H I covariance matrix to separate the foregrounds from the signal in a wavelet (or needlet) space. While it is indeed easier to model the covariance matrix of the foregrounds and signal than the foregrounds or the signal per se, the overall magnitude of the covariance matrix may be highly biased, because the foregrounds are 4–5 orders of magnitude stronger than cosmic 21 cm signal. If we take into account the thermal noise, modeling the covariance matrix can be even more challenging due to the correlated noise between different frequencies and/or between different pixels induced by instrumental effects. These issues impose practical difficulty in applying these methods to the observational data.

In this paper, instead, we will propose a new semiblind PCA-based method for foreground subtraction, dubbed singular vector projection (SVP). The SVP method assumes that the covariance matrix of foregrounds can be modeled a priori by theoretical modeling or from observational data, and then the left and/or right singular vectors of the foregrounds can be obtained by eigen-decomposition of the covariance matrix of the foregrounds in frequency and/or pixel space. Note that the overall magnitude of the covariance matrix affects its eigenvalues, but not the left and right singular vectors. If we can design the SVP estimators in such a manner that they contain the singular vectors, but not the eigenvalues, as will be shown in the paper below, then the estimators are independent of the overall magnitude of the covariance matrix. In other words, the estimators would not be affected if the overall magnitude of the covariance matrix was biased. We will also demonstrate the effectiveness of SVP in reducing the signal loss and foreground mixing in foreground subtraction.

The rest of this paper is organized as follows. In Section 2, we review the PCA and SVD methods for foreground subtraction briefly, show their equivalence, and derive the ideal condition for completely clean separation of foregrounds and the cosmic 21 cm signal (i.e., without any signal loss or foreground mixing). We introduce our SVP method for foreground subtraction in Section 3, and demonstrate its effectiveness with simulation tests in Section 4. We make concluding remarks in Section 5. We leave some technical details (the mathematical proof of the inequalities for signal losses) to the Appendix.

2. The PCA and SVD Methods

In this section, we briefly review the standard PCA and SVD methods for foreground subtraction. We will show their equivalence, and prove the ideal, yet unrealistic, conditions for completely clean separation of foregrounds and cosmic 21 cm signal.

2.1. Problem Setup

The 21 cm observational data is presented as a 3D image cube with two angular directions and one frequency direction. The two angular directions can be combined into a single dimension of size \( p \), where \( p \) is the number of image pixels, and therefore the data set is represented in the form of a matrix \( \mathbf{D} \in \mathbb{R}^{n \times p} \), where \( n \) is the number of frequency bins. This general representation of the data set is valid for the observation of a patch of sky as well as the full sky. Without losing generality, we assume \( n \leq p \) in the analysis below, i.e., the number of frequency bins is equal to or less than the number of pixels in the map. This is often the case, but the conclusions herein are also valid for the case of \( n > p \).

The observational data is a linear combination of the foregrounds, cosmic 21 cm signal, and systematic noise, by writing it as

\[
\mathbf{D} = \mathbf{F} + \mathbf{N},
\]

where the total foregrounds are denoted as \( \mathbf{F} \), and \( \mathbf{N} \) includes both the cosmic 21 cm signal and noise.

2.2. The PCA Method

We follow Masui et al. (2013), Alonso et al. (2015), and Bigot-Sazy et al. (2015) for the review of the PCA method for
foreground subtraction. Consider the covariance matrix $R \in \mathbb{R}^{n \times n}$ of the data set in frequency space, $R = DD^T$, and perform the eigen-decomposition

$$R = U \Lambda U^T.$$  \hfill (2)

Here, $O^T$ denotes the transpose of a matrix $O$. $\Lambda$ is an $n \times n$ diagonal matrix in which the diagonal elements are the eigenvalues $\{\lambda_i\}$ of the matrix $R$. The matrix $U$ is an $n \times n$ real orthogonal matrix in which its $i^{th}$ column is the eigenvector of $R$ corresponding to the $i^{th}$ eigenvalue $\lambda_i$. The magnitude of $\lambda_i$ gives the variance of the corresponding eigenmode, and each eigenvalue measures the contribution of its corresponding eigenvector to the total sky variance. Since the foregrounds dominate the full data overwhelmingly, we can project out the dominant components by picking up the $m$ largest eigenvalues and their corresponding eigenvectors, so the foregrounds and the 21 cm signal can be estimated, respectively, by

$$F_{\text{PCA}} = U \Pi_{n,m} U^T D,$$  \hfill (3)

$$N_{\text{PCA}} = D - F_{\text{PCA}} = U(I - \Pi_{n,m}) U^T D.$$  \hfill (4)

Here, $I_n$ is an $n \times n$ identity matrix, and $\Pi_{n,m}$ is a projection matrix from dimension $n$ to $m$, i.e., an $n \times n$ diagonal matrix in which $m$ diagonal elements are unity if they correspond to the picked eigenvalues, and all other diagonal elements are zero.

### 2.3. The SVD Method

The data set $D$ can be decomposed with SVD as

$$D = U S V^T.$$  \hfill (5)

Here, $S \in \mathbb{R}^{k \times k}$ is a diagonal matrix in which its positive diagonal elements $\{s_i\}$ are the singular values, where the integer $k \leq \min(n, p)$ is the number of singular values of the data set. The matrices $U' \in \mathbb{R}^{n \times k}$ and $V \in \mathbb{R}^{p \times k}$ are the corresponding left and right singular vectors, respectively. These two singular-vector matrices satisfy the conditions, $U'^T U' = I_k$ and $V^T V = I_k$, where $I_k$ is a $k \times k$ identity matrix. However, in general, they do not necessarily meet the following conditions: $U' U'^T = I_k$ (unless $k = n$), and $V^T V = I_p$ (unless $k = p$). For this reason, a matrix like $U'$ and $V$ is called a partial orthogonal matrix, because it contains some columns of an orthogonal matrix.

Using SVD, the foregrounds can be projected out by picking up the largest $m$ singular value modes (Switzer et al. 2013, 2015), similar to the PCA. The foregrounds and the 21 cm signal can be estimated, respectively, by

$$F_{\text{SVD}} = U \Pi_{k,m} V S^T,$$  \hfill (6)

$$N_{\text{SVD}} = D - F_{\text{SVD}} = U'(I_k - \Pi_{k,m}) V S^T.$$  \hfill (7)

### 2.4. The Equivalence of PCA and SVD

Note that the dimension $k$ of the matrix $S$ is also the number of positive eigenvalues for the covariance matrix $R$. The eigenvalues of $R$ are nonnegative, which means that there are $n - k$ zero eigenvalues of $R$. Below, we assume that the $k$ positive eigenvalues of $R$ (or equivalently $k$ singular values of $S$) are in descending order.

Substituting Equation (5) into Equation (2), we get

$$U A U^T = U S^2 U^T.$$  \hfill (8)

It is straightforward to prove that

$$U A U^T = U_k \Lambda_k U_k^T,$$  \hfill (9)

where $\Lambda_k$ is the $k \times k$ subset of $\Lambda$, i.e., the diagonal matrix in which the diagonal elements are the $k$ positive eigenvalues of $R$, and $U_k$ is the $n \times k$ subset of $U$, i.e., the matrix in which its columns are the first $k$ columns of $U$ corresponding to the $k$ positive eigenvalues. So we find $\Lambda_k = S^2$ and $U_k = \pm U$. The undetermined sign is due to the fact that $U S V^T = (-U) S (-V^T)$. Ignoring this sign degeneracy, we have

$$U' = U_k.$$  \hfill (10)

It is straightforward to prove that

$$U \Pi_{n,m} U^T = U_k \Pi_{k,m} U_k^T,$$  \hfill (11)

$$D = U U^T D = U_k U_k^T D.$$  \hfill (12)

The proof uses the identity $U_k^T U_k = I_k$. Substituting Equation (5) into Equations (3) and (4), therefore, we find $F_{\text{PCA}} = F_{\text{SVD}}$ and $N_{\text{PCA}} = N_{\text{SVD}}$. This shows that foreground subtraction results using PCA and SVD methods are equivalent.

As such, the estimators in the PCA/SVD are

$$F_{\text{PCA/SVD}} = U \Pi U^T D = U I S V^T,$$  \hfill (13)

$$N_{\text{PCA/SVD}} = U(I - \Pi) U^T D = U(I - \Pi) S V^T.$$  \hfill (14)

For simplicity, hereafter in this paper, we will drop the prime (') in $U'$ (because of Equation (8)), and drop the subscripts $I$, $\Pi$, and $U$, because the dimensionalities of these matrices can be understood both in the context of PCA (where $U$ is an $n \times n$ matrix) and in the context of SVD (where $U$ is an $n \times k$ matrix) due to Equations (9) and (10), as long as the interpretation of dimensionalities is self-consistent.

### 2.5. Conditions for Completely Cleaned Separation

In this subsection, we attempt to answer this question in the framework of PCA/SVD method: under what condition can we separate the foregrounds from the 21 cm signal in such a completely cleaned manner that there is no signal loss or foreground mixing?

Consider the SVD of the foregrounds and the signal, respectively,

$$F = U_f S_f V_f^T,$$  \hfill (15)

$$N = U_n S_n V_n^T.$$  \hfill (16)

The subscripts $f$ and $n$ denote the foreground and signal (with noise), respectively.

If the foregrounds and the signal are fully separated, then

$$U_f S_f V_f^T = U I S V^T,$$  \hfill (17)

$$U_n S_n V_n^T = U(I - \Pi) S V^T.$$  \hfill (18)

Since the SVD is unique, the completely cleaned separation can be realized if and only if the singular values of $F$ are the largest $m$ positive singular values of $D$, and the columns of $U_f$ ($V_f$) are the corresponding singular vectors in $U$ ($V$), while the singular values of $N$ are the remaining positive singular values of $D$, and the columns of $U_n$ ($V_n$) are the corresponding singular vectors in $U$ ($V$).

Thus, we find the conditions: (1) $U_f^T U_f = 0$; (2) $V_f^T V_f = 0$; (3) $\min(S_f) > \max(S_n)$. The first two conditions are equivalent to
the orthogonality conditions
\[ F^T N = N^T F = 0, \]
\[ F N^T = N F^T = 0. \] (15)

The first condition means that there is no pixel-wise cross-correlation between the foregrounds and the signal, so the pixel covariance can completely separate the contributions from the foregrounds and from the signal, i.e., \( D^T D = F^T F + N^T N \). Similarly, the second condition means that there is no frequency-wise cross-correlation between the foregrounds and the signal.

Equation (15) is the necessary condition for complete separation, but there is another implicit condition in practice—the number \( m \) of PCA/SVD modes of the foregrounds should be known a priori. If \( m \) is known, the foregrounds can be reconstructed from the left and right singular vectors corresponding to the largest \( m \) singular values of \( D \), and the signal is recovered from the remaining singular vectors.

However, we note that these conditions for complete separation of foregrounds and signal are ideal and unsatisfied in most cases. As the result, the signal loss and foreground mixing are unavoidable in practice.

### 2.6. Signal Loss, Foreground Mixing, and Recovery Error

Substituting Equation (1) into Equation (12), the estimated signal with the PCA/SVD is
\[ N_{\text{PCA/SVD}} = U (I - \Pi) U^T F + U (I - \Pi) U^T N = U (I - \Pi) U^T F + \tilde{N}, \]
\[ \text{and the signal loss is} \]
\[ N_{\text{loss}} = U (I - \Pi) U^T N. \] (18)

Since \( U^T U = I \) and \( U^T U_n = 0 \) in general, \( F_{\text{PCA/SVD}}^\text{mix} \) and \( N_{\text{PCA/SVD}}^\text{loss} \) are non-zero; i.e., in the standard PCA/SVD method, there are signal loss and foreground mixing, if the conditions in Section 2.5 are not met.

We define the recovery error as the difference between the true signal and the recovered signal, \( \Delta N = N - \tilde{N} \). For PCA/SVD, the recovery error is
\[ \Delta N_{\text{PCA/SVD}} = N_{\text{PCA/SVD}}^\text{loss} - F_{\text{PCA/SVD}}^\text{mix}. \] (19)

### 3. Singular Vector Projection

In what we call the SVP method, we propose to exploit the information of the singular vectors of the foregrounds, \( U_f \) and/or \( V_f \), if it is known a priori, to improve the accuracy of foreground subtraction. We propose the new estimators in Sections 3.1 and 3.2, provide a pedagogical example in Section 3.3, and discuss the feasibility of obtaining a priori information of \( U_f \) and/or \( V_f \) in Section 3.4.

#### 3.1. The SVP Estimators

We propose four estimators for the data \( D \), as follows:
\[ N_{L} = D - U_f U_f^T D, \] (20)
\[ N_{R} = D - D V_f V_f^T, \] (21)
\[ N_{B} = D - U_f U_f^T D V_f V_f^T, \] (22)
\[ N_{D} = D - U_f (U_f^T D V_f V_f^T)_{\text{diag}} V_f^T. \] (23)

Here, \( O_{\text{diag}} \) denotes a diagonal matrix in which its elements are the same as the diagonal elements of the matrix \( O \).

If only the left (right) singular vector of the foregrounds, \( U_f (V_f) \), is known a priori, then the estimator \( N_L (N_R) \) can be applied. If both left and right singular vectors of the foregrounds are known a priori, then the estimators \( N_L \) and \( N_R \) can be applied. The subscripts “L,” “R,” “B,” and “D” stand for “left,” “right,” “both,” and “diagonal,” respectively.

It is straightforward to prove the following results:
\[ N_{L} = N - U_f U_f^T N, \] (24)
\[ N_{R} = N - N V_f V_f^T, \] (25)
\[ N_{B} = N - U_f U_f^T N V_f V_f^T, \] (26)
\[ N_{D} = N - U_f (U_f^T N V_f V_f^T)_{\text{diag}} V_f^T. \] (27)

This shows that these estimators can project out the foregrounds \( F \) completely, i.e., no foreground mixing. In fact, these are the only four estimators that meet this requirement. This is an advantage against the blind PCA/SVD method, because given that the foreground is several orders of magnitude stronger than the signal, even a small residual foreground mixing can result in large recovery error. The proof of Equations (24) to (26) uses the identity \( U_f U_f^T F = F V_f V_f^T = F, \) which is due to \( U_f U_f^T = I \) and \( V_f V_f^T = I \). The proof of Equation (27) uses the identity \( (U_f^T F V_f)_{\text{diag}} = (S_f)_{\text{diag}} = S_f \), because \( S_f \) is diagonal.

The recovery error and signal loss for these estimators are as follows, respectively:
\[ \Delta N_{L} = N_{L}^\text{loss} = U_f U_f^T N, \]
\[ \Delta N_{R} = N_{R}^\text{loss} = N V_f V_f^T, \]
\[ \Delta N_{B} = N_{B}^\text{loss} = U_f U_f^T N V_f V_f^T, \]
\[ \Delta N_{D} = N_{D}^\text{loss} = U_f (U_f^T N V_f V_f^T)_{\text{diag}} V_f^T. \] (28)

If the condition for the left singular vectors \( U_f U_f^T U_n = 0 \) is met, then \( \Delta N_{L} = \Delta N_{B} = \Delta N_{D} = 0 \), and if the condition for the right singular vectors \( V_n V_f V_f = 0 \) is met, then \( \Delta N_{R} = \Delta N_{B} = \Delta N_{D} = 0 \). Clearly, using the estimators \( N_B \) and \( N_D \) with the information of both left and right singular vectors is more likely to get smaller recovery error than using either \( N_L \) or \( N_R \) with only the left or right singular vector. In fact, roughly speaking, the “magnitude” of recovery error for these estimators has the following relation: \( \Delta N_{D} \leq \Delta N_{B} \leq \Delta N_{L} \) and \( \Delta N_{D} \leq \Delta N_{B} \leq \Delta N_{R} \). We leave it to the Appendix to give an accurate definition of their “magnitudes” and the proof of these relations. Note that these inequality relations are valid for the whole matrix, and do not necessarily hold for each individual frequency bin.
3.2. SVP with Incomplete Singular Vectors

In Section 3.1, we implicitly assume that the left and/or right singular vectors of the foregrounds, \( U_j \in \mathbb{R}^{p \times k} \) and \( V_j \in \mathbb{R}^{p \times k} \), for all \( k \) modes can be well modeled or measured a priori. In practice, this might not be satisfied.

Consider a relaxed condition in which the left and/or right singular vectors of the foregrounds for only a small number of modes that correspond to the \( l \) largest singular values can be modeled a priori, labeled as \( U_{jl} \in \mathbb{R}^{p \times l} \) and \( V_{jl} \in \mathbb{R}^{p \times l} \), and the other \( k - l \) singular vectors of the foregrounds, labeled as \( U_{j(k-l)} \in \mathbb{R}^{p \times (k-l)} \) and \( V_{j(k-l)} \in \mathbb{R}^{p \times (k-l)} \), are not known. The total foregrounds can be written as the sum of two parts, \( F = F_1 + F_2 \), where

\[
F_1 = U_{jl} S_{jl} V_{jl}^T,
\]

\[
F_2 = U_{j(k-l)} S_{j(k-l)} V_{j(k-l)}^T.
\]

Here, \( S_{jl} \in \mathbb{R}^{l \times l} \) is the diagonal matrix of the \( l \) largest singular values, and \( S_{j(k-l)} \in \mathbb{R}^{(k-l) \times (k-l)} \) is the diagonal matrix of the rest \( k - l \) singular values. We assume that the former is significantly larger than the latter.

In this case with incomplete information of singular vectors of the foregrounds, the estimators for the data \( D \) are as follows:

\[
N_l = D - U_{jl} U_{jl}^T D,
\]

\[
N_R = D - D V_{jl} V_{jl}^T,
\]

\[
N_B = D - U_{jl} U_{jl}^T D V_{jl} V_{jl}^T,
\]

\[
N_D = D - U_{jl} (U_{jl}^T D V_{jl} V_{jl}^T)_\text{diag} V_{jl}^T.
\]

Similar to the results in Section 3.1, it is straightforward to prove the following results:

\[
N_l = N - U_{jl} U_{jl}^T N + F_2,
\]

\[
N_R = N - N V_{jl} V_{jl}^T + F_2,
\]

\[
N_B = N - U_{jl} U_{jl}^T N V_{jl} V_{jl}^T + F_2,
\]

\[
N_D = N - U_{jl} (U_{jl}^T N V_{jl} V_{jl}^T)_\text{diag} V_{jl}^T + F_2.
\]

This shows that these estimators project out the major foreground component \( F_1 \), but the unknown, minor foreground component \( F_2 \) is left as the residual foreground mixing, \( F_{\text{mix}} = F_2 \). The proof of Equations (34)–(37) uses the orthogonality relation \( U_{jl}^T U_{jl} = 0 \) and \( V_{jl}^T V_{jl} = 0 \), and the identities \( U_{jl} U_{jl}^T F_1 = F_1 \), \( V_{jl} V_{jl}^T = U_{jl} U_{jl}^T F_1 V_{jl} V_{jl}^T \), \( V_{jl} V_{jl}^T = U_{jl} U_{jl}^T V_{jl} V_{jl}^T \), \( (U_{jl} F_{jl})_\text{diag} V_{jl}^T = F_1 \).

The recovery errors for these estimators are as follows:

\[
\Delta N_l = U_{jl} U_{jl}^T N - F_2,
\]

\[
\Delta N_R = N V_{jl} V_{jl}^T - F_2,
\]

\[
\Delta N_B = U_{jl} U_{jl}^T N V_{jl} V_{jl}^T - F_2,
\]

\[
\Delta N_D = U_{jl} (U_{jl}^T N V_{jl} V_{jl}^T)_\text{diag} V_{jl}^T - F_2.
\]

Finally, we note that there is actually a freedom of normalization in the singular vectors. However, our SVP estimators contain the singular vectors that are always in pairs, such as \( U_{jl} U_{jl}^T \) and \( V_{jl} V_{jl}^T \), so the estimators are independent of this normalization in the singular vectors. Also, the estimators do not depend on the eigenvalues \( \Lambda_f \) of the covariance matrix, so the estimators are not affected by the overall magnitude of the covariance matrix. In other words, even if the covariance matrix of the foregrounds was biased in overall amplitude from observations, the SVP estimators would not be affected.

3.3. Pedagogical Example

In this subsection, we use a pedagogical example to illustrate the SVP. Consider a simple case where the image \( D \) has two frequency bins and two pixels. Suppose the foregrounds have a rank-one matrix, \( F = s_f u_1 v_1^T \), and the signal is a rank-two matrix, \( N = s_n u_{n1} v_{n1}^T + s_n u_{n2} v_{n2}^T \). Here \( u_f(v_f) \) is the (left) (right) singular vector of \( F \), \( u_{n1,n2} \) and \( v_{n1,n2} \) are the (left) (right) singular vectors of \( N \), and perpendicular to each other, i.e., \( u_{n1} \perp u_{n2} \) and \( v_{n1} \perp v_{n2} \). They are all unit vectors of the size \( 2 \times 1 \). In this example, we set \( s_f = 1.0 \), \( s_{n1} = 0.8 \), and \( s_{n2} = 0.8 \), so that the magnitude of \( F \) is about 10 times larger than \( N \). For such a simple case, vectors of the size \( 2 \times 1 \) can be visualized in a 2D plane, so we plot \( F \) by the rescaled vectors \( \sqrt{s_f} u_1 \) and \( \sqrt{s_f} v_1 \), and plot \( N \) by the rescaled vectors \( \sqrt{s_{n1}} u_{n1} \), \( \sqrt{s_{n2}} u_{n2} \), \( \sqrt{s_{n1}} v_{n1} \), and \( \sqrt{s_{n2}} v_{n2} \) in Figure 1. Similar rescaling will be applied when we refer to the “magnitude” of a vector in this subsection.

In the standard SVD analysis, the image is decomposed as \( D = s_f u_{d1} v_{d1}^T + s_f u_{d2} v_{d2}^T \). Here, \( s_{d1} > s_{d2} \), and \( u_{d1} \perp u_{d2} \), \( v_{d1} \perp v_{d2} \). In this decomposition, the first term \( s_{d1} u_{d1} v_{d1}^T \) is identified and thus removed as the foregrounds, but Figure 1 shows that the direction of the left (right) singular vector \( u_{d1} \) (\( v_{d1} \)) is different from that of the left (right) singular vectors \( u_f \) (\( v_f \)) of the true foregrounds \( F \). On the other hand, the second term \( s_{d2} u_{d2} v_{d2}^T \) of the SVD decomposition is identified as the recovered signal \( N_{\text{PCA/SVD}} \), but it is rank-one in this case, while the true signal \( N \) is rank-two. Also, Figure 1 shows that the magnitude of \( N_{\text{PCA/SVD}} \) is much smaller than \( N \), which means a significant signal loss in the SVD foreground removal.

Now, assuming that we know \( u_f \) and \( v_f \) as discussed in Section 3.1, we can project out the foregrounds in four ways. The optimal estimators are \( N_l \) and/or \( N_D \) that exploit the information of both \( u_f \) and \( v_f \). For the pedagogical example considered herein, the two estimators give the same results because \( u_f^T D v_f \) is a \( 1 \times 1 \) matrix, i.e., a number, in this example, so \( u_f^T D v_f = (u_f^T D v_f)_\text{diag} \). We can rewrite the estimated foregrounds as \( F_{B/D} = s_{df} u_f v_f^T \), where we define \( s_{df} \equiv u_f^T D v_f \) for this case. The left (right) singular vector \( \sqrt{s_{df}} u_f \) (\( \sqrt{s_{df}} v_f \)) of the estimated foregrounds \( F_{B/D} \) has the same direction as the left (right) singular vector \( \sqrt{s_f} u_f \) (\( \sqrt{s_f} v_f \)) of the true foregrounds \( F \). In principle, the magnitude of the former is slightly larger than the latter, because \( s_{df} > s_f \), but in this simple example, their magnitudes are very close, as shown in Figure 1. This implies that the subtracted foregrounds contain the full foregrounds \( F \) and some amount of signal. So the recovered signal, \( N_{B/D} = D - F_{B/D} \) is free from contamination of any residual foregrounds, but suffers from some amount of signal loss. Also, \( N_{B/D} = s_{df} u_{df1} v_{df1}^T + s_{df} u_{df2} v_{df2}^T \) is still a rank-two matrix. Figure 1 shows that the recovered signal \( N_{B/D} \) is close to the true signal \( N \). This simple example clearly shows that using the information of the left and right singular vectors of the foregrounds can help improve the performance of foreground subtraction.
3.4. Modeling the Singular Vectors

The left and right singular vectors of the foregrounds, \( U_i \) and \( V_j \), can be obtained by solving for the eigen-decomposition of the frequency and pixel covariance matrix, respectively, i.e.,

\[ F F^T = U_i \Lambda_i U_i^T, \quad \text{and} \quad F^T F = V_j \Lambda_j V_j^T, \]

where \( \Lambda_i = S_i^2 \). Our assumption is that the covariance matrix of the foregrounds can be estimated from modeling or observations a priori. In practice, we may keep only the singular vectors with positive eigenvalues, and drop all modes with nearly zero eigenvalues.

We follow de Oliveira-Costa et al. (2008), which is essentially a PCA method, to estimate the foregrounds. We summarize the approach below, and refer interested readers to de Oliveira-Costa et al. (2008) for details. Typically, we have the data set from a number of available surveys that altogether can cover the full sky, but their overlapping sky coverage is only a patch of sky with \( n_{\text{pix}} \) pixels with data at all \( n_f \) frequencies. For the data subset of the overlapping sky map \( y \in \mathbb{R}^{n_f \times n_{\text{pix}}} \), we begin by estimating its matrix of second moments of the size \( n_f \times n_f \), which is essentially the normalized covariance matrix \( C = y y^T / n_{\text{pix}} \), and then estimating the correlation matrix \( R_{ij} = C_{ij} / \sqrt{C_{ii} C_{jj}} \) that corresponds to the dimensionless correlation coefficients between all pairs of frequencies. Here \( C_{ij} \equiv y_i y_j \) is the rms fluctuations at each frequency. We then perform a standard eigenvalue decomposition to diagonalize the matrix \( R \) as \( R = P \Lambda P^T \), where \( P \) is an orthogonal matrix in which the columns are the eigenvectors (principal components), and \( \Lambda \) is a diagonal matrix containing the corresponding eigenvalues. We determine the first \( n_c \) principal components that best approximate the data using the overlapping sky region with data at all \( n_f \) frequencies according to some accuracy criteria (see de Oliveira-Costa et al. 2008). We then fit for the \( n_c \) principal component maps (the matrix product of the \( n_c \) principal components and the normalized maps, i.e., the input maps rescaled to have rms fluctuations of unity at each frequency) across the entire sky pixel by pixel by using the normalized input maps that have data for that pixel.

To predict sky maps at other frequencies, we can further fit the frequency dependence of both \( \log \sigma_i \) and each of the best \( n_c \) principal components with a cubic spline as a function of \( \log \nu \), and use them and the fitted \( n_c \) principal component maps to reconstruct the sky maps at the required frequencies. Here we implicitly assume that these are smooth, slowly varying functions (as shown in Figure 2). With this machinery, we can construct a map of the foreground emission in the target frequency range with \( n_f \) frequency bins in the target patches of sky with \( p \) pixels, from which we can estimate the frequency (pixel) covariance matrix, and subsequently the left (right) singular vectors of the foregrounds by an eigenvalue decomposition of the covariance matrix. Note that this singular-vector modeling process can be improved as more, higher-quality survey data at different frequencies become available. Advanced techniques presented in Zheng et al. (2017) may be also applied to better account for different survey data that has nonoverlapping regions.

As an example, we apply this method for modeling the singular vectors to the Tianlai array (CHEN 2012; Xu et al. 2015), which operates in the frequency band between 700 and 800 MHz with 256 frequency bins. We use the survey data at the same 11 frequencies (taken from the PyGSM package) as used in de Oliveira-Costa et al. (2008). The first three principal components with the largest three eigenvalues are plotted in the left panel of Figure 2. From this, we can predict the frequency dependence at the frequencies between 700 and 800 MHz where no survey data is available, by fitting each principal component with a cubic spline as a function of \( \log \nu \). Figure 2

\[ \text{https://github.com/telegraphic/PyGSM} \]
Figure 2. (Left) Three principal components that correspond to the largest three eigenvalues of the correlation matrix (from the large to small eigenvalues, in blue, green and red, respectively), using the sky region with data at all 11 frequencies (shown in dots). We fit the frequency dependence with a cubic spline in the frequency band of 700–800 MHz as enveloped by two black dashed lines. (Right) Three left singular vectors of the foregrounds corresponding to the largest three eigenvalues of the frequency covariance matrix that is constructed from the sky map of predicted foreground emission in the frequency band of 700–800 MHz. The eigenvectors are unitless and normalized to unity.

3.5. What if Sky Model Is Biased?

In practical analysis, it is likely that we employ a biased foreground model. In this subsection, we discuss how uncertainties in the sky model impact the foreground subtraction with SVP.

Suppose we employ a foreground model $F'$, which is biased from the true foreground $F$ by $F = F' + \Delta F$. We assume the foregrounds are decomposed using SVD by $F = U_S V^T_F$, $F' = U_S' V^T_F$, and $\Delta F = U_S S' V^T_F$. With this sky model, we apply the biased singular vectors $U'_S$ and $V'_F$, instead of $U_S$ and $V_F$, to the observational data $D$ using Equations (20)–(23)

Since $D = F + N = F' + (N + \Delta F)$, we can simply replace $N$ in Equations (24)–(27) by $N + \Delta F$, and find that

$$N'_{\text{L}} = (N - U'_S V^T_{F'}) N + (\Delta F - U'_S V^T_{F'}\Delta F), \tag{39}$$

$$N'_{\text{R}} = (N - (N - V^T_{F'}) V_{F'}) + (\Delta F - \Delta F V^T_{F'} V_{F'}), \tag{40}$$

$$N'_{\text{B}} = (N - U'_S U^T_{F'} N V^T_{F'}) + ( \Delta F - U'_S U^T_{F'} \Delta F V^T_{F'} V_{F'}), \tag{41}$$

$$N'_{\text{D}} = (N - U'_S (U^T_{F'} N V^T_{F'}))_{\text{diag}} V^T_{F'} + ( \Delta F - U'_S (U^T_{F'} \Delta F V^T_{F'}))_{\text{diag}} V^T_{F'}. \tag{42}$$

The first bracketed terms in Equations (39)–(42) should approximate to the corresponding terms given in Equations (24)–(27), if the foreground model $F'$ is not biased too much from the true foreground $F$. Regarding the second bracketed terms (i.e., foreground mixing terms) in Equations (39)–(42), we first consider two extreme cases.

(i) The singular vectors of the offset $\Delta F$ are aligned with those of the foreground model $F'$, i.e., $U'_S U^T_{F'} U_{\Delta F} = U_{\Delta F}$, $V^T_{\Delta F} V_{\Delta F} = V^T_{\Delta F}$, which means that $U'_S U^T_{F'} \Delta F = \Delta F$, and $\Delta F V^T_{\Delta F} = \Delta F$, and therefore $U'_S U^T_{F'} \Delta F V^T_{\Delta F} = \Delta F$. This means that the second bracketed terms in Equations (39)–(41) vanish (but the second bracketed term in Equation (42) is not necessarily zero).

A stronger example is the case where the foreground model is only off the true foreground by an overall factor, i.e., $F' = (1 - \alpha) F = U'_S [(1 - \alpha) S'_S] V^T_F$ and $\Delta F = \alpha F = U'_S (\alpha S'_S) V^T_F$, where $\alpha$ is a real number. This means that the singular vectors of $F'$, $\Delta F$, and $F$ are just the same, and the overall bias factor of foregrounds only affects the singular values. In this case, $U'_S U^T_{F'} \Delta F = \Delta F V^T_{F'}$ and $U'_S U^T_{F'} \Delta F V^T_{F'} = U'_S U^T_{F'} \Delta F V^T_{F'} = U'_S U^T_{F'} \Delta F V^T_{F'}$, and as such, all second bracketed terms in Equations (39)–(42) vanish, and we get exactly the same results as given in Equations (24)–(27). This just reflects the fact that the overall magnitude of the foreground does not impact the performance of the SVP method.

(ii) The singular vectors of the offset $\Delta F$ are orthogonal to those of the foreground model $F'$, i.e., $U^T_{F'} U_{\Delta F} = 0$, $V^T_{\Delta F} V_{\Delta F} = 0$, which means that $U'_S U^T_{F'} \Delta F = 0$, $\Delta F V^T_{F'} = 0$, and $U'_S U^T_{F'} \Delta F V^T_{F'} = 0$, and thus the second bracketed terms in Equations (39)–(42) are equal to $\Delta F$ due to orthogonality. The incomplete foreground model already discussed in Section 3.2 is such an example, where $F' = F_1$ and $\Delta F = F_2$, and $U'_S = U_S$, $U_{\Delta F} = U_{\Delta F}$. Like in Section 3.2, the “orthogonally” biased foreground model just introduces an additional term $\Delta F$ in the final results.

More generally, the impact of the biased foreground model should be between the above two extreme cases, i.e., the foreground mixing introduced by the SVP method should not be larger than the offset $\Delta F$ of the foreground model. For the rest of this paper, we neglect this systematic error introduced by the biased foreground model.
4. Simulation Test

4.1. Simulation Setup

In this section, we test the performance of SVP in terms of recovery errors with simulation data. We use the CORA\(^8\) (Shaw et al. 2014, 2015) package to generate the simulated data set, which includes the H I 21 cm signal and the mock foregrounds with two dominant components—galactic synchrotron emission and extragalactic point sources. We assume the Planck 2013 cosmological model (Planck Collaboration et al. 2014).

For the H I emission, the 21 cm power spectrum is given by

\[
P_T(k; z, z') = T_b(z)T_b(z')(b + f_j k^2)P_m(k; z, z'),
\]

where \(b\) is the bias, \(f\) is the growth rate, and \(P_m(k; z, z') = P(k)D_*(z)D_*(z')\) is the real-space matter power spectrum, where \(D_*\) is the growth factor normalized such that \(D_*(0) = 1\). The mean brightness temperature takes the form (Chang et al. 2008)

\[
\bar{T}_b(z) = 0.3\left(\frac{\Omega_H}{10^{-32}}\right)^{1/2} \left(\frac{1 + z}{2.5}\right)^{1/2} \times \left[\frac{\Omega_m + (1 + z)^3\Omega_{\Lambda}}{0.29}\right]^{-1/2} \text{mK}.
\]

We adopt the typical values of CORA parameters: \(\Omega_H = 0.2\) and \(b = 6.2 \times 10^{-3}\) (Switzer et al. 2013) and \(b = 1\).

The 21 cm angular power spectrum is given by Datta et al. (2007) \(C_l(\Delta \nu) \propto \int k^2dk_j(k_j\chi_j)P_T(k; z, z')\), where \(\Delta \nu = \nu' - \nu\). Here, \(\chi (\chi')\) is the comoving distance to redshift \(z (z')\) that corresponds to the frequency \(\nu (\nu')\). In the flat-sky approximation that is accurate to the percentage level, the 21 cm angular power spectrum is (Datta et al. 2007; Shaw et al. 2014)

\[
C_l(z, z') = \frac{1}{\pi \chi'} \int_0^\infty dk_j \cos(k_j \Delta \chi) P_T(k; z, z'),
\]

where \(\Delta \chi = \chi - \chi'\). In the integration, the vector \(k = (k_i, k_l = l/\bar{\chi})\), where \(\bar{\chi} = (\chi + \chi')/2\).

To model the foregrounds, for simplicity, we only consider the foregrounds with two main sources at low frequencies—the galactic synchrotron radiation and the extragalactic radio point sources, and ignore other minor sources such as free–free emission and dust emission. The angular power spectra of the foregrounds with these two main sources can be modeled in the form of

\[
C_l(\nu, \nu') = A \left(\frac{1}{100}\right)^{-\alpha} \left(\frac{\nu \nu'}{\nu_0^2}\right)^{-\beta} e^{-\frac{1}{2} \ln^2(\nu/\nu')},
\]

where we choose the pivot frequency \(\nu_0 = 130\) MHz (Santos et al. 2005). In principle, \(\xi\) is a function of \(l\). For simplicity, CORA assumes that \(\xi\) is independent of \(l\). We use the recalibrated model parameters for the 700–800 MHz band of H I intensity mapping experiment. The CORA package also implemented the polarized emission model, but for simplicity we only consider the total intensity model. We follow Shaw et al. (2014) for choosing the valid models, as listed in Table 1.

To generate the galactic synchrotron emission, CORA uses the processed 408 MHz Haslam map (with bright point sources and striping removed; Haslam et al. 1982; Remazeilles et al. 2015) as a template, and extrapolate it to other frequencies using a spectral index from the global sky model (GSM; de Oliveira-Costa et al. 2008), with a Gaussian random realization that is consistent with the angular power spectra of the foregrounds (Equation (46)) and adds fluctuations in frequency and on small angular scales. The extragalactic point-source simulations come from three components: a population of bright point sources \((S > 10\) Jy at 151 MHz\), a synthetic population of dimmer sources down to 0.1 Jy at 151 MHz, and an unresolved background of dimmer sources \((S < 0.1\) Jy\) modeled as a Gaussian random realization from Equation (46) with the point-source model parameters listed in Table 1.

We generate all components of foregrounds in each of 256 frequencies uniformly sampled between 700 and 800 MHz. For visualization, we show these simulated components only at the central frequency 750 MHz in Figure 3.

To include instrumental effects in the real observation, the generated sky maps are convolved with a symmetric circular frequency-dependent Gaussian beam with FWHM = 1.22 \(\lambda/D\), where \(\lambda\) is the observing wavelength, and \(D\) is the diameter of a telescope. We assume \(D = 100\) m, which is about the dish size of the GBT, currently the largest fully steerable telescope (Chang et al. 2010; Masui et al. 2013; Switzer et al. 2013), and also the optimal dish size for the mid-redshift 21 cm intensity mapping experiments (Chang et al. 2008; Seo et al. 2010; Ansari et al. 2012).

4.2. Results

We apply the (blind) PCA/SVD method and the (semiblind) SVP method to the simulated data set, and test their performance using the measures of the \(l_2\)-norm, power spectrum, and Pearson correlation coefficient, as follows. For the PCA/SVD, the largest five PCA modes are removed, because these five modes dominate our data set when instrumental effects are included. For the SVP, we assume that the left and/or right singular vectors \((U_i\) and/or \(V_j\) for all or at least some of the largest modes are known a priori.

4.2.1. \(l_2\)-norm

The \(l_2\)-norm of a \(1 \times p\) vector \(x\) is defined as

\[
\|x\| = \sqrt{x \cdot x^T} = \sqrt{\sum_{i=1}^p x_i^2}.
\]

At a given frequency, the recovery error can be treated as a \(1 \times p\) vector, so we compute the \(l_2\)-norm of recovery error as a function of frequency in Figures 4 and 5, as a statistical measure of the performance of different estimators. A smaller \(l_2\)-norm means better recovery of the true signal.

In Figure 4, we assume that the left and/or right singular vectors \((U_i\) and/or \(V_j\)) for all modes are known a priori. We further consider the scenario of incomplete information of

---

\(\text{Table 1}

| Component     | Polarization | \(A(K^2)\)  | \(\alpha\) | \(\beta\) | \(\xi\) |
|---------------|--------------|-------------|-----------|-----------|--------|
| Galaxy        | TT           | \(6.6 \times 10^{-3}\) | 2.80      | 2.8       | 4.0    |
| Point sources | TT           | \(3.55 \times 10^{-4}\) | 2.10      | 1.1       | 1.0    |

---

\(\text{\(8\) https://github.com/radiocosmology/cora}\)
singular vectors in Figure 5, where we assume that only the largest five left and/or right singular vectors of the foregrounds ($U_f$ and/or $V_f$) are known a priori. Both Figures 4 and 5 show that the $l_2$-norm of recovery error for $N_D$ is the smallest ($\sim 10^{-4}$) at almost all frequencies; the $l_2$-norm of $N_R$ and that of $N_L$ are comparable ($\sim$ a few of $10^{-3}$) but both larger than $N_D$; the $l_2$-norm of $N_L$ and that of PCA/SVD are the largest ($\sim$ a few of $10^{-2}$). This demonstrates that the SVP estimators except for $N_L$ perform better than the PCA/SVD estimator, generally. In particular, the $N_D$ estimator can reduce the $l_2$-norm of recovery error by 2 orders of magnitude over the PCA/SVD method. Also, these results agree with the relations found in the Appendix: $\|\Delta N_D\| \leq \|\Delta N_R\| \leq \|\Delta N_L\|$ and $\|\Delta N_L\| \leq \|\Delta N_R\| \leq \|\Delta N_D\|$. Roughly speaking, this implies that exploiting additional information in the singular vectors can improve the accuracy of signal recovery over the blind PCA/SVD method for foreground subtraction, and exploiting more information (in both the left and right singular vectors) is better than only using partial information (in either the left or right singular vector).

We note that if only a small number of modes of singular vectors corresponding to the largest singular values is exploited in SVP, there is residual foreground mixing (the $F_2$ term in Equation (38)) that contributes to the recovery error. However, comparing the results of Figures 4 and 5, we find the $l_2$-norm of recovery error for the SVP estimators with only the largest five modes of singular vectors is at the same level as that with full information of singular vectors. This is encouraging because in real observations, it is more likely to only obtain the information of the largest few modes than that of all modes.

We further explore the optimal number of retained modes for the SVP estimator $N_D$ in Figure 6, and find that the $l_2$-norm of recovery error converges when the number of largest retained modes exceeds five. This convergence allows for realistic application of the SVP method.

4.2.2. Power Spectrum

In Figure 7, we plot the 1D power spectrum along the line of sight (LOS) of the 21 cm signal, $P_{21}(k_l)$, for the input, true
signal $N$ as the benchmark, and for the recovered signal using the PCA/SVD and SVP estimators. For the SVP estimators, we only exploit the largest five left and/or right singular vectors of the foregrounds, i.e., $U_i$ and $V_i$. We find that these estimators can recover the input power spectrum well at small scales $k_i > 0.1 \, h \, \text{Mpc}^{-1}$. On the other hand, at large scales, while the PCA/SVD method and the $N_L$ estimator lose the power significantly, the $N_D$, $N_B$, and $N_R$ estimators can recover the input power spectrum at high accuracy. In comparison, the $N_D$ estimator performs the best, with the absolute difference $<0.001 \, \text{mK}^2 \, h^{-2} \, \text{Mpc}$ and relative error $<0.01\%$. These results are consistent with the findings in the $l_2$-norm test.

4.2.3. Pearson Correlation Coefficient

The Pearson correlation coefficient $r$ is a statistical measure of the degree of linear correlation between two signals. For the input $1 \times p$ vector $x$ and the recovered vector $\hat{x}$, it is defined as

$$r = \frac{\Delta x \cdot \Delta \hat{x}^T}{\sqrt{\Delta x \cdot \Delta x^T} \sqrt{\Delta \hat{x} \cdot \Delta \hat{x}^T}} = \frac{\sum (x_i - \bar{x})(\hat{x}_i - \bar{\hat{x}})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (\hat{x}_i - \bar{\hat{x}})^2}}. \quad (47)$$

where $\Delta x = x - \bar{x}$, and $\Delta \hat{x} = \hat{x} - \bar{\hat{x}}$. Here, $\bar{x}$ is the mean of $x$, and $\bar{\hat{x}}$ is the mean of $\hat{x}$. The value of $r$ is close to unity if two signals are very correlated.

We compute the Pearson correlation coefficient $r$ between the input, true 21 cm signal and the recovered signal using different estimators. For visualization purposes, we plot the value of $1 - r$ in Figure 8, because the values of $r$ are all close to unity, but how $1 - r$ is close to zero shows the degree of similarity between the recovered signal and the true signal. We find that the values of $1 - r$ for the $N_D$, $N_B$, and $N_R$ estimators are comparable ($\sim 10^{-5} - 10^{-6}$) and significantly smaller than that of the PCA/SVD and the $N_L$ estimator ($\sim 10^{-1} - 10^{-2}$). Here we only exploit the largest five left and/or right singular vectors of the foregrounds for the SVP estimators. These results are consistent with the results in the $l_2$-norm and power spectrum tests.

5. Conclusions

Foreground subtraction is one of the key challenges to the 21 cm observations due to the fact that the foregrounds are 4–5 orders of magnitudes larger than the cosmic 21 cm signal. In this paper, we show that the PCA method and the SVD method, which are widely employed for foreground subtraction, are actually equivalent in principle. We also provide the conditions in which the PCA/SVD method can separate the foregrounds from the 21 cm signal completely, i.e., with zero residual foreground mixing and zero signal loss. Nevertheless, we point out that, in general, the foreground mixing and signal loss are
unavoidable for the PCA/SVD method, because those conditions are hardly satisfied in practice.

In this paper, we propose a new class of semiblind method for foreground subtraction, based on the PCA/SVD method, called the singular vector projection (SVP). The SVP method is semiblind in the sense that it exploits a priori information of the left and/or right singular vectors of the foregrounds—if only the left (right) singular vector is known, then the estimator \( N_L \) (\( N_R \)) can be employed; if both the left and right singular vectors are known, then two estimators, \( N_B \) and \( N_D \), can be employed. The virtue of SVP is that, in principle, the residual foreground mixing is zero for all four SVP estimators, if the information of singular vectors for all modes is known.

We generate the mock maps of the 21 cm signal and the foregrounds from simulations, and use them to test the performance of the SVP estimators and the standard PCA/SVD estimator in terms of the Frobenius norm of recovery error, the LOS power spectrum of the 21 cm signal, and the Pearson correlation coefficient between the input, true signal, and the recovered signal. We find that while the results of \( N_D \) estimator are comparable to those of the PCA/SVD method, the other SVP estimators (\( N_L \), \( N_R \), and \( N_D \)) can significantly improve the accuracy of recovery by orders of magnitude. In particular, the \( N_D \) estimator performs the best in general.

We also consider the more realistic scenario of incomplete foreground information, in which only the largest few modes of the left and/or right singular vectors of the foregrounds are known a priori. In this case, there is residual foreground mixing due to the other small modes of residual foregrounds. However, the accuracy of recovery with the SVP estimators, if the largest five modes of the foregrounds are given in our demonstration, is not degraded with respect to the case with the information of all-mode singular vectors. This indicates that the SVP estimator reaches a balance between signal loss and residual foreground mixing.

Regarding the availability of a priori information, the left and right singular vectors can be solved by eigen-decomposition of the frequency and pixel covariance matrices of the foregrounds, respectively. Specifically, an accurate frequency covariance matrix of the foregrounds can be obtained with low cost by observations at other frequency bands. Our test results reflect the fact that the frequency spectrum information of the foregrounds is insufficient for the 21 cm experiments to reach high precision, and the spatial information of the foregrounds should be taken into account as well.

However, accurate foreground modeling is still a challenge to 21 cm observations. For example, Haslam and other maps may have offsets and/or incorrect scaling and may typically double-count unresolved or pointlike sources (Monsalve et al. 2021). With the SVP, the impact of the biased foreground model is restricted to be smaller than the offset between the foreground model and the true foreground. In particular, the SVP estimators are independent of the overall magnitude of the covariance matrix, which can be highly biased. This may be an advantage against some other semiblind methods that depend on the overall magnitude of the foreground covariance matrix, e.g., the KL transform method (Shaw et al. 2014, 2015).

When such a priori information of the foregrounds is available, our paper suggests that the SVP estimators can improve the recovery results significantly over the standard PCA/SVD method, which is blind against the prior information. In particular, the right singular vector of the foregrounds can help improve the foreground subtraction in a more effective manner than the left singular vector. Furthermore, combining both the left and right singular vectors, the \( N_D \) estimator works the best. These SVP estimators provide a new, effective approach for 21 cm observations to remove foregrounds and uncover the physics in the cosmic 21 cm signal.

This work is supported by the National SKA Program of China (grant No. 2020SKA0110401), NSFC (grant Nos. 11821303 and 11633004), National Key R&D Program of China (grant No. 2018YFA0404502), the MOST inter-government cooperation program China-South Africa Cooperation Flagship project (grant No. 2018YFE010800), the Chinese Academy of Sciences (CAS) Frontier Science Key Project (grant No. QYZDJ-SSW-SLH017), and the CAS Strategic Priority Research Program (grant No. XDA15020200). We thank Fengquan Wu and Yichao Li for useful discussions and their help. We acknowledge the Tsinghua Astrophysics High-Performance Computing platform at Tsinghua University for providing computational and data storage resources that have contributed to the research results reported within this paper.

Software: CORA (Shaw et al. 2014), PyGSM (de Oliveira-Costa et al. 2008; Zheng et al. 2017), hSpy (Collette et al. 2021), healpy (Zonca et al. 2019), Matplotlib (Hunter 2007), NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020).

Appendix

Inequalities for Signal Loss

In this section, we prove some inequalities for the signal loss using the SVP estimators. These inequalities focus on the comparison of the “magnitude” of signal loss.

We quantify the “magnitude” of signal loss using the Frobenius norm (Noble & Daniel 1977). For an \( m \times n \) matrix...
A, its Frobenius norm $\|A\|_F$ is defined as

$$\|A\|_F = \sqrt{\text{Tr}(A^T A)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}. \quad (A1)$$

We first prove three inequalities of Frobenius norm.

(1) $$\|AB\|_F^2 \leq \|A\|_F^2 \|B\|_F^2. \quad (A2)$$

The proof is as follows:

$$\|AB\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left[ \sum_{k=1}^{p} a_{ik} b_{kj} \right]^2$$

$$\leq \sum_{i=1}^{m} \sum_{j=1}^{n} \left[ \left( \sum_{k=1}^{p} |a_{ik}|^2 \right) \left( \sum_{l=1}^{p} |b_{lj}|^2 \right) \right]$$

$$= \left( \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right) \left( \sum_{i=1}^{m} \sum_{j=1}^{n} |b_{ij}|^2 \right)$$

$$= \|A\|_F^2 \|B\|_F^2. \quad (A3)$$

The second line in Equation (A3) uses the Cauchy-Schwarz inequality.

(2) For a square matrix $S$, it is straightforward by its definition to prove

$$\|S_{\text{diag}}\|_F \leq \|S\|_F. \quad (A4)$$

(3) For a partial orthogonal matrix $U,$

$$\|AU\|_F \leq \|A\|_F. \quad (A5)$$

The proof is as follows:

$$\|AU\|_F^2 = \text{Tr}(U^T A^T A U) = \text{Tr}(A^T A U U^T)$$

$$= \sum_{i} \sum_{j} (A^T A)_{ij} (U U^T)_{ij}$$

$$\leq \sum_{i} \sum_{j} (A^T A)_{ij} \delta_{ij} = \sum_{i} (A^T A)_{ii}$$

$$= \text{Tr}(A^T A) = \|A\|_F^2. \quad (A6)$$

Using these properties, we have

$$\|N_{\text{loss}}^D\|_F^2 = \|U_f U_f^T N_f V_f V_f^T\|_F^2$$

$$= \text{Tr}(V_f V_f^T N_f^* U_f U_f^T N_f V_f V_f^T)$$

$$= \text{Tr}(V_f V_f^T N_f^* U_f U_f^T N_f V_f V_f^T)$$

$$= \|U_f^T N_f V_f\|_F^2 \quad (A7)$$

$$\leq \|U_f^T N_f\|_F^2 = \|N_{\text{loss}}^L\|_F^2 \quad (A8)$$

Similarly,

$$\|N_{\text{loss}}^D\|_F^2 = \|U_f^T N_f V_f\|_F^2 \leq \|N_f\|_F^2 = \|N_{\text{loss}}^R\|_F^2. \quad (A9)$$

Also,

$$\|N_{\text{loss}}^D\|_F^2 = \|U_f (U_f^T N_f V_f)_{\text{diag}} V_f^T\|_F^2$$

$$= \text{Tr}(V_f (U_f^T N_f V_f)_{\text{diag}} U_f^T (U_f^T N_f V_f)_{\text{diag}} V_f^T)$$

$$= \|U_f^T N_f V_f\|_F^2 \quad (A10)$$

Altogether, the following inequalities are proven:

$$\|N_{\text{loss}}^D\|_F \leq \|N_{\text{loss}}^L\|_F \leq \|N_{\text{loss}}^R\|_F. \quad (A11)$$

$$\|N_{\text{loss}}^D\|_F \leq \|N_{\text{loss}}^L\|_F \leq \|N_{\text{loss}}^R\|_F. \quad (A12)$$

Since there is no foreground mixing for the SVP estimators if all modes of the left and/or right singular vectors of foregrounds are known a priori, the inequalities for signal loss are equivalent to the inequalities for recovery error, $\|\Delta N_D\|_F \leq \|\Delta N_L\|_F \leq \|\Delta N_R\|_F$.
