Received 22 December 2022; accepted 12 January 2023. Date of publication 25 January 2023; date of current version 15 February 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3238772

Fronthaul Compression for Uplink Massive MIMO Using Matrix Decomposition

P. ASWATHYLAKSHMI (Graduate Student Member, IEEE), AND RADHA KRISHNA GANTI (Member, IEEE)

Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai 600036, India

CORRESPONDING AUTHOR: P. ASWATHYLAKSHMI (e-mail: aswathylakshmi@ee.iitm.ac.in)

This work was supported in part by the Department of Telecommunications (DoT), Government of India, through the project titled, "Next Generation Wireless Research and Standardization on 5G and Beyond" and in part by the ANSYS Software Pvt. Ltd., through their Doctoral Fellowship Award Program.

A preliminary version of this work was presented at the IEEE Wireless Communications and Networking Conference, Austin, TX, USA, in April 2022 [1] [DOI: 10.1109/WCNC51071.2022.9771783].

ABSTRACT Massive multiple-input-multiple-output (MIMO) is a key enabler for obtaining higher data rates in the next generation wireless technology. While it has the power to transform cellular communication, with potential for spatial diversity and multiplexing, a bottleneck that often gets overlooked is the fronthaul capacity. The fronthaul link that connects a massive MIMO Remote Radio Head (RRH) and carries in-phase and quadrature (IQ) samples to the Baseband Unit (BBU) of the base station can throttle the network capacity/speed if appropriate data compression techniques are not applied, particularly in the uplink. This paper proposes an iterative technique for fronthaul load reduction in the uplink for massive MIMO systems utilizing the convolution structure of the received signals. The proposed algorithm provides compression ratios of about 30-50×. This work provides extensive analysis of the performance of the proposed method for a plethora of practical scenarios and constraints, such as different channel parameters and models, receive antenna correlation, and under imperfect channel information. It also discusses the numerical convergence and complexity of the proposed algorithm and compares the performance against other existing compression techniques.

INDEX TERMS Massive MIMO, 5G networks, fronthaul compression, iterative technique, alternating minimization, blind matrix deconvolution.

I. INTRODUCTION

MASSIVE MIMO is slated to change wireless communication experience with unprecedented speeds, network capacity and coverage [2]. The 5G New Radio (NR) wireless standards support massive MIMO for a wide variety of applications and deployment scenarios [3]. Next generation wireless systems will likely capitalise on this framework and use massive MIMO to enhance user experience and improve system performance with the use of large-scale antenna arrays in a variety of configurations [2]. Large scale antenna arrays can provide huge diversity gain and beamforming opportunities for millimeter wave applications as well as spatial multiplexing to support a large number of users or offer massive data rates with several MIMO layers for a few users [4]. However, one major roadblock to achieving these goals is the fronthaul data rate [5].

The fronthaul connects the radio frequency (RF) components of the base station, termed the Remote Radio Head (RRH), to its baseband processing unit (BBU), typically using optical fibre cables (Fig. 1) [6]. The separation between the RRH and the BBU can range from a few metres to several kilometres depending on the deployment scenario or network architecture. For example, massively distributed MIMO deployments offer enhanced coverage for edge users by locating antennas over a wide area which are connected to a central server [2]. Such an architecture not only reduces the capital expenditure for network operators, since antenna installations are inexpensive and easy to maintain, but it also allows for cooperative interference management strategies in

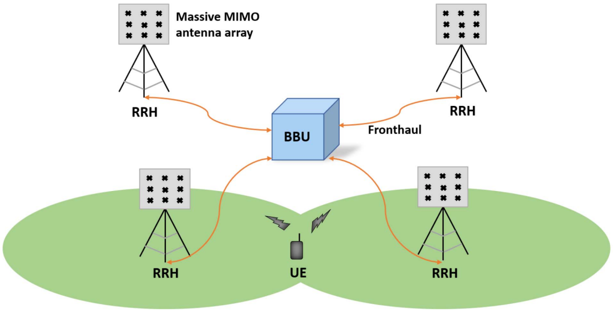


FIGURE 1. Massive MIMO architecture with several Remote Radio Heads (RRH) connected via fronthaul links to a central Baseband Unit (BBU), which can coordinate interference management and offer improved coverage to cell-egde user equipments (UE).

the network [6]. Since the fronthaul load scales up with the number of RRH antennas, laying high speed optical fibres for increasing antenna elements will increase the cost for network operators significantly. For instance, sending raw IQ samples from the RRH to the BBU requires a data rate upto 236 Gbps for 100 MHz bandwidth for a 64-antenna base station [7]. Given that the economical optical fibre cables and optical electronics used for fronthaul data transport have capacities in the range of only 10-40 Gbps, approximately 10× compression is required to make such MIMO architectures economically feasible for the operators.

Numerous compression techniques have been proposed to reduce fronthaul load. In the downlink, data is to be compressed at the BBU and transmitted to the RRH via the fronthaul. The BBU has complete knowledge of the data being transmitted and its parameters, such as the modulation order, the pilot symbols, the subcarrier allocation for data and pilots, the number of users, the number of MIMO layers, etc., which can be effectively utilized to achieve high compression rates. However, in the uplink direction, such knowledge is not available for compression at the RRH unless the entire receiver processing chain is implemented at the RRH. This makes achieving high compression rates for uplink fronthaul a challenging task.

An important factor that affects the fronthaul data rate is the functional split between the RRH and the BBU [7]. The Open Radio Access Network (O-RAN) Alliance outlines multiple ways in which the transmitter/receiver processing chain can be split between the BBU and the RRH and how each of these splits can be implemented/managed for different use cases. Fig. 2 outlines the different functional splits possible in the uplink [8]. The higher level splits are simpler to implement since the RRH does not require expensive computational resources for these splits. However, the further one moves down the receiver processing chain, the higher the compression ratios that can be obtained as one gets closer to decoding the received data. The highest compression ratio is obtained if the receive chain is fully implemented at the RRH, but this implies one cannot take advantage of cooperative interference management strategies such as mitigating pilot contamination that a centralized BBU can offer. The most popular split is the 7.2 split as it offers a good trade-off between cost, compression ratio and centralization, and is used in [9]. But further increase in antenna elements as envisioned for 6G renders even 5× compression [9] insufficient. The ideal compression method would be one that offers both the high compression ratio provided by near completion of the receive chain at the RRH, but can also take advantage of centralized processing. This is precisely the aim of this work.

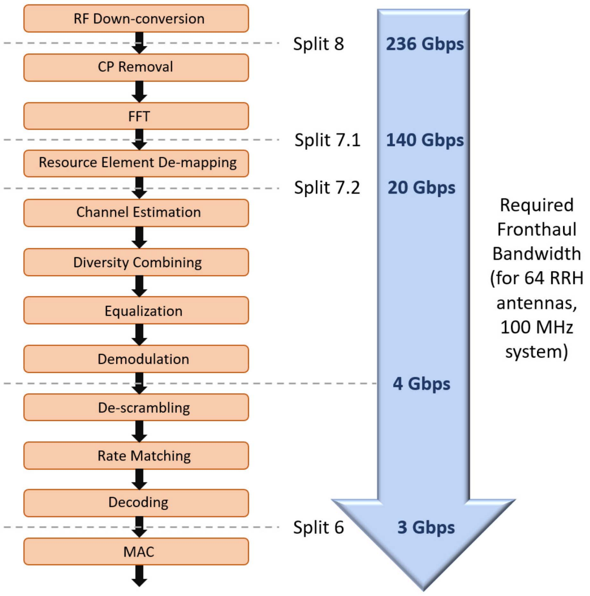


FIGURE 2. Possible functional splits between the RRH and the BBU and the resulting fronthaul bandwidths required.

Most fronthaul load reduction techniques for the uplink exploit either the redundancies in the received waveform [10], [11], a priori knowledge of the characteristics of the received signal [12] or the correlation between the base station antennas [9], [13]. A technique that achieves a compression rate of 1/2 by removing redundant spectrum and reducing quantization bit-width is proposed in [10]. A lossy compression method is discussed in [11] which applies Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT) to the received signals and discards low power frequency coefficients. Though these methods have the advantage of low implementation complexity, they do not offer high compression ratios. On the other hand, compressive sensing techniques that make use of the sparsity of uplink signals offer slightly better but varying compression ratios depending upon the frequency of bursty transmissions from the user equipments (UEs) [12]. Antenna selection, wherein a subset of antennas is selected for reception at a time [14], can be applied to reduce the number of RF chains and consequently the fronthaul bandwidth required in the uplink. Although this can offer significant compression ratios, it can lead to loss in diversity or multiplexing capacity since the antenna array is not fully utilized. The principal component analysis (PCA) compression performed on the matrix of received signals in [13] utilizes the spatio-temporal correlation of the signals. Similarly, low-rank approximation of the received signal matrix via QR decomposition is used in [9] to reduce the fronthaul load. Although the methods in [9], [13] achieve better compression ratios than those proposed in [10], [11] or [12] at the cost of increased computations, the highest compression ratio they achieve for 64 antennas is 5×.

The goal in this work is to reduce the number of samples that have to be transmitted from the RRH to the BBU via fronthaul. For this, a block of N complex time-domain samples received at the N_r antennas at the RRH are considered together to exploit their joint characteristics for the compression. These samples can be arranged in the form of an $N \times N_r$ matrix. For example, N can be the number of samples received in the duration of one orthogonal frequency division multiplexing (OFDM) symbol, which is the FFT size used for the OFDM. In this paper, a massive MIMO system in which N_r is higher than the number of users simultaneously served (in a transmission time interval) is considered [15]. The columns of the received signal matrix at the RRH are the result of the user data sequence convolved with a different multi-path channel response corresponding to each receive antenna in the time domain. In the frequency domain, this received signal matrix can be expressed as the product of an N-dimensional diagonal matrix of user data and the Fourier transform of the $L \times N_r$ channel matrix, where L is the number of significant multi-paths. If $L \ll N_r$, as is typical for a well-designed massive MIMO system, representing the received signal matrix in terms of the N-length user data sequence and the $L \times N_r$ channel matrix allows its recovery at the BBU using far fewer samples than NN_r. Towards this end, a blind deconvolution of the received signal matrix needs to be performed at the RRH. Noting that the N-point Fourier transform of the $L \times N_r$ channel matrix is low rank, given $L < \min\{N, N_r\}$, an iterative algorithm consisting of alternating minimization similar to the one in [16] used for low rank matrix sensing is presented in this work. However, the proposed algorithm is constructed differently from [16] to reflect the fact that only one of the deconvolved matrices (the channel matrix) is low rank while the other (the user data matrix) is a full rank diagonal matrix. Deconvolution using non-convex minimization has been considered in [17]. However, a different approach is adopted here to solve the objective function since a matrix of multiple sequences has to be deconvolved rather than a single sequence as in [17].

The problem of fronthaul compression has also been studied in the context of a Cloud Radio Access Network (C-RAN), where several RRHs are connected to a central BBU. A distributed spatial filtering scheme and joint optimization of users' power allocation, RRHs' spatial filter design and quantization bits allocation, as well as BBU's receive beamforming to maximise the minimum signal-to-interference-plus-noise ratio (SINR) of all the users in the uplink in a C-RAN is proposed in [18]. Reference [19] investigates the joint design of fronthaul compression and precoding, and solves the optimization problem of maximising the ergodic capacity for two downlink functional splits in C-RAN, compression-after-precoding (CAP) and compression-before-precoding (CBP). In [20], the optimal design for the multiplex-and forward (MF) and the decompress-process-and-recompress (DPR) backhaul schemes is investigated, with the aim of maximising the sum-rate under the backhaul capacity constraints for a multihop topology in a C-RAN. Reference [21] studies the joint design of precoding and backhaul compression strategies for downlink C-RAN. It aims to maximise the weighted sum-rate with respect to both the precoding matrix and the joint correlation matrix of the quantization noises subject to power and backhaul capacity constraints. These works ([18], [19], [20], [21]) look at the fronthaul compression problem with the aim of maximising the information theoretic capacity of multiplexed fronthaul links, subject to power/capacity constraints within a C-RAN framework. In contrast, the framework in this paper considers uplink data compression from a signal processing perspective rather than an information theoretic one. It aims to reduce the total number of digital samples that need to be sent in a single BBU-RRH fronthaul link. Moreover, the compression scheme presented here can be used at each RRH independent of the BBU or any number of other RRHs connected to the BBU. Therefore, the algorithm proposed in this paper can be applied to any type of network deployment/architecture.

The main contribution of this paper is to propose a novel fronthaul compression method for uplink massive MIMO that exploits the underlying convolution structure of the received signals. Although previous works have identified and utilized the low rank nature of the signals for their compression [9], [12], [13], these methods are agnostic to the convolution structure of the signals which gives rise to this low rank nature. The proposed method takes into account this signal structure to decompose the received signal into an approximate user signal matrix and a channel matrix. As a result, it offers a compression rate comparable to that achieved if the receive chain is fully implemented at the RRH, while still retaining the ability to leverage the advantages offered by centralization. Thus, it provides both orders of magnitude higher compression ratios than that offered by existing methods and a flexible functional split between the RRH and BBU as the iterative technique proposed in this paper can also be used for the blind demodulation of massive MIMO OFDM signals.

This paper is organized as follows: Section II outlines the system model, Section III presents the compression algorithm for single user and multi-user cases, and Sections IV–VI present the results of link level simulations of the algorithm for various scenarios. Section IV analyzes the robustness of the proposed method under different channel conditions, examines the impact of receive antenna correlation on the proposed method and evaluates the performance of the method for realistic channel models.

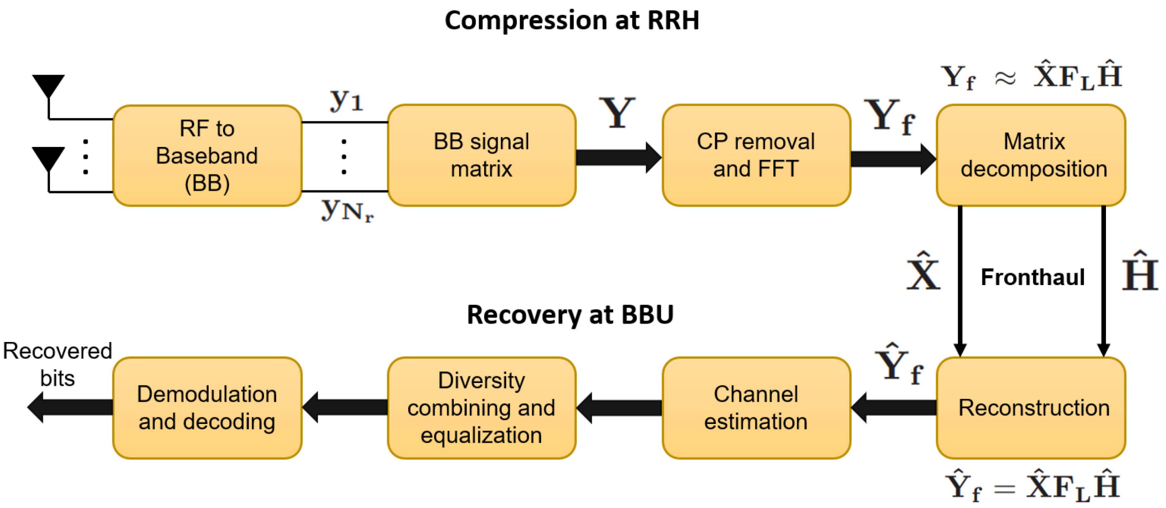


FIGURE 3. **Block diagram of the proposed fronthaul compression scheme, where the signals received over a duration of time across the N_r RRH antennas are compressed into an approximate user data matrix and a channel matrix, which are then used at the BBU for reconstruction and recovery of the received signals.**

Section V discusses the convergence and complexity of the proposed compression algorithm. Section VI compares its performance vis-á-vis the PCA compression in [13] and the uncompressed system. Finally, Section VII concludes the work.

II. SYSTEM MODEL

This paper considers a massive MIMO 5G base station RRH that has N_r antennas and receives signals from N_u users in the uplink at a given time slot. Since the uplink multiple access scheme in 5G is orthogonal frequency division multiple access (OFDMA) [22], the bit-stream from each user is mapped to an M-QAM (Quadrature Amplitude Modulation) symbol constellation followed by subcarrier mapping, inverse Fourier transform, and cyclic prefix (CP) addition. Let N denote the size of the FFT at the receiver. It is assumed that the channel between each user and each antenna has a maximum of L significant multi-paths in time-domain, and this L is known at the RRH. After sampling and removal of the CP, the signal received at antenna r at a sampling instant n is

$$y_r[n] = \sum_{u=1}^{N_u} x_u[n] \circledast h_{r,u}[n] + w_r[n],$$

where $x_u[n]$ is the OFDM symbol of user u, $h_{r,u}[n]$ is the multi-path channel response between user u and antenna r, $\circledast$ represents circular convolution resulting from the cyclic prefix in OFDM, and w_r is the additive white circularly symmetric complex Gaussian noise (AWGN) at antenna r with variance σ^2.

Consider a block of N samples in the time-domain received at the RRH after CP removal. The channel is assumed to be constant for the duration of these N samples. Let N be the size of FFT (or the size of one OFDM symbol). The signal received at the RRH across all the antennas is

$$\mathbf{Y} = \begin{bmatrix} y_1[1] & y_2[1] & . & . & . & y_{N_r}[1] \\ y_1[2] & y_2[2] & . & . & . & y_{N_r}[2] \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ y_1[N] & y_2[N] & . & . & . & y_{N_r}[N] \end{bmatrix}_{N \times N_r}.$$

Here, each column of $\mathbf{Y}$ represents the received signal at each antenna over N sampling instants. The data $\mathbf{Y}$ needs to be sent from the RRH to the BBU via the fronthaul link for its faithful reconstruction at the BBU.

III. COMPRESSION USING MATRIX DECOMPOSITION

This paper considers the convolution structure inherent in the received signal matrix $\mathbf{Y}$ for its compression at the RRH and reconstruction at the BBU, which is outlined in Fig. 3. This structure depends on the manner in which the N subcarriers are allocated to different users. Users can be allocated subcarriers in one of the following three ways:

1) A single user is allocated all the N subcarriers.
2) Multiple users share the same set of N subcarriers (overlapping allocation).
3) Multiple users are allocated subsets of the N subcarriers in a non-overlapping manner.

In the frequency domain, the matrix $\mathbf{Y}$ can be decomposed in different ways corresponding to the above three cases. This decomposition reveals the low rank nature of $\mathbf{Y}$ when the number of significant multi-paths, $L < \min\{N, N_r\}$, and forms the crux of the compression proposed in this work. For ease of analysis, compression in the above cases for users with single antenna is discussed first. Later in the section, it is shown how these can be extended to users with multiple antennas.

A. SINGLE USER CASE

The matrix $\mathbf{Y}$ can be expressed in the frequency domain by applying FFT to $\mathbf{Y}$, and using the subscript f to denote quantities in the frequency domain, as

$$\mathbf{Y_f} = \mathbf{X_f}\mathbf{H_f} + \mathbf{W_f}, \tag{1}$$

where $\mathbf{X_f}$ is the $N \times N$ diagonal matrix with the N-length M-QAM user data as its diagonal, $\mathbf{H_f}$ is the $N \times N_r$ multipath channel matrix in the frequency domain, and $\mathbf{W_f}$ is the noise in the frequency domain.

In a massive MIMO setting, the phenomenon of channel hardening decreases the delay spread of the channel [15]. This means that the channel has very few resolvable multipath components in the time-domain. If the channel is assumed to have L resolvable taps in the time-domain, and $\mathbf{H_t}$ denotes the $L \times N_r$ time-domain channel response for the N_r antennas, then the frequency domain channel, $\mathbf{H_f}$ can be expressed as the product $\mathbf{F_L}\mathbf{H_t}$, where $\mathbf{F_L}$ denotes L columns of the $N \times N$ Discrete Fourier Transform (DFT) matrix. Thus, the frequency-domain channel, $\mathbf{H_f}$ in (1) is the result of oversampling in the frequency domain due to the large FFT/OFDM symbol size used in 5G. Equation (1) can be expressed using the above as

$$\mathbf{Y_f} = \mathbf{X_f}\mathbf{F_L}\mathbf{H_t} + \mathbf{W_f}. \tag{2}$$

If $\mathbf{Y_f}$ is transmitted on the fronthaul without any compression, then NN_r complex samples are required to be sent via the fronthaul. However, (2) reveals that the signal component of $\mathbf{Y_f}$ can be reconstructed with the N non-zero samples corresponding to $\mathbf{X_f}$ and the LN_r samples corresponding to $\mathbf{H_t}$. Note that $L \ll N$ due to the channel hardening effect. Therefore, if $\mathbf{Y_f}$ can be decomposed into its signal component $\mathbf{X_f}$ and the *time-domain* channel $\mathbf{H_t}$, only $N + LN_r$ samples need to be sent from the RRH to the BBU, to recover $\mathbf{Y_f}$. Therefore, this gives the compression ratio for a single user (CR_{SU}) as

$$\text{CR}_{\text{SU}} = \frac{NN_r}{N + LN_r}. \tag{3}$$

An obvious way of obtaining $\mathbf{X_f}$ and $\mathbf{H_t}$ from $\mathbf{Y_f}$ is coherent demodulation of $\mathbf{Y_f}$. This would require pilots and modulation order to be known and the entire receive chain to be present at the RRH [23]. However, in general, this information is not available at the RRH [8]. The idea in this paper is to decompose $\mathbf{Y_f}$ into matrices $\hat{\mathbf{X}}$ and $\hat{\mathbf{H}}$ (not necessarily equal to $\mathbf{X_f}$ and $\mathbf{H_t}$) so that $\mathbf{Y_f} \approx \hat{\mathbf{X}}\mathbf{F_L}\hat{\mathbf{H}}$, without any knowledge of pilots/modulation order. Thus, for compressing $\mathbf{Y_f}$, a diagonal matrix $\hat{\mathbf{X}}$ and an $L \times N_r$ matrix $\hat{\mathbf{H}}$ that minimise $||\mathbf{Y_f} - \hat{\mathbf{X}}\mathbf{F_L}\hat{\mathbf{H}}||_F^2$ have to be found.

Minimising $||\mathbf{Y_f} - \hat{\mathbf{X}}\mathbf{F_L}\hat{\mathbf{H}}||_F^2$, where both $\hat{\mathbf{X}}$ and $\hat{\mathbf{H}}$ are unknown, is a non-convex optimization problem. However, fixing one of the unknowns reduces solving for the other into a simpler, convex problem. Therefore, this paper proposes using an iterative alternating minimization approach to find $\hat{\mathbf{X}}$ and $\hat{\mathbf{H}}$. The algorithm begins with an initial guess for $\hat{\mathbf{X}}$, denoted by $\hat{\mathbf{X}}_\mathbf{0}$. Then, in the first iteration, the estimate for $\hat{\mathbf{H}}$ is found by solving

$$\hat{\mathbf{H}}_\mathbf{k} = \underset{\mathbf{H}}{\operatorname{argmin}}\, ||\mathbf{Y_f} - \hat{\mathbf{X}}_\mathbf{k-1}\mathbf{F_L}\mathbf{H}||_F^2, \tag{4}$$

where k denotes the iteration number. This reduces to finding $\mathbf{H}$ such that $\mathbf{Y_f} = \hat{\mathbf{X}}_\mathbf{0}\mathbf{F_L}\mathbf{H}$. This can be solved using linear least squares as

$$\hat{\mathbf{H}}_\mathbf{k} = \left(\hat{\mathbf{X}}_\mathbf{k-1}\mathbf{F_L}\right)^{\dagger}\mathbf{Y_f}, \tag{5}$$

where † denotes the Moore-Penrose inverse. This can be interpreted as channel estimation. In the next step, this $\hat{\mathbf{H}}_\mathbf{k}$ is used to update the estimate of $\hat{\mathbf{X}}$ by solving

$$\hat{\mathbf{X}}_\mathbf{k} = \underset{\mathbf{X}}{\operatorname{argmin}}\, ||\mathbf{Y_f} - \mathbf{X}\mathbf{F_L}\hat{\mathbf{H}}_\mathbf{k}||_F^2. \tag{6}$$

The norm in (6) can be simplified by denoting $\mathbf{B} = \mathbf{F_L}\hat{\mathbf{H}}_\mathbf{k}$ for ease of notation, and using the fact that $\hat{\mathbf{X}}_\mathbf{k}$ needs to be diagonal. This leads to

$$\begin{aligned} ||\mathbf{Y_f} - \mathbf{X}\mathbf{F_L}\hat{\mathbf{H}}_\mathbf{k}||_F^2 &= ||\mathbf{Y_f} - \mathbf{X}\mathbf{B}||_F^2, \\ &= \sum_{r=1}^{N_r} ||\mathbf{y_r} - \mathbf{X}\mathbf{b_r}||_2^2, \\ &= \sum_{n=1}^{N}\sum_{r=1}^{N_r} |y(n,r) - x(n)b(n,r)|^2, \end{aligned} \tag{7}$$

where $\mathbf{y_r}$ and $\mathbf{b_r}$ denote the r-th columns of $\mathbf{Y_f}$ and $\mathbf{B}$, respectively, $y(n,r)$ and $b(n,r)$ denote the elements at n-th row and r-th column of $\mathbf{Y_f}$ and $\mathbf{B}$, respectively, and $x(n)$ denotes the n-th element on the diagonal of the matrix $\mathbf{X}$. The inner sum $\sum_{r=1}^{N_r} |y(n,r) - b(n,r)|^2$ can be optimized independently for each $n \in 1, 2, \ldots, N$ to obtain the n-th element of $\hat{\mathbf{X}}_\mathbf{k}$ as

$$x_k(n) = \frac{\left(\sum_{r=1}^{N_r} y(n,r)b^*(n,r)\right)}{\left(\sum_{r=1}^{N_r} |b(n,r)|^2\right)}, \tag{8}$$

where $b^*(n,r)$ denotes the complex conjugate of $b(n,r)$. The expression in (8) is equivalent to maximal-ratio combining (MRC).

At each iteration, the optimization problems in (4) and (6) are solved alternately using the closed form solutions given in (5) and (8) until the optimal solution to faithfully reconstruct $\mathbf{Y_f}$ within a pre-defined error tolerance ϵ is found. The above steps are summarized as Algorithm 1. Now, only the N samples corresponding to $\hat{\mathbf{X}}$ and the LN_r samples corresponding to $\hat{\mathbf{H}}$ are sent to the BBU to reconstruct $\mathbf{Y_f}$, leading to the compression ratio given in (3).

Algorithm 1 Alternating Minimization for Fronthaul Compression for Single User

```
Input $\mathbf{Y_f}$
Define error tolerance $\epsilon$
Initialize $\mathrm{diag}(\hat{\mathbf{X}}_\mathbf{0}) = \mathbf{y_1}$, $k = 1$, $\mathbf{B} = [\ ]$
repeat
    $\hat{\mathbf{H}}_\mathbf{k} \leftarrow (\hat{\mathbf{X}}_\mathbf{k-1}\mathbf{F_L})^\dagger \mathbf{Y_f}$
    $\mathbf{B} \leftarrow \mathbf{F_L}\hat{\mathbf{H}}_\mathbf{k}$
    for $n \leftarrow 1$ to $N$ do
        $x_k(n) = \dfrac{\left(\sum_{r=1}^{N_r} y(n,r)b^*(n,r)\right)}{\left(\sum_{r=1}^{N_r} |b(n,r)|^2\right)}$
    end for
    $k \leftarrow k + 1$
until $||\mathbf{Y_f} - \hat{\mathbf{X}}_\mathbf{k}\mathbf{F_L}\hat{\mathbf{H}}_\mathbf{k}||_F / ||\mathbf{Y_f}||_F < \epsilon$
return $\hat{\mathbf{X}}_\mathbf{k}$, $\hat{\mathbf{H}}_\mathbf{k}$
```

$\mathbf{y_1}$ denotes the first column of matrix $\mathbf{Y_f}$, † denotes matrix pseudo-inverse, $*$ denotes complex conjugation, $a(n, r)$ denotes element at row n and column r of matrix $\mathbf{A}$ and $x_k(n)$ denotes $n-$th diagonal element of $\hat{\mathbf{X}}_\mathbf{k}$.

B. MULTI-USER CASE - OVERLAPPING ALLOCATION

In this section, the signal model and fronthaul compression for multiple users who share the same set of N subcarriers (overlapping allocation) is discussed. When there are N_u users scheduled in the same set of N subcarriers, $\mathbf{Y}$ can be expressed in the frequency domain as

$$\mathbf{Y_f} = [\mathbf{X_f(1)}\mathbf{X_f(2)}, \ldots \mathbf{X_f(N_u)}] \begin{bmatrix} \mathbf{H_f(1)} \\ \mathbf{H_f(2)} \\ \vdots \\ \mathbf{H_f(N_u)} \end{bmatrix} + \mathbf{W_f}, \tag{9}$$

where $\mathbf{X_f(u)}$ is the $N \times N$ diagonal matrix, with the N-length M-QAM data of user u as its diagonal, and $\mathbf{H_f(u)}$ is the $N \times N_r$ multi-path channel matrix in the frequency domain for user u. Similar to the single user case, each $\mathbf{H_f(u)}$ can be expressed as the matrix product $\mathbf{F_L}\mathbf{H_t(u)}$, where $\mathbf{H_t(u)}$ is the $L \times N_r$ time-domain multi-path channel response for user u. Let $\mathrm{diag}(\mathbf{F_L})$ denote the block matrix of the form

$$\mathrm{diag}(\mathbf{F_L}) = \begin{bmatrix} \mathbf{F_L} & 0 & . & . & . & 0 \\ 0 & \mathbf{F_L} & . & . & . & 0 \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 0 & 0 & . & . & . & \mathbf{F_L} \end{bmatrix}_{NN_u \times LN_u} .$$

Then,

$$\begin{bmatrix} \mathbf{H_f(1)} \\ \mathbf{H_f(2)} \\ \vdots \\ \mathbf{H_f(N_u)} \end{bmatrix}_{NN_u \times N_r} = \mathrm{diag}(\mathbf{F_L}) \begin{bmatrix} \mathbf{H_t(1)} \\ \mathbf{H_t(2)} \\ \vdots \\ \mathbf{H_t(N_u)} \end{bmatrix}_{LN_u \times N_r} .$$

Equation (9) can be written using the above as

$$\mathbf{Y_f} = [\mathbf{X_f(1)}\mathbf{X_f(2)}, \ldots \mathbf{X_f(N_u)}]\mathrm{diag}(\mathbf{F_L}) \begin{bmatrix} \mathbf{H_t(1)} \\ \mathbf{H_t(2)} \\ \vdots \\ \mathbf{H_t(N_u)} \end{bmatrix} + \mathbf{W_f}. \tag{10}$$

Let $\hat{\mathbf{X}} = [\hat{\mathbf{X}}\mathbf{(1)}\hat{\mathbf{X}}\mathbf{(2)} \ldots \hat{\mathbf{X}}\mathbf{(N_u)}]$, where $\hat{\mathbf{X}}\mathbf{(u)}$ are $N \times N$ diagonal matrices, and $\hat{\mathbf{H}} = [\hat{\mathbf{H}}\mathbf{(1)}^\mathbf{T}\hat{\mathbf{H}}\mathbf{(2)}^\mathbf{T} \ldots \hat{\mathbf{H}}\mathbf{(N_u)}^\mathbf{T}]^T$ where $\hat{\mathbf{H}}\mathbf{(u)}$ are size $L \times N_r$ matrices, for $u = 1, 2, .., N_u$. The same iterative approach is used to find $\hat{\mathbf{X}}$ and $\hat{\mathbf{H}}$ that form the signal component of $\mathbf{Y_f}$ as was used for the single user case.

An initial point $\hat{\mathbf{X}}_\mathbf{0}\mathbf{(u)}$ is used in the first iteration, $k = 1$, in the minimization problem

$$\hat{\mathbf{H}}_\mathbf{k} = \underset{\mathbf{H}}{\mathrm{argmin}}\, ||\mathbf{Y_f} - \hat{\mathbf{X}}_\mathbf{k-1}\mathrm{diag}(\mathbf{F_L})\mathbf{H}||_F^2, \tag{11}$$

whose linear least squares solution is given by

$$\hat{\mathbf{H}}_\mathbf{k} = \left(\hat{\mathbf{X}}_\mathbf{k-1}\mathrm{diag}(\mathbf{F_L})\right)^\dagger \mathbf{Y_f}. \tag{12}$$

This $\hat{\mathbf{H}}_\mathbf{k}$ is then used to find

$$\hat{\mathbf{X}}_\mathbf{k} = \underset{\mathbf{X}}{\mathrm{argmin}} \left\| \mathbf{Y_f} - \mathbf{X}\mathrm{diag}(\mathbf{F_L})\hat{\mathbf{H}}_\mathbf{k} \right\|_F^2 . \tag{13}$$

Let $\mathbf{B(u)} = \mathbf{F_L}\hat{\mathbf{H}}_\mathbf{k}\mathbf{(u)}$ and $\mathbf{B} = \mathrm{diag}(\mathbf{F_L})\hat{\mathbf{H}}_\mathbf{k} = [\mathbf{B(1)}^\mathbf{T}\mathbf{B(2)}^\mathbf{T} \ldots \mathbf{B(N_u)}^\mathbf{T}]^T$ for each user $u \in 1, 2, \ldots, N_u$. Since $\mathbf{X}$ has to be in the form of N_u diagonal $N \times N$ matrices stacked horizontally, the norm in (13) can be simplified as

$$\begin{aligned} \left\| \mathbf{Y_f} - \mathbf{X}\mathrm{diag}(\mathbf{F_L})\hat{\mathbf{H}}_\mathbf{k} \right\|_F^2 &= \|\mathbf{Y_f} - \mathbf{XB}\|_F^2, \\ &= \sum_{n=1}^{N} \left\| \mathbf{y_n^T} - \mathbf{x^T(n)}\mathbf{B_n} \right\|_2^2, \end{aligned} \tag{14}$$

where $\mathbf{y_n^T}$ is the n-th row of $\mathbf{Y_f}$, $\mathbf{x^T(n)} = [x(n, 1)x(n, 2) \cdots x(n, N_u)]$, where $x(n, u)$ is the n-th diagonal element of $\mathbf{X(u)}$, and $\mathbf{B_n}$ is the $N_u \times N_r$ matrix obtained by stacking the n-th row of $\mathbf{B(u)}$, for $u = 1, 2, \ldots, N_u$, vertically. Here, $\mathbf{y_n^T}$ is the received signal at the N_r antennas for the n-th subcarrier, $\mathbf{x^T(n)}$ represents the estimated transmitted signal on the n-th subcarrier for all the users, and $\mathbf{B_n}$ is the estimated frequency domain channel matrix for the n-th subcarrier for all the users. Thus, for each subcarrier $n \in 1, 2, \ldots, N$, this leads to the linear least squares solution

$$\mathbf{x_k^T(n)} = \mathbf{y_n^T}(\mathbf{B_n})^\dagger. \tag{15}$$

In each iteration, the minimization problems in (11) and (13) are solved alternately using the expressions in (12) and (15), respectively, until the optimal solution $\{\hat{\mathbf{H}}, \hat{\mathbf{X}}\}$ to faithfully reconstruct $\mathbf{Y_f}$ within a pre-defined error tolerance ϵ is obtained. Then the NN_u samples corresponding to $\hat{\mathbf{X}}$ and the LN_rN_u samples corresponding to $\hat{\mathbf{H}}$ are sent from the RRH to the BBU, leading to the compression ratio for multi-user system ($\mathrm{CR_{MU}}$) as

$$\mathrm{CR_{MU}} = \frac{NN_r}{N_u(N + LN_r)}. \tag{16}$$

C. MULTI-USER CASE - NON-OVERLAPPING ALLOCATION

The signal model for the scenario when multiple users are allocated subsets of the N subcarriers in one OFDM symbol in a non-overlapping manner can be expressed as a special case of the model for overlapping allocation. In (9) for $\mathbf{Y}$ in the frequency domain,

$$\mathbf{Y_f} = [\mathbf{X_f(1)X_f(2)}, \ldots \mathbf{X_f(N_u)}]\mathrm{diag}(\mathbf{F_L})\begin{bmatrix} \mathbf{H_t(1)} \\ \mathbf{H_t(2)} \\ \vdots \\ \mathbf{H_t(N_u)} \end{bmatrix} + \mathbf{W_f},$$

each user data matrix $\mathbf{X_f(u)}$, for $u = 1, 2, \ldots, N_u$, will have non-zeros diagonal elements only on the subcarriers allocated to that user, and zeros elsewhere. With this signal model, for the alternating minimization algorithm, the solution for the optimization problem in (11) remains the same as

$$\mathbf{\hat{H}_k} = \left(\mathbf{\hat{X}_{k-1}}\mathrm{diag}(\mathbf{F_L})\right)^{\dagger}\mathbf{Y_f}.$$

However, since only one user is scheduled in each subcarrier, the expression for the norm in (13) is modified to

$$\begin{aligned} \left\|\mathbf{Y_f} - \mathbf{X}\mathrm{diag}(\mathbf{F_L})\mathbf{\hat{H}_k}\right\|_F^2 &= \|\mathbf{Y_f} - \mathbf{XB}\|_F^2, \\ &= \sum_{n=1}^{N}\left\|\mathbf{y_n^T} - x(n)\mathbf{b_n^T}\right\|_2^2, \end{aligned} \tag{17}$$

where $\mathbf{b_n^T}$ is the n-th row of the estimated frequency domain channel matrix $\mathbf{B(u)}$ of the user allocated the n-th subcarrier. This leads to the same solution as in the single user case in (8):

$$x_k(n) = \frac{\left(\sum_{r=1}^{N_r} y(n,r)b^*(n,r)\right)}{\left(\sum_{r=1}^{N_r} |b(n,r)|^2\right)},$$

where $x_k(n)$ is the estimated data symbol of the user allocated the n-th subcarrier. The alternating minimization algorithm solves (11) and (13) using (12) and (8), respectively until $\{\mathbf{\hat{H}}, \mathbf{\hat{X}}\}$ meeting the error tolerance is found. Since there are only N non-zero elements in $\mathbf{\hat{X}}$, these N samples and the LN_rN_u samples corresponding to $\mathbf{\hat{H}}$ are transmitted on the fronthaul, resulting in a compression ratio of

$$\mathrm{CR_{MU}}(\text{non-overlapping}) = \frac{NN_r}{N + LN_rN_u}. \tag{18}$$

Algorithm 2 Alternating Minimization for Fronthaul Compression for Multiple Users With Overlapping Subcarrier Allocation

```
Input $\mathbf{Y_f}$
Define error tolerance $\epsilon$
Initialize diagonals of $\mathbf{\hat{X}_0(u)}$ to be any $N_u$ columns of
    $\mathbf{Y_f}$ for $u = 1, 2, \ldots, N_u$, $k = 1$, $\mathbf{B} = [\ ]$
repeat
    $\mathbf{\hat{H}_k} \leftarrow (\mathbf{\hat{X}_{k-1}}\mathrm{diag}(\mathbf{F_L}))^{\dagger}\mathbf{Y_f}$
    $\mathbf{B} \leftarrow \mathrm{diag}(\mathbf{F_L})\mathbf{\hat{H}_k}$
    for $n \leftarrow 1$ to $N$ do
        $\mathbf{x_k^T(n)} = \mathbf{y_n^T}(\mathbf{B_n})^{\dagger}$
    end for
    $k \leftarrow k + 1$
until $||\mathbf{Y_f} - \mathbf{\hat{X}_k}\mathrm{diag}(\mathbf{F_L})\mathbf{\hat{H}_k}||_F/||\mathbf{Y_f}||_F < \epsilon$
return $\mathbf{\hat{X}_k}, \mathbf{\hat{H}_k}$
```

$\mathbf{x_k^T(n)} = [x(n,1)\ x(n,2)\ \ldots\ x(n,N_u)]$, where $x(n,u)$ is the n^{th} diagonal element of $\mathbf{\hat{X}_k(u)}$, $\mathbf{y_n^T}$ is n^{th} row of $\mathbf{Y_f}$, $\mathbf{B_n}$ is the $N_u \times N_r$ sub-matrix of $\mathbf{B}$ for n-th sub-carrier for the N_u users.

D. MULTI-ANTENNA USERS

Multiple transmit antennas can be used for either diversity to improve error rate or spatial multiplexing to improve data rate. Suppose a user in the uplink has N_t antennas. If the N_t antennas are used to transmit the same data (diversity), then the received signal matrix can be expressed as

$$\begin{aligned} \mathbf{Y_f} &= \mathbf{X_fF_LH_t^1} + \mathbf{X_fF_LH_t^2} + \ldots + \mathbf{X_fF_LH_t^{N_t}} + \mathbf{W_f}, \\ &= \mathbf{X_fF_L}\left(\mathbf{H_t^1} + \mathbf{H_t^2} + \ldots + \mathbf{H_t^{N_t}}\right) + \mathbf{W_f}, \end{aligned} \tag{19}$$

where the superscript denotes the user antenna number, i.e., $\mathbf{H_t^i}$ is the time-domain channel for transmit antenna i of the user. Letting $\mathbf{H_t^1} + \mathbf{H_t^2} + \cdots + \mathbf{H_t^{N_t}} = \mathbf{H_t}$ allows (19) to be of the same form as (2) for the single user with single antenna case. Therefore, Algorithm 1 can be used for compression and the compression ratio remains the same as in (3).

When the multiple antennas are employed by the user to transmit different data streams (spatial multiplexing), then

$$\mathbf{Y_f} = \left[\mathbf{X_f^1X_f^2}, \ldots \mathbf{X_f^{N_t}}\right]\mathrm{diag}(\mathbf{F_L})\begin{bmatrix} \mathbf{H_t^1} \\ \mathbf{H_t^2} \\ \vdots \\ \mathbf{H_t^{N_t}} \end{bmatrix} + \mathbf{W_f}, \tag{20}$$

where $\mathbf{X_f^i}$ is the $N \times N$ diagonal matrix of data symbols transmitted from antenna i of the user. This is similar to (10) for the multi-user case. Thus, the multiple antennas of a user can be treated as additional virtual users when they transmit different data streams. Therefore, the compression ratio for a single user employing N_t antennas for multiplexing is the same as given in (16), with N_t replacing N_u, and Algorithm 2 can be used for compression.

The above can be similarly extended to multiple users, each with N_t antennas, by replacing N_u by the total number of data streams, $N_s = N_u N_t$ in (10) and (16).

E. RECOVERY AT BBU

Upon receiving the samples corresponding to $\hat{\mathbf{X}}$ and $\hat{\mathbf{H}}$ at the BBU, the matrix $\mathbf{Y_f}$ is reconstructed as

$$\hat{\mathbf{Y}}_\mathbf{f} = \hat{\mathbf{X}}\mathbf{F_L}\hat{\mathbf{H}}, \tag{21}$$

for the single user case, and as

$$\hat{\mathbf{Y}}_\mathbf{f} = \hat{\mathbf{X}}\mathrm{diag}(\mathbf{F_L})\hat{\mathbf{H}}, \tag{22}$$

for the multi-user case.

Once $\mathbf{Y_f}$ is reconstructed, the standard pilot-based channel estimation can be used at the BBU for data recovery. The estimated channel coefficients are used to apply either MRC in the single user case or zero-forcing (ZF) equalization in the multi-user case to $\hat{\mathbf{Y}}_\mathbf{f}$ in order to recover the user data $\mathbf{X_f}$.

The $\hat{\mathbf{X}}$ received from RRH are not directly used as the estimate of the user data $\mathbf{X_f}$ for the following two reasons:

1) The solution $\{\hat{\mathbf{X}}, \hat{\mathbf{H}}\}$ obtained in Algorithms 1 and 2 are unique only up to a scalar constant, and the product in (21) or (22) can converge to $\mathbf{Y_f}$ even when $\hat{\mathbf{X}}$ and $\hat{\mathbf{H}}$ do not individually converge to the actual $\mathbf{X_f}$ and $\mathbf{H_t}$. A detailed discussion on this is given in the Appendix.
2) In the case of network architectures where several MIMO RRHs share a single BBU pool, the BBU can have knowledge of interferers which allows it to get a better estimate of the channel than the $\hat{\mathbf{H}}$ received from the RRH.

IV. PERFORMANCE ANALYSIS UNDER DIFFERENT PROPAGATION CONDITIONS

In this section, the robustness of the proposed method is assessed under different channel conditions by analyzing how various channel parameters impact its error performance and compare it against that of an uncompressed system for reference. The analysis is confined to single user system since the proposed method has the same iterative minimization approach for both single user and multi-user systems. The multi-path channel can be modeled as a finite impulse response (FIR) filter characterized by the following three parameters:

- Number of channel taps
- Power of the channel taps
- Delay of the channel taps

Another factor that affects error performance is the correlation between the receive antennas. Monte Carlo simulations are used to evaluate the effect of the above parameters on the uncoded symbol error rates (SER) when the fronthaul data is compressed using the proposed method. The performance of the method under realistic channel models is also assessed in this section. The common simulation parameters are summarized in Table 1.

TABLE 1. Simulation parameters.

Modulation scheme (M)	64-QAM
No. of RRH antennas (N_r)	64
FFT size (N)	1024
No. of pilots (P)	64

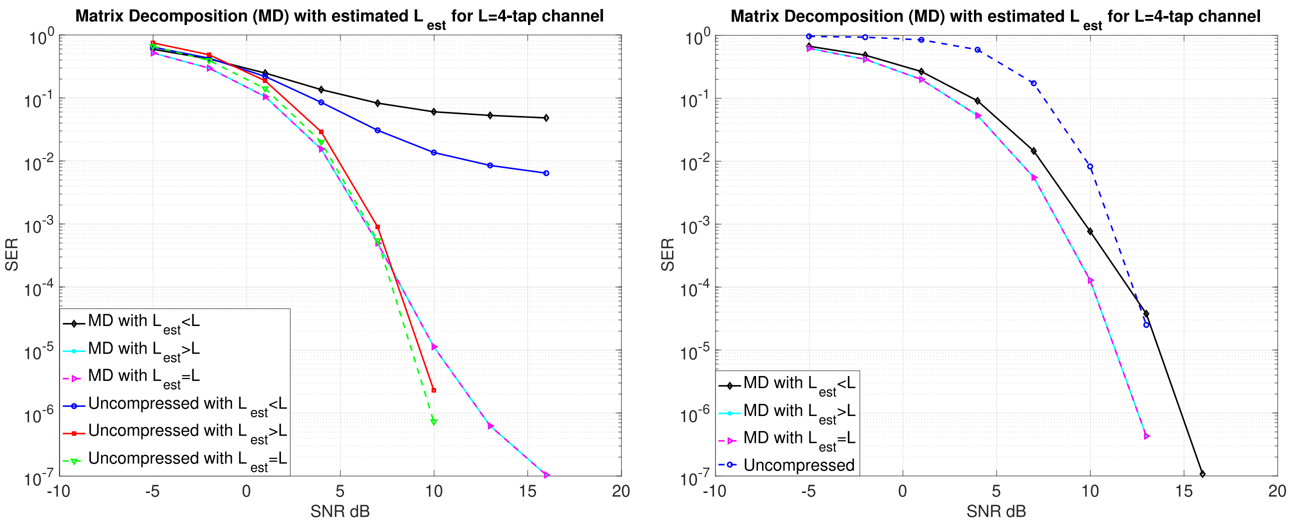
After reconstruction of $\mathbf{Y_f}$, two different interpolation methods are used for channel estimation at the BBU: (a) Linear interpolation, where the frequency-domain channel estimates obtained at the pilot subcarriers are linearly interpolated to the rest of the subcarriers, and (b) FFT interpolation, where the pilot-based frequency-domain channel estimates are first converted to the time-domain and then interpolated to all the subcarriers using an FFT operation.

A. NUMBER OF TAPS

The number of taps determines the dimension, L of the multi-path channel matrix, $\mathbf{H_t}$ in the decomposition $\mathbf{Y_f} = \mathbf{X_f}\mathbf{F_L}\mathbf{H_t} + \mathbf{W_f}$. In the 3rd Generation Partnership Project (3GPP) 5G standards, the Physical Random Access Channel (PRACH) can estimate the number of taps in the channel. This estimate is used to fix the dimension of $\hat{\mathbf{H}}$ produced by the proposed method, which affects the achieved compression ratio (CR), i.e., greater this dimension, lower the achieved CR as per (3), (16), and (18). The scenario of interest is when this estimate is not exact.

The error performance of the proposed method is examined when the estimate of the number of taps in the channel, L_{est} available at the RRH is not equal to its true value, L. This affects the dimension of $\hat{\mathbf{H}}$ output by the proposed fronthaul compression algorithm at the RRH. It can also affect the channel estimation at the BBU when the pilot-based estimates are converted to the time-domain in the FFT interpolation method.

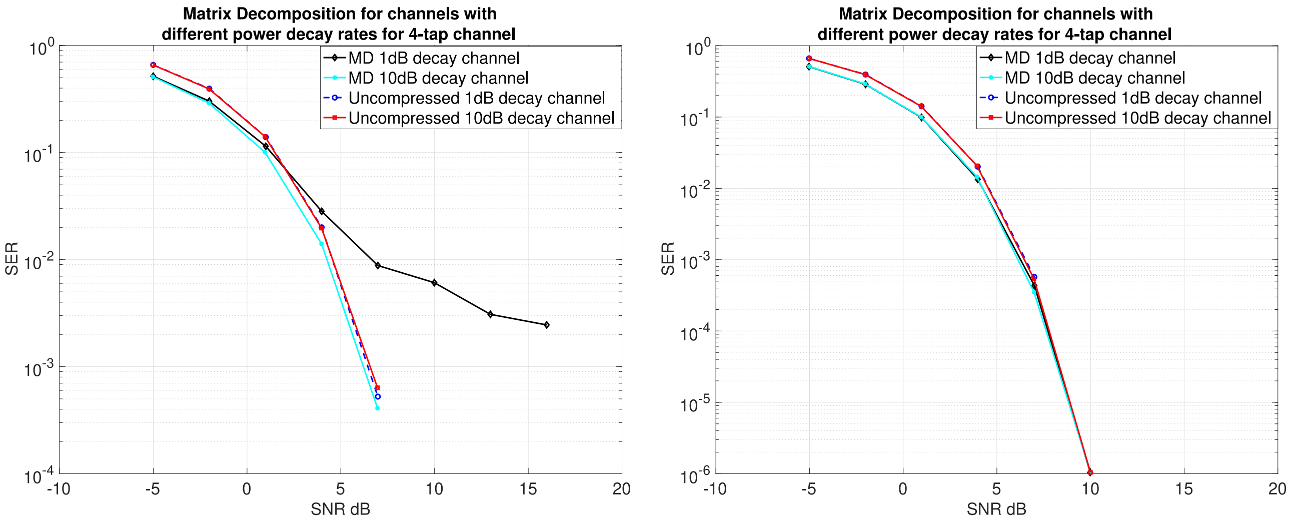
Two cases are considered here: (i) when $L_{est} < L$ and (ii) when $L_{est} \geq L$. Fig. 4(a) compares the uncoded SER of the proposed matrix decomposition method for these two cases against the uncompressed system when FFT interpolation is used at the BBU for pilot-based channel estimation. It is observed that when $L_{est} \geq L$, the proposed method performs slightly better than the uncompressed system, whereas when $L_{est} < L$, it performs worse. Fig. 4(b) shows the SER for both the proposed method and uncompressed system with linear interpolation at the BBU for channel estimation after reconstruction of $\mathbf{Y_f}$. There is only one curve for the uncompressed system as it is not affected by the error in L_{est}. Here, it is observed that matrix decomposition has nearly 4 dB gain over the uncompressed system. The uncompressed system suffers 4 dB loss in performance with linear interpolated channel estimates versus the FFT interpolated estimates in Fig. 4(a), whereas matrix decomposition suffers only 2 dB loss. It can also be concluded from both these figures that as long as $L_{est} \geq L$, matrix decomposition offers a small denoising gain similar to that observed in other low rank approximation methods such as PCA. However,



(a)

(b)

FIGURE 4. Uncoded SERs of the proposed method (after 10 iterations of Algorithm 1) and the uncompressed system when $L_{est} \neq L$ for (a) FFT interpolation, and (b) linear interpolation used at the BBU for channel estimation. SER for $L_{est} = L$ plotted for reference. In case (a), the proposed method offers a small denoising gain over the uncompressed system if $L_{est} \geq L$ at low signal-to-noise ratios (SNRs). In case (b), it offers good denoising gain for $L_{est} \geq L$, but when $L_{est} < L$, this denoising effect is lost at higher SNRs.



(a)

(b)

FIGURE 5. Uncoded SERs of the proposed method and the uncompressed system with 1 dB and 10 dB decay rates in channel tap powers for (a) 10 iterations of Algorithm 1, and (b) 50 iterations of Algorithm 1. FFT interpolation used at BBU for channel estimation. The proposed method performs well if the channel taps have large differences in power, but takes more number of iterations to provide comparable performance when the tap powers are similar.

when $L_{est} < L$, some of the diversity provided by multi-path propagation is lost.

B. POWER DELAY PROFILE

The effect of the power of the channel taps on the proposed method is examined next. The tap powers of the channel filter are modeled as a decaying exponential and the error performance for different decay rates are studied. Fig. 5(a) shows the uncoded SER for 10 iterations of Algorithm 1 for two decay rates, 1 dB/tap and 10 dB/tap. It is observed that the method performs well if the difference in tap powers is large; it performs poorly at higher SNR if the tap powers are too close together.

Fig. 5(b) shows the SER for 50 iterations of the algorithm and it can be observed that the 1 dB decay curve now matches the 10 dB decay curve, and both provide a small denoising gain over the uncompressed system. This implies that it is harder for the algorithm to distinguish the different multi-paths for the estimate $\hat{\mathbf{H}}$ when their powers are similar; however, given enough time, the algorithm manages to identify them correctly.

C. TAP DELAYS

The matrix $\mathbf{F_L}$ in the signal model $\mathbf{Y_f} = \mathbf{X_f F_L H_t} + \mathbf{W_f}$ denotes the L columns of the $N \times N$ DFT matrix

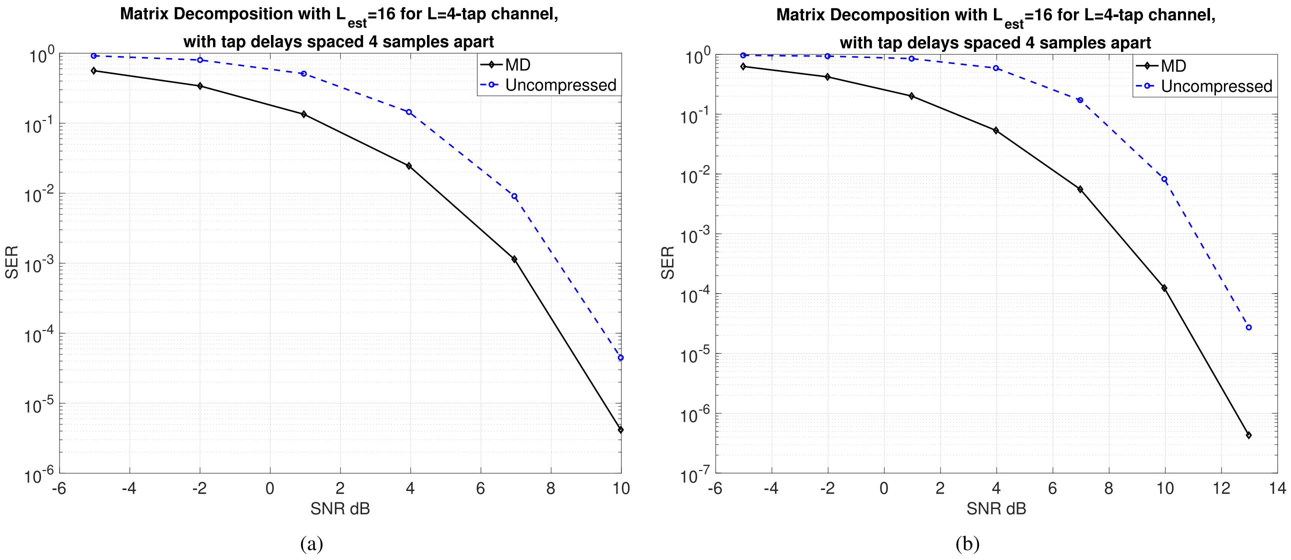


FIGURE 6. Uncoded SER of the proposed method (10 iterations) for a 4-tap channel ($L = 4$), all taps spaced four samples apart and $\mathbf{F}_{\mathbf{L}_{\mathbf{est}}}$ with $L_{est} = 16$ used instead of $\mathbf{F}_{\mathbf{L}}$ in Algorithm 1, with (a) FFT interpolation, and (b) linear interpolation used at BBU for channel estimation. In both cases, the proposed method provides denoising gain compared to the uncompressed system.

corresponding to the channel tap delays in units of sampling duration. This should match the $\mathbf{F_L}$ used in (4) of the proposed alternating minimization algorithm for the best results. However, in practice, the tap delays are not known at the receiver. It was noted in Section IV-A that as long as $L_{est} > L$, the method performs well. Therefore, choose a value of L_{est} that is large enough to correspond to the channel tap with the largest delay and use $\mathbf{F}_{\mathbf{L}_{\mathbf{est}}}$, the first L_{est} columns of the $N \times N$ DFT matrix in place of $\mathbf{F_L}$ in (4). This changes the compression ratio (CR) in (3) to

$$\mathrm{CR_{SU}} = \frac{NN_r}{N + L_{est}N_r}. \tag{23}$$

This would work well in practice since multi-paths which are extremely delayed have negligible power and can be ignored while choosing the value of L_{est}. This circumvents the problem of lack of information on the tap delays at the cost of only slightly lower CR.

Fig. 6 shows the SER performance of the proposed method and the uncompressed system for a channel with four taps and a delay of four samples between each tap. $\mathbf{F}_{\mathbf{L}_{\mathbf{est}}}$ with $L_{est} = 16$ (corresponding to the delay of the last tap) is used in place of $\mathbf{F_L}$ in Algorithm 1. The proposed method performs better than the uncompressed system while still providing a CR of 32.

It can be concluded from this section that the proposed method can provide good error performance even when none of the channel parameters are known, as long some of the compression ratio is sacrificed to fix a large enough value of L_{est}. This method can also be used to perform an almost blind demodulation of MIMO OFDM symbols by using just a single pilot to estimate the scalar λ in Lemma 1 in the Appendix.

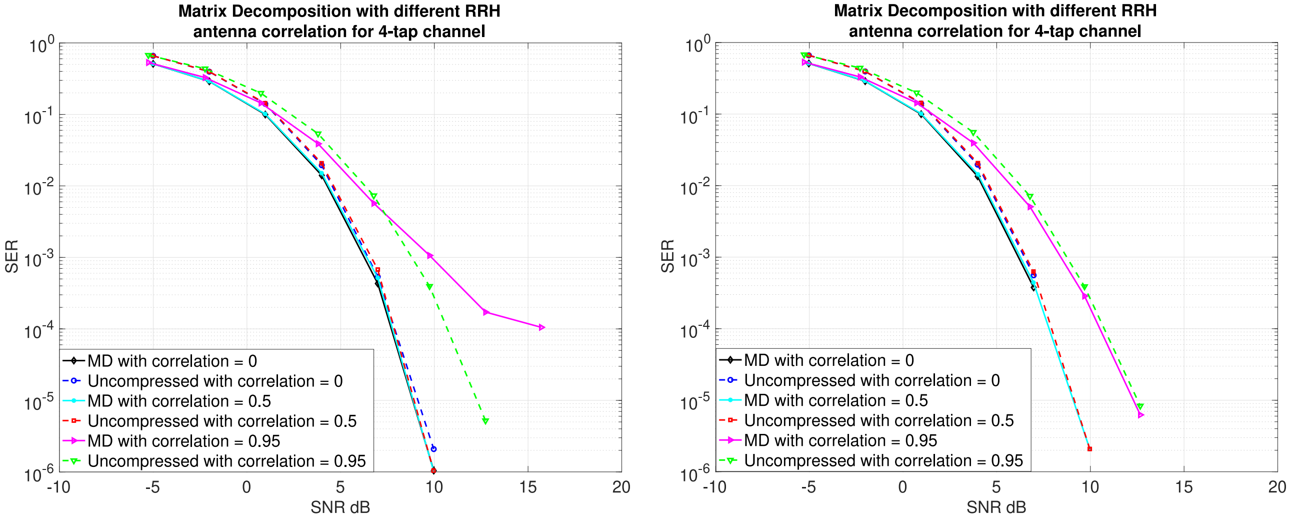
D. IMPACT OF RECEIVE ANTENNA CORRELATION

Massive MIMO antenna arrays have antenna elements spaced close together, making correlation between the antenna elements unavoidable in practical channels. Spatial correlation decreases the MIMO gain and capacity. The most widely used antenna correlation model is the exponential correlation model [24], which is also adopted by the 3GPP standards for 5G NR [25]. This model makes physical sense in most scenarios as correlation between antenna elements varies inversely with the distance between them. In this section, the impact of spatial correlation between the RRH antenna elements on the SER of the proposed method is evaluated. A uniform linear array (ULA) is assumed at the RRH with correlation matrix $\mathbf{R}$ given by

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & \rho^2 & . & . & \rho^{N_r-1} \\ \rho & 1 & \rho & . & . & \rho^{N_r-2} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \rho^{N_r-1} & \rho^{N_r-2} & . & . & . & 1 \end{bmatrix}_{N_r \times N_r},$$

where ρ is the correlation coefficient with $0 \le \rho \le 1$.

Fig. 7 shows the SER for three different levels of correlation, $\rho = 0$ or no correlation, $\rho = 0.5$ or moderate correlation, and $\rho = 0.95$, very high correlation, for the proposed method and the uncompressed system. Fig. 7(a) shows the SER for 10 iterations of Algorithm 1. It is observed that the proposed matrix decomposition method matches the uncompressed system performance when the correlation is not high, but performs worse at very high correlation. This is because when the antennas are highly correlated, the algorithm struggles to distinguish between the different antennas,



(a)

(b)

FIGURE 7. Uncoded SERs of the proposed method and the uncompressed system for different correlation coefficients between the antennas for (a) 10 iterations of Algorithm 1, and (b) 50 iterations of Algorithm 1. FFT interpolation used at BBU for channel estimation. The proposed method performs well when the correlation is not high, but takes more number of iterations to provide comparable performance when the correlation is extremely high.

similar to the trend observed in Section IV-B for tap powers. However, when the algorithm is given sufficient time (iterations), it again manages to distinguish between the antennas. This is demonstrated in Fig. 7(b) where the SER of matrix decomposition for very high correlation matches that of the uncompressed curve when the number of iterations is increased to 50.

E. PERFORMANCE IN PRACTICAL CHANNELS

In this section, the performance of the proposed matrix decomposition method in realistic multi-path fading propagation environments is examined. The 3GPP technical specifications for 5G NR provide channel models developed from real channel measurements for use in link level simulations [22], [26]. The tapped delay line (TDL) models for sub-6 GHz given therein are used here. These models are used to test the performance of the proposed compression method in low, medium and high delay spread environments. Linear interpolation is used at the BBU for channel estimation in the simulation of all three cases.

Fig. 8 shows the SER performance of the proposed method and the uncompressed system for the NR channel model TDLA30, which represents a low delay spread environment. SERs for two values of $L_{est} = 2$ and 6 are plotted, which provide compression ratios (CRs) of 57 and 51, respectively. It can be observed that the method performs well and provides about 3 dB denoising gain compared to the uncompressed system when the delay spread is low. In medium and high delay spread channels, the channel has more number of taps due to the richer multi-path propagation conditions. Therefore, L_{est} has to be increased to obtain comparable SERs. This is demonstrated in Fig. 9 and Fig. 10. Fig. 9 shows the SER for the medium delay spread channel model, TDLB100. With $L_{est} = 6$ and 12, matrix decomposition offers CRs of 51 and 36.6, respectively. The denoising gain is reduced for the lower value of L_{est} at higher SNR. In Fig. 10, $L_{est} = 10$ and 20 provide CRs of 39.4 and 28.4, respectively, for the high delay spread channel model, TDLC300. It can be observed that the denoising gain is further reduced at higher SNRs despite the increase in L_{est}.

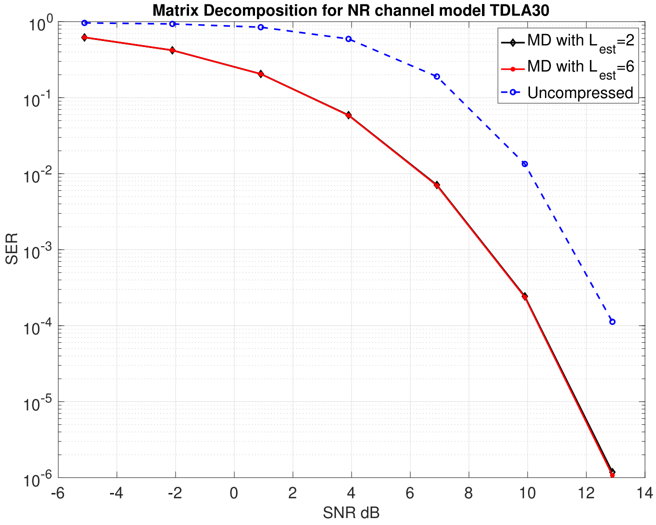


FIGURE 8. Uncoded SER of the proposed method (10 iterations of Algorithm 1) compared against the uncompressed system in low delay spread channel model TDLA30.

The algorithm is also tested in E-UTRA channel models, Extended Pedestrian A model (EPA) and Extended Vehicular A model (EVA) given in [22]. The values of L_{est} are chosen to be 5 and 30, giving compression ratios of 48.8 and 22.3 for a single user for the EPA and EVA models, respectively. Fig. 11 shows that the method provides very good denoising gain for both models, especially at low SNR, but the

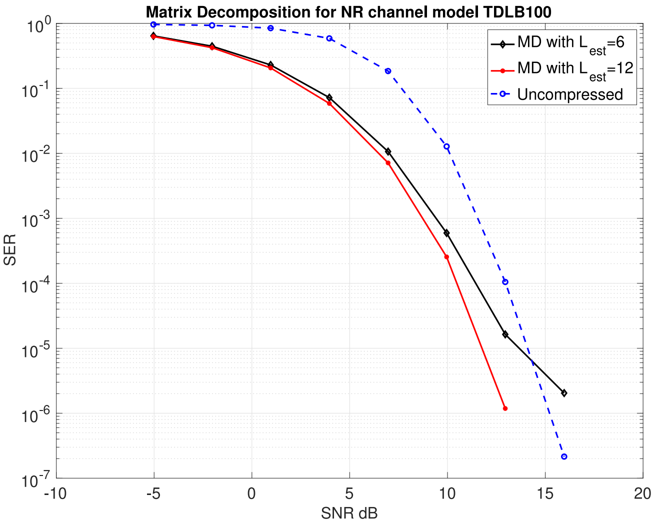


FIGURE 9. Uncoded SER of the proposed method (10 iterations of Algorithm 1) compared against the uncompressed system in medium delay spread channel model TDLB100.

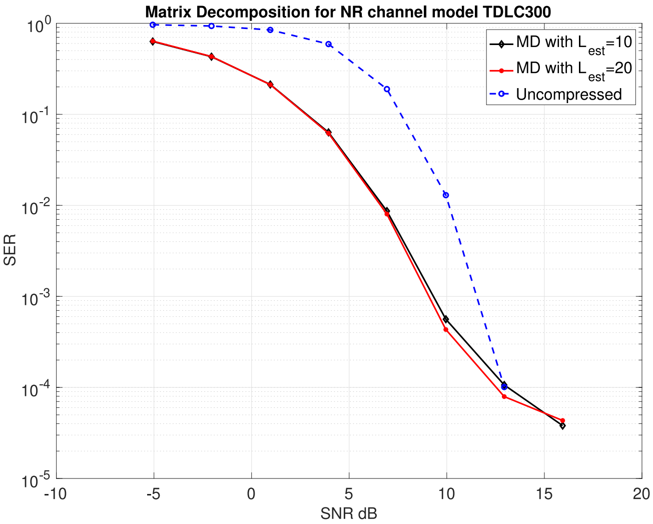


FIGURE 10. Uncoded SER of the proposed method (10 iterations of Algorithm 1) compared against the uncompressed system in high delay spread channel model TDLC300.

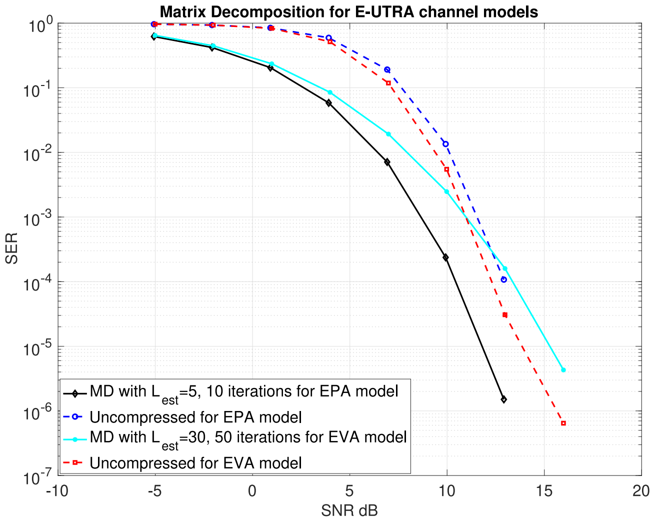


FIGURE 11. Uncoded SER of the proposed method compared against the uncompressed system for E-UTRA channel models, EPA (10 iterations of Algorithm 1) and EVA (50 iterations).

performance under EVA model deteriorates at high SNR even after 50 iterations of Algorithm 1, because its delay spread is larger, in line with the observations from Section IV-A.

V. ALGORITHM CONVERGENCE AND COMPLEXITY

The numerical convergence of the algorithm is evaluated for the different channel models discussed above. The normalized error between the received data matrix, $\mathbf{Y_f}$ and the matrix reconstructed at the BBU, $\hat{\mathbf{Y}}_\mathbf{f}$ is plotted against the iteration number at 0 dB SNR in Fig. 12(a) and at 20 dB SNR in Fig. 12(b). It can be observed that the algorithm converges fast (in 2 iterations) for the low delay spread channel, TDLA30. It takes about 7 iterations for TDLB100 model and converges the slowest (20 iterations) for the high delay spread model, TDLC300. For reference, the noise variance is equal to one for the 0 dB case and the proposed method converges to a value of error lower than this, demonstrating the denoising gain provided by the algorithm as observed in the previous section.

The above results suggest that the proposed method performs well in low/medium delay spread environments. However, for high delay spread channels, L_{est} needs to be very high and more number of iterations are required to achieve good performance. When L_{est} is high, even though it offers higher compression ratios than the existing methods [9], [13], algorithm complexity considerations come into play, which is discussed in detail next.

The computational complexity of the proposed compression method can be analyzed as a function of the dimensions N, L and N_r. It can be observed from Algorithm 1 that the most computationally intensive operation is step 5, to calculate

$$\hat{\mathbf{H}}_\mathbf{k} = \left(\hat{\mathbf{X}}_{\mathbf{k-1}}\mathbf{F_L}\right)^{\dagger}\mathbf{Y_f}. \tag{24}$$

The matrix pseudo-inverse is usually calculated via singular value decomposition (SVD), which has a time complexity of the order of LN^2 computations, i.e., $\mathcal{O}(LN^2)$, since $\hat{\mathbf{X}}_\mathbf{k}\mathbf{F_L}$ has dimension $N \times L$ and $N > L$. However, considering the fact that typically $N \gg L$ and noting that the matrix $\hat{\mathbf{X}}_\mathbf{k}$ is diagonal, expanding (24) as follows

$$\left(\hat{\mathbf{X}}_{\mathbf{k-1}}\mathbf{F_L}\right)^{\dagger}\mathbf{Y_f} = \left(\mathbf{F_L^H}\hat{\mathbf{X}}_\mathbf{k}^\mathbf{H}\hat{\mathbf{X}}_\mathbf{k}\mathbf{F_L}\right)^{-1}\mathbf{F_L^H}\hat{\mathbf{X}}_\mathbf{k}^\mathbf{H}\mathbf{Y_f} \tag{25}$$

reveals that replacing the matrix pseudo-inverse calculation via SVD with the expanded expression in (25) can lead to significant savings in computation. This arises from the fact that $\mathbf{F_L^H}\hat{\mathbf{X}}_\mathbf{k}^\mathbf{H}\hat{\mathbf{X}}_\mathbf{k}\mathbf{F_L}$ is only an $L \times L$ matrix, whose inverse can be calculated with significantly less number of computations than the SVD of an $N \times L$ matrix, when $N \gg L$.

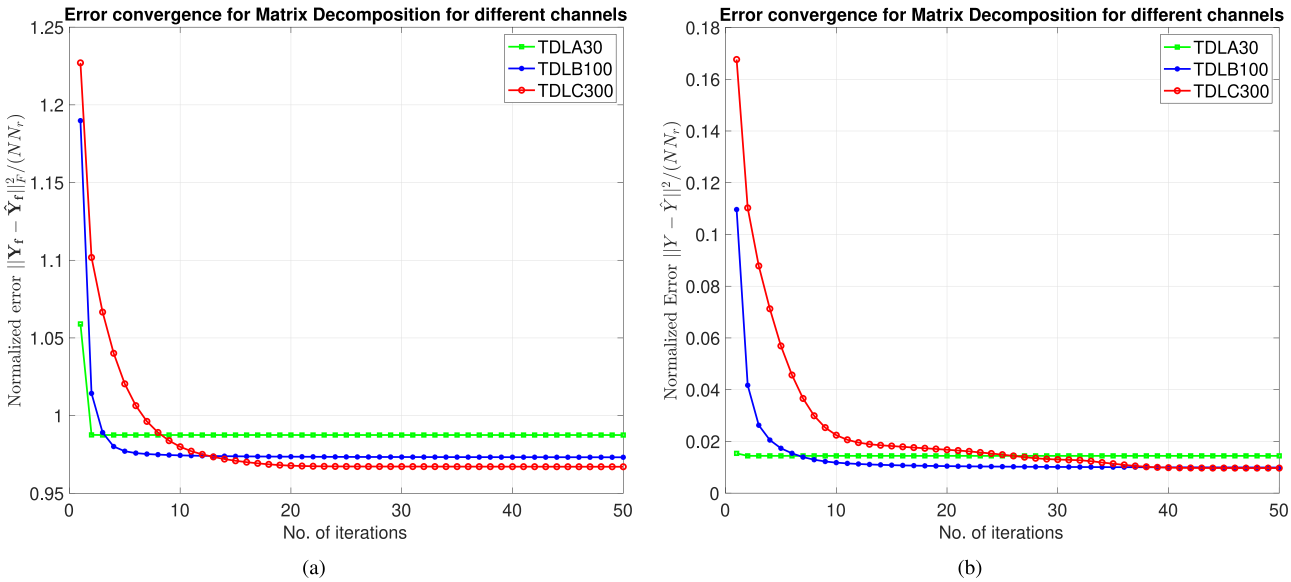


FIGURE 12. Convergence of normalized error between the received data matrix and reconstructed matrix from the proposed method at (a) 0 dB SNR, (b) 20dB SNR.

Thus, (25) can be split into the following steps:

1) Compute and store the matrix $\mathbf{F_L^H\hat{X}_k^H}$, which requires LN multiplications, same as a matrix-vector product since $\mathbf{\hat{X}_k^H}$ is diagonal.
2) Compute the product $(\mathbf{F_L^H\hat{X}_k^H})(\mathbf{\hat{X}_kF_L})$, which requires L^2N multiplications. Denote this matrix as $\mathbf{U}$.
3) Compute the inverse of the $L \times L$ matrix $\mathbf{U}$ which requires $\mathcal{O}(L^3)$ operations.
4) Compute the product $(\mathbf{F_L^H\hat{X}_k^H})\mathbf{Y_f}$ using the $\mathbf{F_L^H\hat{X}_k^H}$ stored in step 1. Denote this as the matrix $\mathbf{V}$. This requires LNN_r multiplications since $\mathbf{Y_f}$ is an $N \times N_r$ matrix.
5) Compute the product of $\mathbf{U}_{L\times L}^{-1}$ and $\mathbf{V}_{L\times N_r}$, which requires L^2N_r multiplications.

Considering that $N \gg N_r \gg L$, the approximate complexity of one iteration in Algorithm 1 is $\mathcal{O}(LNN_r)$, which is comparable to the complexity of $\mathcal{O}(N_r^3)$ of the SVD step in PCA compression [13]. However, multiple iterations are required for Algorithm 1 to converge to $\mathbf{Y_f}$, which makes it computationally more intensive than PCA compression in [13].

The break down of (25) into steps 1-5 above shows that the algorithm is only memory-intensive but not processor-intensive. This is due to the fact that all the steps except Step 4 involve only multiplications and additions. Step 4 calculates the inverse of an $L\times L$ matrix; however, since the size L is small (typically $L < 10$) because of channel hardening in the massive MIMO setting [15], it can even be implemented on a field programmable gate array (FPGA) [27]. In contrast, the PCA compression in [13] needs to compute the SVD of a large $N\times N_r$ matrix, which involves the calculation of square roots, a much more processor-intensive operation compared to multiplications and additions.

VI. PERFORMANCE EVALUATION AGAINST OTHER COMPRESSION METHODS

The QR decomposition-based compression method [9] and the PCA-based compression method [13] offer the best compression ratios and performance in existing literature. Therefore, these methods are used to benchmark the performance of the proposed compression method based on two criteria: Compression Ratio (CR) and Symbol Error Rate (SER).

A. COMPARISON OF ACHIEVED COMPRESSION RATIOS

For a single user in the system, the CR of the proposed method is given by (3). The CR for PCA/QR-based compression in [9], [13] is given by

$$\mathrm{CR}_{\mathrm{SU,PCA}} = \frac{NN_r}{L(N+N_r)}. \tag{26}$$

It can be observed from (3) and (26) that for large values of N, which is the case when N is the OFDM symbol length for large bandwidths, having $N \gg \max\{L, N_r\}$ implies

$$\frac{\mathrm{CR}_{\mathrm{SU}}}{\mathrm{CR}_{\mathrm{SU,PCA}}} = \frac{LN + LN_r}{N + LN_r} \approx \frac{LN}{N},$$

which gives

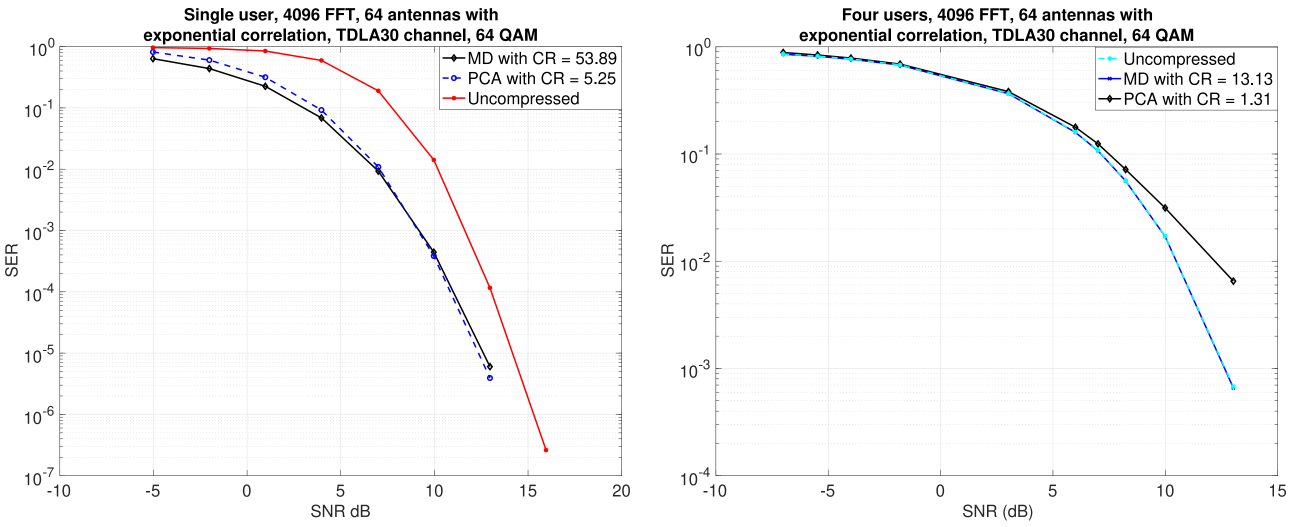
$$\mathrm{CR}_{\mathrm{SU}} \approx L \times \mathrm{CR}_{\mathrm{SU,PCA}}. \tag{27}$$

For the case when multiple users share the same set of N subcarriers (overlapping allocation),

$$\mathrm{CR}_{\mathrm{MU,PCA}} = \frac{NN_r}{LN_u(N+N_r)}. \tag{28}$$

Comparing this with (16) when $N \gg \max\{L, N_r, N_u\}$ gives

$$\frac{\mathrm{CR}_{\mathrm{MU}}}{\mathrm{CR}_{\mathrm{MU,PCA}}} = \frac{LN_uN + LN_uN_r}{N_uN + LN_uN_r} \approx L.$$



(a)

(b)

FIGURE 13. Uncoded SERs of the proposed method (after 10 iterations of alternating minimization), PCA compression and uncompressed system for (a) single user in the system, and (b) four users sharing the same N subcarriers in the system. For a single user, the proposed method achieves above 50× compression and matches the SER of PCA compression, and both provide denoising gain over the uncompressed system. For the multi-user case, the proposed method achieves above 10× compression and its SER matches that of the uncompressed system while the SER of PCA compression degrades at high SNR.

TABLE 2. Compression ratios for the proposed matrix decomposition (MD) and PCA methods for $N_r = 64$, $L = 12$, $N_u = 1$.

Method	$N = 1024$	$N = 4096$
MD	36.6	53.9
PCA	5.0	5.2

TABLE 3. Compression ratios for the proposed MD and PCA methods for $N_r = 64$, $L = 12$, $N_u = 4$ (multi-user MIMO).

Method	$N = 1024$	$N = 4096$
MD	9.2	13.5
PCA	1.2	1.3

TABLE 4. Simulation parameters.

Modulation scheme (M)	64-QAM
No. of RRH antennas (N_r)	64
FFT size (N)	4096
Multi-path Channel length (L)	12
Channel Model	TDLA30

The CRs for different values of N, N_r, L, N_u for both the methods are given in Tables 2 and 3. The TDLA30 channel with $L = 12$ taps [22] is used for the calculation of the CRs. Since $L \ll N$, the FFT size, the decomposition of $\mathbf{Y}$ into approximate factors $\hat{\mathbf{X}}$ of length N, and $\hat{\mathbf{H}}$ of size $L \times N_r$ using the proposed method results in a compression of 30-50×. It is noted that the proposed compression method can give nearly an order of magnitude higher CRs than PCA compression in [9], [13] as illustrated by equations (27) and (28). This means that for a 64 antenna, 100 MHz system, the proposed compression method reduces the fronthaul data rate required from 236 Gbps to approximately 5 Gbps for a single user system, and 20 Gbps for four users in a multi-user MIMO system.

B. SYMBOL ERROR RATE PERFORMANCE

Monte Carlo simulations are used to evaluate the uncoded SER of the proposed method, PCA compression and the uncompressed system. The simulation parameters are summarized in Table 4. The exponential correlation model [24] with correlation coefficient 0.7 is used to model the antenna correlation for a uniform linear array at the RRH. For the multi-user case, $N_u = 4$ users are assumed to share the same set of N subcarriers (overlapping allocation).

Fig. 13 compares the uncoded SERs of the proposed method and PCA compression for the TDLA30 channel model. SERs for the uncompressed system are also plotted as baseline. After quantizing the channel tap delays for the resolution given by 4096-point FFT and 30 kHz subcarrier spacing, the value of $L_{est} = 12$ was chosen. For a single user system, this gives compression ratios of 53.9 and 5.25, for the proposed matrix decomposition method and PCA method, respectively. Fig. 13(a) shows the SER for a single user in the system. It is observed that the proposed method performs as well as PCA while providing more than 10× its compression. Both the proposed method and PCA also provide a denoising gain of approximately 2dB over the uncompressed system. Fig. 13(b) shows the SERs for $N_u = 4$ users in the system (multi-user MIMO) with overlapping subcarrier allocation. The proposed method provides a compression ratio of 13.13 and matches the SER of the uncompressed system while the SER of PCA degrades at high SNR even though it provides only a tenth of the compression provided by the proposed method.

VII. CONCLUSION

The capacity of the fronthaul optical link is a major bottleneck in the implementation of OFDM massive MIMO networks with split architecture as specified by O-RAN. In this work, a method of fronthaul compression for the uplink MIMO that exploits the convolution structure of the data received at the RRH was proposed. An iterative alternating minimization approach was used at the RRH to approximate the received signal as the product of a diagonal user data matrix and a low rank channel response matrix, allowing the received signal to be reconstructed at the BBU using fewer samples. The performance of this method was analyzed for different channel parameters and its robustness was tested in many simulated realistic channel models. The method can be tailored to both single-user and multi-user MIMO systems, and link level simulations show that it provides the same or even better symbol error rates compared to an uncompressed system. It can offer nearly an order of magnitude higher compression ratios than the existing compression methods. The proposed method can also be used for almost blind demodulation of MIMO OFDM symbols.

APPENDIX

UNIQUENESS OF THE SOLUTION

The optimal solution to (4) and (6), $\{\hat{\mathbf{H}}, \hat{\mathbf{X}}\}$, respectively, obtained via Algorithm 1 is unique up to a scalar constant. This is proved in the following lemma.

Lemma 1: If $\{\hat{\mathbf{H}}, \hat{\mathbf{X}}\}$ and $\{\tilde{\mathbf{H}}, \tilde{\mathbf{X}}\}$ are two solutions to (4) and (6), then they are related to each other by the scalar transform

$$\tilde{\mathbf{X}} = \left(\frac{1}{\lambda}\right)\hat{\mathbf{X}} \text{ and } \tilde{\mathbf{H}} = \lambda\hat{\mathbf{H}}, \tag{29}$$

where $\lambda \in \mathbb{C}$, the set of complex numbers.

Proof: If $\{\hat{\mathbf{H}}, \hat{\mathbf{X}}\}$ and $\{\tilde{\mathbf{H}}, \tilde{\mathbf{X}}\}$ are both solutions to (4) and (6), then

$$\hat{\mathbf{X}}\mathbf{F_L}\hat{\mathbf{H}} = \tilde{\mathbf{X}}\mathbf{F_L}\tilde{\mathbf{H}}. \tag{30}$$

Let $\tilde{\mathbf{X}} = \hat{\mathbf{X}}\mathbf{X}^*$ and $\tilde{\mathbf{H}} = \mathbf{H}^*\hat{\mathbf{H}}$, where $\mathbf{X}^*$ is an $N \times N$ matrix and $\mathbf{H}^*$ is an $L \times L$ matrix. Then,

$$\mathbf{X}^*\mathbf{F_L}\mathbf{H}^* = \mathbf{F_L} \tag{31}$$

is needed to satisfy (30). If $\mathbf{X}^*$ is a rank-N diagonal matrix, then $\mathbf{F_L}\mathbf{H}^* = (\mathbf{X}^*)^{-\mathbf{1}}\mathbf{F_L}$. If $\mathbf{f_m}$ denotes the m^{th} row of $\mathbf{F_L}$ and λ_m denotes the m^{th} diagonal element of $(\mathbf{X}^*)^{-\mathbf{1}}$, then the above can be written as

$$\mathbf{f_m}\mathbf{H}^* = \lambda_m\mathbf{f_m}, m = 1, 2, \ldots, N. \tag{32}$$

Thus, the rows of $\mathbf{F_L}$ form the left eigen vectors of the matrix $\mathbf{H}^*$. $\mathbf{F_L}$ is a rank-L matrix, therefore $\mathbf{f_{L+1}}$ can be expressed as

$$\mathbf{f_{L+1}} = \sum_{i=1}^{L} a_i\mathbf{f_i}. \tag{33}$$

Combining (32) and (33), this leads to

$$\mathbf{f_{L+1}}\mathbf{H}^* = \sum_{i=1}^{L} a_i\lambda_i\mathbf{f_i} = \lambda_{L+1}\sum_{i=1}^{L} a_i\mathbf{f_i},$$

which can hold true only when all λ_i's are equal. This gives

$$\mathbf{X}^* = \left(\frac{1}{\lambda}\right)\mathbf{I_N} \text{ and } \mathbf{H}^* = \lambda\mathbf{I_L}, \tag{34}$$

where $\lambda_i = \lambda$, for $i = 1, 2, \ldots, N$ and $\mathbf{I_K}$ denotes identity matrix of dimension K.

Note that $\{\hat{\mathbf{H}}\mathbf{H}^*, \hat{\mathbf{X}}\mathbf{X}^*\}$ is also a solution to (4) and (6), respectively, where $\mathbf{H}^* = \lambda\mathbf{I_{N_r}}$. To see this, use $\mathbf{A} = \mathbf{F_L}\hat{\mathbf{H}}$ in (30). Then the condition to be satisfied, $\mathbf{X}^*\mathbf{A}\mathbf{H}^* = \mathbf{A}$ is of the form in (31) and the result follows. ■

REFERENCES

[1] P. Aswathylakshmi and R. K. Ganti, "Fronthaul compression for uplink massive MIMO using matrix decomposition," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 2524–2529.

[2] (Next G Alliance, Washington, DC, USA). *Next G Alliance Report: 6G Technologies*. (2022). [Online]. Available: https://www.nextgalliance.org/white_papers/6g-technologies/

[3] S. Ahmadi, *5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards*. London, U.K.: Academic, 2019.

[4] E. G. Larsson and L. Van der Perre, "Massive MIMO for 5G," *IEEE 5G Tech Focus*, vol. 1, no. 1, pp. 1–4, Mar. 2017.

[5] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 708–769, 1st Quart., 2018.

[6] "Release 15 description, technical specification group services and system aspects," 3GPP, Sophia Antipolis, France, Rep. 3GPP TR 21.915, Sep. 2019.

[7] *eCPRI Specification V1.0*, CPRI Consortium, Bengaluru, India, Aug. 2017.

[8] "O-RAN use cases and deployment scenarios," ORAN-Alliance, Herndon, VA, USA, White Paper, 2020.

[9] P. Aswathylakshmi and R. K. Ganti, "QR approximation for fronthaul compression in uplink massive MIMO," in *Proc. IEEE Globecom Workshops*, 2019, pp. 1–7.

[10] B. Guo, W. Cao, A. Tao, and D. Samardzija, "LTE/LTE-A signal compression on the CPRI interface," *Bell Labs Tech. J.*, vol. 18, no. 2, pp. 117–133, Sep. 2013.

[11] B. Drvenica and G. Luz, "Compression analysis of massive MIMO uplink," M.S. thesis, Dept. Signals Syst., Chalmers Univ. Technol., Gothenburg, Sweden, 2016.

[12] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, 3rd Quart., 2016.

[13] J. Choi, B. L. Evans, and A. Gatherer, "Space-time fronthaul compression of complex baseband uplink LTE signals," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–6.

[14] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[15] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[16] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[17] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Appl. Comput. Harmonic Anal.*, vol. 47, no. 3, pp. 893–934, 2019.

[18] L. Liu and R. Zhang, "Optimized uplink transmission in multi-antenna C-RAN with spatial compression and forward," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5083–5095, Oct. 2015.

[19] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for C-RANs over ergodic fading channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5022–5032, Jul. 2016.

[20] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Multihop backhaul compression for the uplink of cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3185–3199, May 2016.

[21] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[22] *V15.4.0, NR; User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone (Release 15)*, 3GPP Standard TS 38.101-1, Dec. 2018.

[23] A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, and X. Chen, *5G Physical Layer: Principles, Models and Technology Components*. London, U.K.: Academic, 2018.

[24] S. L. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369–371, Sep. 2001.

[25] "V16.1.0. 5G; study on channel model for frequencies from 0.5 to 100 GHz (release 16)," 3GPP, Sophia Antipolis, France, Rep. 3GPP TR 38.901, Nov. 2020.

[26] *3GPP, V16.4.0, 5G; NR; Base Station (BS) Radio Transmission and Reception (Release 16)*, 3GPP Standard TS 38.104, Jul. 2020.

[27] S. Chetan, J. Manikandan, V. Lekshmi, and S. Sudhakar, "Hardware implementation of floating point matrix inversion modules on FPGAs," in *Proc. 32nd Int. Conf. Microelectron. (ICM)*, 2020, pp. 1–4.

P. ASWATHYLAKSHMI (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the University of Calicut, India, in 2016. She is currently pursuing the Integrated M.S.-Ph.D. degree in electrical engineering with the Indian Institute of Technology Madras, Chennai. She has worked with the Wireless Networking and Communications Group, The University of Texas at Austin from April 2022 to June 2022. Her research interests include massive MIMO systems, signal processing, wireless communication, and 5G and beyond systems. She was the recipient of the ANSYS Doctoral Fellowship Award from 2020 to 2022.

RADHA KRISHNA GANTI (Member, IEEE) received the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology Madras, Chennai, India, and the master's degree in applied mathematics and the Ph.D. degree in electrical engineering from the University of Notre Dame in 2009. He is an Associate Professor with the Indian Institute of Technology Madras. He is a coauthor of the monograph, *Interference in Large Wireless Networks* (NOW Publishers, 2008). His doctoral work focused on the spatial analysis of interference networks using tools from stochastic geometry. He received the 2014 IEEE Stephen O. Rice Prize, the 2014 IEEE Leonard G. Abraham Prize, and the 2015 IEEE Communications Society Young Author Best Paper Award. He was also awarded the 2016–2017 Institute Research and Development Award (IRDA) by IIT Madras. In 2019, he was awarded the TSDSI Fellow for Technical Excellence in standardisation activities and contribution to LMLC use case in ITU. He was the Lead PI from IITM involved in the development of 5G base stations for the 5G Testbed Project funded by the Department of Telecommunications, Government of India.