2D Beam Domain Statistical CSI Estimation for Massive MIMO Uplink

An-An Lu, Member, IEEE, Yan Chen, and Xiqi Gao, Fellow, IEEE

Abstract—In this paper, we investigate the beam domain statistical channel state information (CSI) estimation for the two-dimensional (2D) beam-based statistical channel model (BSCM) in massive multi-input multi-output (MIMO) systems. The problem is to estimate the beam domain channel power matrices (BDCPMs) based on multiple received pilot signals. A received signal model showing the relation between the statistical properties of the received pilot signals and the BDCPMs is derived. On the basis of the received signal model, we formulate an optimization problem with the Kullback-Leibler (KL) divergence. By solving the optimization problem, a novel method to estimate the statistical CSI without the estimates of instantaneous CSI is proposed. We further reduce the complexity of the proposed method by utilizing the circulant structures of particular matrices in the algorithm. We also showed the generality of the proposed method by introducing another application, i.e., estimation of the angle domain channel power matrix. Simulation results show that the proposed method has good convergence and can obtain sparse BDCPMs. Compared with the regularized multiple measurement vector focal underdetermined system solver (RM-FOCUSS) algorithm, the proposed algorithm obtains overall more accurate statistical CSI with much lower complexity and brings significant performance gains when used in channel estimation.

Index Terms—Statistical channel state information (CSI), massive multi-input multi-output (MIMO), beam-based statistical channel model (BSCM), beam domain channel power matrices (BDCPMs), KL-divergence.

I. INTRODUCTION

MASSIVE multiple-input multiple-output (MIMO) [1], [2], [3], [4], [5], [6] has been one of the key enabling technologies of the fifth-generation (5G) wireless communications networks. It provides enormous capacity gains and achieves high energy efficiency by employing a large number of antennas at the base station (BS). In massive MIMO systems, multi-user MIMO (MU-MIMO) [7] transmissions on the same time and frequency resource are enhanced significantly. Furthermore, massive MIMO also brings many new applications and services [8], [9], [10]. For the antenna array equipped in the BS, the uniform planar array (UPA) is widely used in practical massive MIMO systems since it has a compact size. In this paper, we investigate the three-dimensional (3D) massive MIMO systems equipped with UPA.

For massive MIMO systems with UPA, the beam-based statistical channel model (BSCM) is first used in the literature for robust linear precoder design [11], [12], [13], [14], [15]. The BSCM is extended from the unitary-independent-unitary (UIU) model [16] or the jointly correlated channel model [17], [18], [19] with the eigen-matrices being replaced by the over-sampled discrete Fourier transform (DFT) matrices. Because the number of antennas at each column or each row in practical massive MIMO with UPA is usually limited, the BSCM is more accurate than the beam domain channel model based on the DFT-based beams [14]. By assuming the statistical channel state information (CSI) is known, a posterior BSCM can be established to characterize the imperfect CSI caused by channel aging and other factors. The robust linear precoders based on the posterior BSCM achieve significant performance gains compared with the widely used regularized zero-forcing (RZF) [20] and signal-to-leakage and noise ratio (SLNR) [21] precoders under imperfect CSI [12], [13], [14], [15]. Furthermore, they also outperform the robust precoders based on the posterior beam domain channel model as shown in [14] and [15] because the BSCM is more accurate.

The model is then extended to the angle-delay domain as a two-dimensional (2D) BSCM [22], which is established by using over-sampled steering vectors in the space and frequency domain. The 2D BSCM can exploit the 2D channel sparsity more thoroughly, and thus greatly reduces the dimension of nonzero channel elements. With the 2D BSCM, significant performance gains can be achieved by the minimum mean square error (MMSE) channel estimation or the information geometry based channel estimation approach in [22]. However, the statistical CSI or the statistical parameters of the 2D BSCM need to be known before the instantaneous channel estimation. In massive MIMO systems, statistical CSI also has many applications other than channel estimation and robust precoding. For example, it is exploited to schedule user terminals (UTs) [23], [24], and is also used for transmitter eigen-beamforming and space-time block coding [25]. Thus, accurate statistical CSI plays an important role in improving the performance of massive MIMO systems. Although the statistical CSI of the

Manuscript received 20 May 2022; revised 2 February 2023 and 12 May 2023; accepted 22 May 2023. Date of publication 8 June 2023; date of current version 9 January 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1801103, in part by the National Natural Science Foundation of China under Grant 61801113 and Grant 61906206006, in part by the Jiangsu Province Basic Research Project under Grant BK201920602, in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022067, and in part by the Huawei Cooperation Project. The associate editor coordinating the review of this article and approving its publication was O. Tirkkonen.

(Corresponding author: Xiqi Gao.)

The authors are with the National Mobile Communications Research Laboratory (NCRL), Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 211111, China (e-mail: aalu@seu.edu.cn; yan.chen@seu.edu.cn; xqgao@seu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2023.3281841.

Digital Object Identifier 10.1109/TWC.2023.3281841

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
BSCM is very important, the problem of estimating them has not been addressed. Therefore, we consider the problem of estimating the statistical CSI for the 2D BSCM based on the received pilot signals in this paper.

In the literature, the statistical CSI is often obtained based on the estimated instantaneous CSI [19], or obtained through the expectation-maximization (EM) algorithms [26] which iteratively estimate the instantaneous and statistical CSI. There are also works [27] that obtain the covariance matrix directly without instantaneous CSI being involved. For the 2D BSCM, the problem becomes obtaining the beam domain channel power matrices (BDCPMs), which has not been addressed in the literature. In the 2D BSCM, the angle-delay domain or the beam domain channel coefficient is sparse due to the limited number of resolvable multi-paths. The considered problem can also be viewed as a multiple measurement vectors (MMV) problem [28], which is a classic compressive sensing problem [29]. The MMV focal under-determined system solver (M-FOCUSS) algorithm [28] can be applied to solve the problem and obtain the instantaneous CSI, which are then used to compute the statistical CSI. The M-FOCUSS algorithm is verified to have good performance at high signal-to-noise ratio (SNR), but poor performance at low SNR. Although the regularized M-FOCUSS (RM-FOCUSS) is developed for noisy scenarios and performs better than the M-FOCUSS algorithm, its performance is still not ideal under low SNR.

However, low SNR scenarios have to be considered in the practical massive MIMO systems. Besides, the computational complexity of the RM-FOCUSS method is not satisfactory due to the inversion of large dimensional matrices and estimating instantaneous CSI at each iteration. Furthermore, the statistical CSI of the 2D BSCM can also be utilized to improve the estimation performance of instantaneous CSI in practical massive MIMO systems [22]. Interference caused by pilot contamination was believed to create a finite capacity limit. In [30], this is proved to be a misunderstanding if the pilot-sharing UEs have asymptotically linearly independent covariance matrices and these matrices are known before channel estimation. Thus, it is better to obtain the statistical CSI for the considered problem before estimating the instantaneous CSI. In conclusion, we need a new method with lower complexity to estimate the statistical CSI for the 2D BSCM without estimating instantaneous CSI.

To achieve this goal, we first derive a theorem which gives the relation between the statistical properties of the considered channel matrices and the BDCPMs. Then, we derive a received signal model for the BDCPMs based on this relation. Based on the derived model, the statistical parameters of the 2D BSCM can be estimated directly without involving the instantaneous CSI. Furthermore, non-orthogonal pilots (NPs) are used due to the limited pilot resources in practical systems. To estimate the BDCPM, we then formulate a new optimization problem based on the Kullback-Leibler (KL) divergence. By solving the problem, we propose a novel method to obtain the BDCPM for the 2D BSCM. The proposed algorithm has a much lower complexity than the RM-FOCUSS method. Furthermore, we further reduce the complexity of the proposed method by utilizing the circulant structures of certain matrices in the algorithm. We also show the generality of the proposed method by presenting another application, i.e., estimation of the angle domain channel power matrix.

The main contributions of this paper are summarized as follows:

1) We derive a received signal model for the multi-user BDCPMs of the 2D BSCM. The received signal model can be used to estimate the statistical CSI of multiple users directly from the received pilot signals without estimating the instantaneous CSI.

2) We propose a novel method to obtain the 2D BDCPMs based on the received signal model and the KL divergence. Compared with the RM-FOCUSS method, the proposed method has much lower complexity and better performance in the low SNR regime.

3) We present a simplified method to obtain the angle domain channel power matrix for frequency flat fading channels.

The rest of this article is organized as follows. Section II introduces the system model and formulates the problem. Section III presents the estimation of the BDCPMs. Section IV provides simulation results. Section V draws the conclusion. Proofs of theorems and the corollary are provided in the Appendices.

A. Notations

Throughout this paper, uppercase and lowercase boldface letters are used for matrices and vectors, respectively. Superscripts (·)*, (·)T and (·)H denote the conjugate transpose, transpose and conjugate transpose operations, respectively. The operator \( \mathbb{E}\{\cdot\} \) denotes the mathematical expectation operator. In some cases, where it is not clear, we will employ subscript to emphasize the definition. Operators \( \circ \) and \( \otimes \) represent the Hadamard and Kronecker product, respectively. We use \( 0_{N,M} \) and \( 1_{N,M} \) to denote \( N \times M \) matrices or vectors of all zeros and all ones, respectively. The \( N \times N \) identity matrix is denoted by \( I_N \), and \( I_{N,M} \) is used to denote \( I_N \otimes 0_{M-N} \) when \( N < M \) and \( I_M \otimes 0_{N-M} \) when \( N > M \). The subscripts of 0, 0 and 1 can sometimes be omitted for convenience. We use \( [A]_{ij} \) to denote the \( (i,j) \)-th entry of the matrix \( A \). Operators \( \text{tr}(\cdot) \) and \( \text{det}(\cdot) \) represent the matrix trace and determinant, respectively. We use \( \text{diag}(X) \) to denote a column vector composed of the main diagonal elements of a square matrix \( X \), and \( \text{diag}(x) \) to denote the diagonal matrix with \( x \) along its main diagonal. A \( N \)-dimensional normalized DFT matrix is denoted as \( F_N \).

We further define the permutation matrix as

\[
\Pi_N^\infty = \begin{bmatrix}
0 & I_{(N-n)N}
\end{bmatrix},
\]

where \( \langle n \rangle_N \) denotes the modulo-\( N \)-operation.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a 3D massive MIMO system with frequency selective fading channels. The system consists of one BS equipped with a UPA and \( K \) UTs with single antennas. The
number of the antennas at the BS is $M_r$, where the numbers of antennas at each vertical column and horizontal row are $M_{r,z}$ and $M_{r,x}$, respectively. The orthogonal frequency division multiplexing (OFDM) [31] modulation is used to transform the frequency-selective fading channel into multiple parallel channels. Thus, the considered system is a massive MIMO-OFDM system. The number of subcarriers in the massive MIMO-OFDM system is $M_s$, and $M_p$ subcarriers are used for uplink pilot signal transmission. Lengths of the cyclic prefix (CP) and the sampling interval are denoted as $M_f$ and $T_s$.

We restrict our considerations to stationary channels and use the BSCM to describe the space-frequency correlations of each channel. We denote the polar and azimuthal angles of each channel. We denote the polar and azimuthal angles of arrival at the BS by $\theta_r, \phi_r$. Let $d_z$ and $d_x$ be the antenna spacing of each row and each column of the UPA. Let $\Delta_z = \frac{d_z}{\lambda}$, $\Delta_x = \frac{d_x}{\lambda}$ where $\lambda$ is the wavelength. The symbols $u_r$ and $v_r$ denote the directional cosines with respect to the $z$ axis and $x$ axis, respectively. Then, we have $u_r = \cos \theta_r$ and $v_r = \sin \theta_r \cos \phi_r$. The space steering vector at the BS side is given by

$$
\mathbf{a}_r(u_r, v_r) = \mathbf{v}_z(u_r) \otimes \mathbf{v}_x(v_r),
$$

(2)

where

$$
\mathbf{v}_z(u_r) = \begin{bmatrix} e^{-j2\pi \Delta_z u_r} & \cdots & e^{-(M_s-1)2\pi \Delta_z u_r} \end{bmatrix}^T, \quad
\mathbf{v}_x(v_r) = \begin{bmatrix} e^{-j2\pi \Delta_x v_r} & \cdots & e^{-(M_c-1)2\pi \Delta_x v_r} \end{bmatrix}^T.
$$

(3)

(4)

In this paper, both $d_z$ and $d_x$ are assumed to be equal to $\frac{\lambda}{2}. Then, we obtain that $\Delta_z = \Delta_x = \frac{1}{2}. Let \mathbf{V}$ be the matrix of sampled space steering vectors defined as

$$
\mathbf{V} = \mathbf{V}_z \otimes \mathbf{V}_x \in \mathbb{C}^{M_f \times N_r},
$$

(5)

where $N_r = N_{r,z} \times N_{r,z}$, $N_{r,z}$ and $N_{r,x}$ are the numbers of the sampled vertical and horizontal cosines respectively, the matrices $\mathbf{V}_z \in \mathbb{C}^{M_{r,z} \times N_{r,z}}$ and $\mathbf{V}_x \in \mathbb{C}^{M_{r,x} \times N_{r,x}}$ are given by

$$
\mathbf{V}_z = \begin{bmatrix} \mathbf{v}_z(u_{r,1}) & \mathbf{v}_z(u_{r,2}) & \cdots & \mathbf{v}_z(u_{r,N_s}) \end{bmatrix},
$$

(6)

$$
\mathbf{V}_x = \begin{bmatrix} \mathbf{v}_x(v_{r,1}) & \mathbf{v}_x(v_{r,2}) & \cdots & \mathbf{v}_x(v_{r,N_s}) \end{bmatrix},
$$

(7)

and $u_{r,i}$ and $v_{r,j}$ are sampled vertical and horizontal cosines respectively. We define $N_{a,z} = \frac{M_f}{M_s}$ and $N_{a,x} = \frac{M_f}{M_c}$ as the vertical and horizontal angle domain fine factors (FFs), respectively. Matrices $\mathbf{V}_z$ and $\mathbf{V}_x$ are over-sampled DFT matrices when $N_{a,z}$ and $N_{a,x}$ are integers, and the directional cosines are uniformly sampled in the range of $-1$ to $1$. Then, $\mathbf{V}_z$ and $\mathbf{V}_x$ can be represented as $\mathbf{V}_z = \mathbf{I}_{M_{r,z}} \mathbf{F}_{N_{r,z}}$ and $\mathbf{V}_x = \mathbf{I}_{M_{r,x}} \mathbf{F}_{N_{r,x}}$, respectively.

The frequency basis vector $\mathbf{b}_r(\tau_r) \in \mathbb{C}^{M_f \times 1}$ is given by

$$
\mathbf{b}_r(\tau_r) = \begin{bmatrix} e^{-j2\pi \Delta_f \tau_r} & \cdots & e^{-(M_f-1)2\pi \Delta_f \tau_r} \end{bmatrix}^T,
$$

(8)

where $\Delta_f = \frac{1}{M_{t,T}}$ is the frequency spacing between neighboring carriers and $\tau_r$ is the delay. We define the matrix of sampled frequency basis vectors as

$$
\mathbf{U} = [\mathbf{b}_r(\tau_{r,1}) \; \mathbf{b}_r(\tau_{r,2}) \; \cdots \; \mathbf{b}_r(\tau_{r,N_f})] \in \mathbb{C}^{M_f \times N_f},
$$

(9)

where $N_f$ is the number of the sampled delays and $\tau_{r,\ell}$ is the sampled delay. The delay domain fine factor is defined as $N_{a,p} = \frac{N_f}{M_f}$. Similarly, $\mathbf{U}$ is an over-sampled DFT matrix when $N_{a,p}$ is an integer and $\tau_r$ is uniformly sampled as

$$
\tau_{r,\ell} = \frac{\ell - 1}{N_f \Delta_f}, \quad \ell = 1, 2, \ldots, N_f.
$$

(10)

In this case, we have that $\mathbf{U} = \mathbf{I}_{M_f} \mathbf{F}_{N_f}$. We assume that the delay is within the guard interval, i.e., $\tau_r \leq M_f T_s$. We define $M_f = \left\lceil \frac{M_s M_c}{M_f} \right\rceil$ and $N_f = N_{a,p} M_f$, where the notation $\lceil \cdot \rceil$ represents rounding upwards. Then, the vector set $\{\mathbf{b}_r(\tau_{r,1}), \mathbf{b}_r(\tau_{r,2}), \ldots, \mathbf{b}_r(\tau_{r,N_f})\}$ is enough to contain all the sampled frequency basis vectors since $\frac{N_f}{N_r \Delta_f} \geq M_f T_s$. Thus, we further define $\mathbf{U}_f \in \mathbb{C}^{M_f \times N_f}$ as

$$
\mathbf{U}_f = \mathbf{U} \mathbf{I}_{N_{r,f}} = [\mathbf{b}_r(\tau_{r,1}) \; \mathbf{b}_r(\tau_{r,2}) \; \cdots \; \mathbf{b}_r(\tau_{r,N_f})].
$$

(11)

By using the 2D BSCM, the space-frequency domain channel matrix between the $k$-th UE and the BS for the $t$-th OFDM symbol can be modeled as [12], [13], [14], and [32]

$$
\mathbf{H}_{k,t} = \mathbf{V}(\mathbf{M}_k \odot \mathbf{W}_{k,t}) \mathbf{U}_f^T,
$$

(12)

where the matrix $\mathbf{M}_k$ is an $N_r \times N_f$ real deterministic matrix and remains unchanged in different OFDM symbols, and $\mathbf{W}_{k,t}$ is a complex Gaussian random matrix consisting of independent and identically distributed (i.i.d.) elements with zero mean and unit variance. We also assume that $\mathbf{W}_{k,t}$ and $\mathbf{W}_{k',t}$ are independent of each other when $k \neq k'$. Define $\mathbf{G}_{k,t} = \mathbf{M}_k \odot \mathbf{W}_{k,t}$, which is the angle-delay domain channel matrix and also called the 2D beam domain channel matrix (BDCM). In the channel model, the deterministic matrices $\mathbf{V}$ and $\mathbf{U}_f$ are generally non-unitary matrices. They are the same for all users. The rows of the deterministic matrix $\mathbf{V}$ decide which spatial beams or which angle of arrivals are used, whereas the columns of the deterministic matrix $\mathbf{U}_f$ decides which delays are used. The channel model is closer to the physical channel when more spatial beams and delays are used, and thus becomes more accurate. The BDCPM of the $k$-th user is defined as $\Omega_k = \mathbf{M}_k \odot \mathbf{M}_k$, which is a sparse matrix since most of the channel power is distributed in a limited number of resolvable spatial directions and time delays.

**B. Problem Formulation**

In the two-dimensional channel model (12), $\mathbf{U}_f$ and $\mathbf{V}$ are deterministic matrices. The unknown statistical parameter is the matrix $\mathbf{M}_k$ or equivalently the BPCMs $\Omega_k$. The statistical CSI or the matrix $\Omega_k$ can be exploited to schedule UTs and improve the estimation performance of instantaneous CSI, which will bring significant system performance gain. Thus, it is very important to obtain the statistical CSI or the matrix $\Omega_k$.

To estimate $\Omega_k$, we use the received pilot signals. We now consider the statistical CSI estimation based on the uplink pilot transmission. We use the pilot signal sequence in [33] as

$$
x_{q,p} = \mathbf{x}_q \odot \mathbf{b}_r(\tau_{r,(p-1)N_f}),
$$

(13)

where $q$ and $p$ denote the root coefficient and cyclic shift respectively, and $\tau_{r,(p-1)N_f}$ is the sampled delay defined
in (10). The sequence $\tilde{x}_q$ is the Zadoff-Chu (ZC) sequence with root coefficient $q$, which is specifically represented as [33]

$$[\tilde{x}_q]_n = \exp \left\{ -j \frac{\pi q n (n + 1)}{N} \right\}, \quad n = 0, \ldots, M_p - 1,$$  

where $N_f$ is the largest prime number such that $N_f < M_p$.

In particular, the pilots degenerate into orthogonal pilots (OPs) [34] when there is only one root. However, the overhead of orthogonal pilots is relatively large. There might not be enough pilot resources for the orthogonal pilots as the number of users increases in the massive MIMO systems. Thus, we use the non-orthogonal pilots (NOPs) in [33] to schedule more UTs in an OFDM symbol. We denote the number of roots and the number of UTs on the $q$-th root as $Q$ and $P_q$. Let the matrices $X_{q,p}$, $\tilde{X}_q$ and $B_r(\tau_r)$ denote $\text{diag}(x_{q,p})$, $\text{diag}(\tilde{x}_q)$ and $\text{diag}(b_r(\tau_r))$, respectively. We use the subscript $q,p$ to replace $k$ for convenience. The received pilot signal $Y_t \in \mathbb{C}^{M_x \times M_p}$ at the BS for the $t$-th OFDM symbol is given by [22] and [23]

$$Y_t = \sum_{k=1}^{K} H_{k,t} X_k + Z_t$$

$$= \sum_{q=1}^{Q} \sum_{p=1}^{P_q} H_{q,p} B_r(\tau_r,(p-1)N_f) \tilde{X}_q + Z_t,$$  

where $Z_t$ is a complex Gaussian noise matrix consisting of i.i.d. elements with zero mean and variance $\sigma_z^2$, which is assumed to be known. Substituting the channel model (12) into (15), we obtain that

$$Y_t = \sum_{q=1}^{Q} \sum_{p=1}^{P_q} VG_{q,p} U_T B_r(\tau_r,(p-1)N_f) \tilde{X}_q + Z_t$$

$$= \sum_{q=1}^{Q} V \left( \sum_{p=1}^{P_q} G_{q,p} I_{N_f,N_p} \Pi^{(p-1)N_f} \right) U_T \tilde{X}_q + Z_t,$$  

where the derivation of step (a) is provided in Appendix A. Let the matrix $G_{q,t}$ be defined as

$$G_{q,t} = \sum_{p=1}^{P_q} G_{q,p} I_{N_f,N_p} \Pi^{(p-1)N_f}.$$  

From the property that $B I \Pi^N$ is equivalent to the right cyclic shift of matrix $B$ by $n$ columns, we get

$$G_{q,p} I_{N_f,N_p} \Pi^{(p-1)N_f} = \begin{bmatrix} 0_{M_r, (p-1)N_f} & G_{q,p} & 0_{M_r, M_p - pN_f} \end{bmatrix}.$$  

To avoid mutual aliasing of UTs on the same root, the number of UTs on each root satisfies $P_q \leq \left\lceil \frac{N_f}{M_p} \right\rceil - \left\lceil \frac{M_r}{M_f} \right\rceil$. In this case, we obtain that

$$\tilde{G}_{q,t} = [G_{q,1,t} G_{q,2,t} \cdots G_{q,P_q,t} 0] \in \mathbb{C}^{N_f \times N_p},$$  

where the dimension of the zero matrix is $M_r \times (N_p - P_qN_f)$, and its subscript is omitted for convenience. According to (17), the received signal model becomes

$$Y_t = \sum_{q=1}^{Q} VG_{q,t} U_T \tilde{X}_q + Z_t.$$  

Let the matrices $G_t$ and $P$ be defined as

$$G_t = [\tilde{G}_{1,t} \tilde{G}_{2,t} \cdots \tilde{G}_{Q,t}],$$

$$P = [X_T^T U X_2^T U \cdots X_N^T U]^T.$$  

Then, the received signal model can be rewritten as

$$Y_t = VG_t P + Z_t.$$  

As with the way $G_t$ is constructed, we define $M$ and $\Omega$ as

$$M = \begin{bmatrix} M_1 & M_2 & \cdots & M_Q \end{bmatrix},$$

$$\Omega = \begin{bmatrix} \Omega_1 \Omega_2 \cdots \Omega_Q \end{bmatrix},$$

where $M_q$ and $\Omega_q$ are defined as

$$M_q = [M_{q,1} M_{q,2} \cdots M_{q,P_q} 0] \in \mathbb{R}^{N_f \times N_p},$$

$$\Omega_q = [\Omega_{q,1} \Omega_{q,2} \cdots \Omega_{q,P_q} 0] \in \mathbb{C}^{N_f \times N_p}.$$  

In the received signal model (28), the matrices $V$ and $P$ are deterministic matrices. The unknown statistical parameter is the matrix $M$ or equivalently the matrix $\Omega$. Since the statistical CSI varies slowly and remains the same over a short period of time, we can use multiple received pilot signals to estimate it. Thus, the problem is to estimate $\Omega$ based on multiple received pilot signals $Y_t$. The problem can also be viewed as an MMV problem [28]. Multiple instantaneous beam domain channel coefficients $G_t$ can be obtained by using the RM-FOCUSS method [28] and then be used to obtain the statistical $\Omega$. However, the RM-FOCUSS method still performs poorly at low SNR because it does not exactly establish the relationship between the received signal statistics and the noise statistics. Furthermore, the complexity of the RM-FOCUSS method is also not satisfactory because it needs to calculate the inversions of large dimensional matrices and estimate instantaneous CSI at each iteration. Thus, we need a new method to estimate the statistics directly.

III. ESTIMATION OF BEAM DOMAIN POWER MATRICES

A. 2D BDCPM Acquisition Algorithm Based on KL Divergence Minimization

Since the statistical parameters to be estimated are on the two-dimensional angle-delay domain, it is natural to convert the received pilot signals into signals on the angle-delay domain. According to (23), by left multiplying $Y_t$ with $V^H$
and right multiplying it with $P^H$, we obtain the received pilot signal on the angle-delay domain as

$$V^H Y_t P^H = V^H V G_t P P^H + V^H Z_t P^H. \quad (29)$$

The relation between the received pilot signal and the 2D BDCPM is not clear yet. To figure out their relation, we present an important property of the product of one random matrix with zero mean and independent entries and two deterministic matrices.

**Theorem 1**: Let $C_1 \in \mathbb{C}^{N_1 \times N_2}$ and $C_2 \in \mathbb{C}^{N_3 \times N_4}$ be constant matrices and $R \in \mathbb{C}^{N_2 \times N_3}$ be a random matrix with zero mean and independent entries, then

$$\mathbb{E}\{(C_1 R C_2) \odot (C_1 R C_2)^*\} = T_1 \mathbb{E}\{R \odot R^*\} T_2, \quad (30)$$

where $T_1 = C_1 \odot C_1^*$ and $T_2 = C_2 \odot C_2^*$.

**Proof**: The proof is provided in Appendix B. \(\square\)

Let $\Phi$ denote the expectation of the receive power matrix on the angle-delay domain as

$$\mathbb{E}\{\{(V^H Y P^H) \odot (V^H Y P^H)^*\}\}.$$ 

From Theorem 1, we obtain the received signal model of the statistical parameter $\Omega$ as

$$\Phi = T_0 \Omega T_f + N, \quad (31)$$

where $T_0$, $T_f$ and $N$ are known deterministic matrices defined as

$$T_0 = (V^H V) \odot (V^H V)^*, \quad (32)$$

$$T_f = (P P^H) \odot (P P^H)^*, \quad (33)$$

$$N = \mathbb{E}\{(Z \odot Z^*)\}(P^H \odot P^T) = M_1 M_2 \sigma^2_1 I, \quad (34)$$

To estimate the 2D BDCPM $\Omega$ from $\Phi$, we can solve (31) directly to obtain an exact solution. However, it is not an easy task since the elements of $\Omega$ need to be nonnegative. Even more, (31) might not have an exact solution. Thus, we propose to construct an optimization problem to obtain an approximate solution. To achieve this goal, we need an objective function first. This means that we need to choose a type of divergence or distance between the matrix $\Phi$ and the sum $T_0 \Omega T_f + N$ according to (31). Since the matrix $\Phi$ can be viewed as the observed angle-delay domain power spectrum, the Kullback-Leibler (KL) divergence for the spectral density functions [35] extended from that of the probability functions can be used. Because the sum of elements in $\Phi$ is not the same as that of $T_0 \Omega T_f + N$, we use the KL divergence for positive measures in [36] as the objective function.

Meanwhile, optimizing $\Omega$ directly is still complicated since it has the constraint that its elements are non-negative. From the relation $\Omega = M \odot M$, we know that estimating the matrix $M$ is equivalent to estimating the matrix $\Omega$. Thus, we choose to optimize $M$ instead of $\Omega$ because it has no constraint. We define the function $f(M)$ as the KL divergence between the matrices $\Phi$ and $T_0 \Omega T_f + N$, i.e., [36]

$$f(M) = \sum_{ij} \frac{[\Phi]_{ij}}{[T_0(M \odot M)T_f + N]_{ij}} \log \frac{[\Phi]_{ij}}{[T_0(M \odot M)T_f + N]_{ij}} + \sum_{ij} [T_0(M \odot M)T_f + N]_{ij} - \sum_{ij} [\Phi]_{ij}. \quad (35)$$

Using the KL divergence $f(M)$, we are now able to formulate an unconstrained optimization problem as

$$M^* = \arg \min_M f(M). \quad (36)$$

To solve the optimization problem in (36), the gradient method [37] can be used. Thus, we calculate the gradient of $f(M)$ with respect to $M$ first. In the following theorem, we provide the gradients of two items in $f(M)$.

**Theorem 2**: The gradient of $f(M)$ is obtained as

$$\frac{\partial f(M)}{\partial M} = 2(T_0 \text{1T}_f) \odot M - 2(T_0 Q T_f) \odot M,$$

where the matrix $Q$ is the Hadamard division of two matrices with the elements being defined as

$$[Q]_{ij} = \frac{[\Phi]_{ij}}{[T_0(M \odot M)T_f + N]_{ij}}. \quad (38)$$

**Proof**: The proof is provided in Appendix C. \(\square\)

With the obtained gradient, we can apply the gradient method to obtain the optimal $M$ as

$$M^{d+1} = M^d - \delta^d \frac{\partial f(M^d)}{\partial M^d}, \quad (39)$$

where the superscript $d$ represents the iteration number and $\delta^d$ is the step size which can be obtained by the line search method [38]. Recall that $\Phi$ denotes $\mathbb{E}\{\{(V^H Y P^H) \odot (V^H Y P^H)^*\}\}$. Thus, it is not possible to obtain the matrix $\Phi$ directly in practice. Instead, we use the sample average $\frac{1}{T} \sum_{t=1}^{T} (V^H Y_t P^H) \odot (V^H Y_t P^H)^*$, where $T$ is the number of samples. In the practical system, the channel realizations are usually correlated in the time domain. Thus, the effective sample size [39] will be smaller than $T$. The obtained algorithm is summarized as Algorithm 1.

Since the computational complexity of products is much higher than that of additions, we use the number of complex products as the computational complexity. Then, the complexity of Algorithm 1 is dominated by the matrix product $T_0 Q T_f$ in (37), whose complexity is $O(Q N_r N_p (N_r + Q N_p))$. The complexity is much lower than that of the RM-FOCUSS algorithm, which is of order $O((Q N_r N_p)^3)$. However, the complexity of Algorithm 1 is still not satisfactory for the 2D BSCM since $Q N_r N_p (N_r + Q N_p)$ is still very large. Thus, we need to further reduce the complexity of the proposed algorithm.

### B. Low-Complexity 2D BDCPM Acquisition Algorithm

In the previous subsection, we provide a received signal model that can be utilized to estimate the beam domain channel power matrix $\Omega$. However, the dimensions of $T_0$
Algorithm 1 2D BDCPM Acquisition Algorithm Based on KL Divergence Minimization

1: Use $\tilde{T}$ received pilot signals to calculate $\Phi$ as $\frac{1}{T} \sum_{t=1}^{T} \left( \mathbf{V}^H \mathbf{Y}_t \mathbf{P}^H \right) \odot \left( \mathbf{V}^H \mathbf{Y}_t \mathbf{P}^H \right)$.
2: **Initialization:** set $d = 0$ and the maximum number of iterations as $D$, select appropriate $\delta^0$, $\delta_{\text{min}} < \delta^0$ and $\alpha \in (0, 1)$, and initialize $\Omega^0 = \frac{1}{q_N N_p} \Phi$ and $M^0 = \sqrt{\Omega}$.
3: **Repeat**
4: Calculate the gradient $\frac{\partial f(M)}{\partial M}$ as (37).
5: while $\delta^d > \delta_{\text{min}}$
6: Update $M^{d+1} = M^d - \delta^d \frac{\partial f(M^d)}{\partial M^d}$
7: if $f(M^{d+1}) \geq f(M^d)$ then $\delta^d = \alpha \delta^d$ and $M^{d+1} = M^d$
8: else break
9: Set $d = d + 1$
10: until $\delta^d \leq \delta_{\text{min}}$ or $d = D$
11: Calculate $\Omega^d = M^d \odot M^d$

and $T_f$ are too large such that the matrix product $T_a \Omega T_f$ will cause high computational complexity. To reduce the computational complexity of Algorithm 1, the structure of $T_a$ and $T_f$ provided in the following theorem and corollary can be utilized.

**Theorem 3:** Let the matrix $\mathbf{A} = \mathbf{1}_{M,N} \mathbf{F}_N$ be an oversampled DFT matrix and $\mathbf{D} = \text{diag}(\mathbf{d})$ be an $M$-dimensional diagonal matrix. Then, $(\mathbf{A}^H \mathbf{D}) \odot (\mathbf{A}^H \mathbf{D})^*$ is a circulant matrix, given by

$$ (\mathbf{A}^H \mathbf{D}) \odot (\mathbf{A}^H \mathbf{D})^* = \mathbf{F}_N^H \mathbf{A} \mathbf{F}_N, $$

where $\Lambda$ is a diagonal matrix defined as

$$ \Lambda = \frac{1}{N} \text{diag} \left( \mathbf{F}_N \left( (\mathbf{F}_N^H \mathbf{d}) \odot (\mathbf{F}_N^H \mathbf{d})^* \right) \right), $$

$$ \tilde{\mathbf{d}} = [\mathbf{d}^T \ 0_{N-M1}^T]^T. $$

**Proof:** The proof is provided in Appendix D. □

**Theorem 3** is obtained based on the properties of circulant matrices [40]. From Theorem 3, we then obtain the following corollary.

**Corollary 1:** The matrices $T_a$ and $T_f$ can be written as

$$ T_a = (\mathbf{F}_{N_p} \odot \mathbf{F}_{N_q})^H (\mathbf{A}_z \odot \mathbf{A}_x) (\mathbf{F}_{N_p} \odot \mathbf{F}_{N_q}), $$

$$ T_f = \mathbf{I}_Q \odot \mathbf{F}_{N_p}^H \mathbf{\Sigma} (\mathbf{I}_Q \odot \mathbf{F}_{N_q}), $$

where $\mathbf{A}_z$ and $\mathbf{A}_x$ are diagonal matrices, defined as

$$ \mathbf{A}_z = \frac{1}{N_z} \text{diag} \left( \mathbf{F}_{N_z} \left( (\mathbf{F}_N^H \mathbf{d}_z) \odot (\mathbf{F}_N^H \mathbf{d}_z)^* \right) \right), $$

$$ \mathbf{A}_x = \frac{1}{N_x} \text{diag} \left( \mathbf{F}_{N_x} \left( (\mathbf{F}_N^H \mathbf{d}_x) \odot (\mathbf{F}_N^H \mathbf{d}_x)^* \right) \right), $$

$$ \mathbf{d}_z = [\mathbf{1}^T_{M_z} 0_{N_z-M_z1}^T]^T, $$

$$ \mathbf{d}_x = [\mathbf{1}^T_{M_x} 0_{N_x-M_x1}^T]^T, $$

and $\Sigma$ is a block matrix with diagonal matrices being its elements, defined as

$$ \Sigma = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,Q} \\ \vdots & \ddots & \vdots \\ \Sigma_{Q,1} & \cdots & \Sigma_{Q,Q} \end{bmatrix}. $$

**Proof:** The proof is provided in Appendix E. □

Based on the structures of $T_a$ and $T_f$ provided in Corollary 1, we can reduce the complexity of the proposed algorithm. In the following, we analyze the complexity of the proposed algorithm after utilizing the structure provided in Corollary 1.

We write the matrix $\mathbf{Q}$ as $[\mathbf{Q}_1 \  \mathbf{Q}_2 \  \cdots \  \mathbf{Q}_Q]$ for convenience, where $\mathbf{Q}_q \in \mathbb{C}^{N_q \times N_p}, \forall q$. According to Corollary 1, the matrix product $T_a \mathbf{Q} T_f$ can be written in (52), shown at the bottom of the next page.

There are three kinds of matrix products in (52). For convenience, we analyze their complexities after utilizing the structure as follows.

1) The first kind of product is the product between an $M \times N$ matrix and an $N$-dimensional diagonal matrix. Its complexity is $O(NM)$.

2) The second kind of product is the product between an $M \times N$ matrix and an $N$-dimensional DFT matrix, which can be implemented by using the fast Fourier transform (FFT) and has complexity of $O(NM \log_2 N)$.

3) The third kind of product is the product between a matrix and the Kronecker product of two DFT matrices. For example, we consider the matrix product $(\mathbf{F}_{N_1} \odot \mathbf{F}_{N_2}) \mathbf{A}$, where $\mathbf{A}$ can be written for convenience as

$$ \mathbf{A} = [\text{vec}(\mathbf{A}_1) \ \text{vec}(\mathbf{A}_2) \ \cdots \ \text{vec}(\mathbf{A}_M)] \in \mathbb{C}^{N \times M} $$

and $\mathbf{A}_n \in \mathbb{C}^{N_2 \times N_1}, \forall n$. It can be calculated in the following way as

$$ (\mathbf{F}_{N_1} \odot \mathbf{F}_{N_2}) \mathbf{A} = [\text{vec}(\mathbf{F}_{N_2} \mathbf{A}_1 \mathbf{F}_{N_1}) \ \cdots \ \text{vec}(\mathbf{F}_{N_2} \mathbf{A}_M \mathbf{F}_{N_1})]. $$

Its complexity after using FFT is $O(NM \log_2 N)$.

By utilizing the structure provided in Corollary 1, we can obtain a low-complexity version of Algorithm 1, the details of which are omitted for brevity. Based on the complexity analysis of all three kinds of products provided above, the computational complexity of the low-complexity method after utilizing the structure is calculated as $O(QN_q N_p \log_2(N_q N_p))$, which is much lower than that of using the direct matrix product.

**C. Angle Domain BDCPM Acquisition Method in Frequency-Flat Fading Channels**

In the previous subsections, we have introduced the method of obtaining 2D BDCPM based on the model in (23).
The proposed method is applicable as long as the received signal model satisfies the following form

$$\mathbf{Y} = \mathbf{AGB} + \mathbf{Z},$$

(53)

where $\mathbf{A}$ and $\mathbf{B}$ are deterministic matrices, $\mathbf{G}$ is a random matrix with independent entries, $\mathbf{Y}$ is the received matrix, and $\mathbf{Z}$ is a noise matrix. The matrix $\mathbf{\Omega} = \mathbb{E}\{\mathbf{G} \odot \mathbf{G}^*\}$ is the statistical parameter to be estimated.

In this subsection, we present another application of the proposed method. We consider massive MIMO transmission over frequency-flat fading channels. It can be seen as a narrow-band sub-carrier of the considered massive MIMO-OFDM system. In this system, $K$ UTs equipped with $M_t$ antennas send pilot signals to a base station equipped with $M_r$ antennas. Let $\mathbf{X}_k \in \mathbb{C}^{M_r \times T}$ denote the uplink pilot signal transmitted by the $k$-th user. We assume the pilot signals of different users are orthogonal to each other for simplicity.

The received channel matrix $\mathbf{H}_{k,m} \in \mathbb{C}^{M_r \times M_t}$ on the $m$-th slot at the BS can be written as

$$\mathbf{H}_{k,m} = \mathbf{V}_r \mathbf{G}_{k,m} \mathbf{V}_t^T,$$

(54)

where $\mathbf{G}_{k,m}$ denotes the angle domain channel matrix, $\mathbf{V}_r$ and $\mathbf{V}_t$ are the conversion matrices from the angle domain to the space domain at the transmitter and receiver, respectively. Similarly to the 2D BSCM, $\mathbf{V}_r$ and $\mathbf{V}_t$ need not be unitary matrices.

The received pilot signal $\mathbf{Y}_m \in \mathbb{C}^{M_r \times T}$ on the $m$-th slot at the BS can be written as

$$\mathbf{Y}_m = \sum_{k=1}^{K} \mathbf{H}_{k,m} \mathbf{X}_k + \mathbf{Z}_m = \sum_{k=1}^{K} \mathbf{V}_r \mathbf{G}_{k,m} \mathbf{V}_t^T \mathbf{X}_k + \mathbf{Z}_m,$$

(55)

where $\mathbf{Z}_m$ is the noise matrix whose elements are i.i.d. complex Gaussian random variables with zero mean and variance $\sigma^2$.

Under the assumption that the pilot matrices of different UTs are orthogonal to each other, it is easy to obtain that

$$\mathbf{V}_r^H \mathbf{Y}_m \mathbf{X}_k \mathbf{V}_t^* = (\mathbf{V}_r^H \mathbf{V}_t) \mathbf{X}_k + (\mathbf{V}_r^H \mathbf{Z}_m) \mathbf{X}_k \mathbf{V}_t^* + (\mathbf{V}_r^H \mathbf{G}_{k,m} \mathbf{V}_t^T) \mathbf{X}_k + (\mathbf{V}_r^H \mathbf{Z}_m) \mathbf{X}_k \mathbf{V}_t^*,$$

(56)

By defining $\mathbf{T}_r$, $\mathbf{T}_t$ and $\mathbf{N}$ as

$$\mathbf{T}_r = (\mathbf{V}_r^H \mathbf{V}_t) \mathbf{X}_k \mathbf{V}_t^*,$$

(57)

$$\mathbf{T}_t = (\mathbf{V}_t^H \mathbf{X}_k \mathbf{V}_t^*) \mathbf{X}_k \mathbf{V}_t^*,$$

(58)

$$\mathbf{N} = \sigma^2 (\mathbf{V}_r^H \mathbf{V}_t) \mathbf{X}_k \mathbf{V}_t^* + (\mathbf{V}_r^H \mathbf{Z}_m) \mathbf{X}_k \mathbf{V}_t^*,$$

(59)

we obtain from Theorem 1 that

$$\mathbb{E}_{p_{\mathbf{X}}} \{ (\mathbf{V}_r^H \mathbf{Y}_m \mathbf{X}_k \mathbf{V}_t^*) \mathbf{X}_k \mathbf{V}_t^* \} = \mathbf{T}_r \mathbf{\Omega} \mathbf{T}_t + \mathbf{N},$$

(60)

where the matrix $\mathbf{\Omega}_k$ is the angle domain channel power matrix of the $k$-th user, which is defined as $\mathbb{E}\{\mathbf{G}_{k,m} \odot \mathbf{G}_{k,m}^*\}$. It is obviously that the received signal model has the same structure as (31). Therefore, the proposed method can be used here to estimate the angle domain channel power matrix.

IV. Simulation Results

In this section, we provide simulation results to show the performance of the proposed algorithm. The antenna spacing on the base station side is set to half wavelength and all UTs are equipped with single antennas. The major parameters of OFDM system are summarized in Table I. According to these parameters, we can obtain $M_f = \frac{M_r M_t}{M_p} = 9$. Therefore, the maximum number of UTs that can be scheduled by orthogonal pilots in an OFDM symbol is $\frac{M_r M_t}{M_f} = 13$. In the simulations, we select two scenarios with different numbers of UTs, one of which is $K = Q \times P = 1 \times 12$ and the other is $K = Q \times P = 2 \times 12$, where $Q$ and $P$ are the number of roots and the number of UTs per root, respectively. We set the fine factors as $N_a = N_a = N_a = 2$, which is enough to obtain good performance [22]. Furthermore, the power of the transmitted pilots is set to 1. The space-frequency domain channel $\mathbf{H}_{k,t}$ has been normalized as $\|\mathbf{H}_{k,t}\|_F = M_r M_p$, where $\|\cdot\|_F$ represents the Frobenius norm. The signal-to-noise ratio (SNR) is given by $\text{SNR} = \frac{\eta}{\sigma^2}$.

We adopt two methods to generate channels for simulation. The first one is to generate the channels according to the 2D BSCM in (12). In this case, the accurate BDCPMs are known, and thus can be used to evaluate the accuracy of the estimated BDCPMs. The second one is to generate the channels by the widely used channel model QuaDRiGa [41], which is close to the physical channel and will be used to verify the convergence of the algorithm, the sparsity of the estimated BDCPMs and the performance gain achieved by the channel estimation based on the estimated BDCPMs.

A. Channels Generated According to the 2D BSCM

We evaluate the accuracy of estimated BDCPMs using the channels generated from the 2D BSCM with given BDCPMs.

| Parameters          | Value               |
|---------------------|---------------------|
| Carrier frequency $f_c$ | 4.5 GHz             |
| Subcarrier spacing $\Delta f$ | 30 kHz             |
| Guard interval $M_g$     | 124                |
| Number of subcarriers $M_s$ | 2048               |
| Number of pilot subcarriers $M_p$ | 120                |

\[ \mathbf{T}_a \mathbf{Q} \mathbf{T}_f = (\mathbf{F}_{N_c} \otimes \mathbf{F}_{N_h})^H (\mathbf{A}_v \otimes \mathbf{A}_h) (\mathbf{F}_{N_c} \otimes \mathbf{F}_{N_h}) \mathbf{Q} (\mathbf{I}_Q \otimes \mathbf{F}_{N_p})^H \mathbf{\Sigma} (\mathbf{I}_Q \otimes \mathbf{F}_{N_p}) \]

\[ \sum_{q=1}^{Q} (\mathbf{F}_{N_c} \otimes \mathbf{F}_{N_h})^H (\mathbf{A}_v \otimes \mathbf{A}_h) (\mathbf{F}_{N_c} \otimes \mathbf{F}_{N_h}) \mathbf{Q}_{q}^H \mathbf{\Sigma}_{q} \mathbf{F}_{N_p} \]

\[ = \sum_{q=1}^{Q} (\mathbf{F}_{N_c} \otimes \mathbf{F}_{N_h})^H (\mathbf{A}_v \otimes \mathbf{A}_h) (\mathbf{F}_{N_c} \otimes \mathbf{F}_{N_h}) \mathbf{Q}_{q}^H \mathbf{\Sigma}_{q} \mathbf{F}_{N_p} \mathbf{g} \]
Results of the proposed algorithm are used to compare with that of the RM-FOCUSS algorithm, which has been verified to have high accuracy under high SNR among a series of compressed sensing algorithms. The massive MIMO channel with $M_{r,z} = 8$, $M_{r,x} = 16$ is used for simulation. Since the accurate BDCPMs are known, we can use the normalized mean squared error (NMSE) between the estimated BDCPM $\hat{\Omega}_k$ and the accurate BDCPM $\Omega_k$ to evaluate the accuracy. The NMSE in dB is defined as

$$\text{NMSE\,(dB)} \triangleq 10 \log_{10} \left( \frac{1}{K} \sum_{k=1}^{K} \frac{\|\hat{\Omega}_k - \Omega_k\|_2^2}{\|\Omega_k\|_F^2} \right)$$  \hspace{1cm} (61)$$

Simulation results of the NMSE performance of estimated BDCPMs are shown in Fig. 1, where the massive MIMO system is with $M_{r,z} = 8$, $M_{r,x} = 16$, $T = 80$. It can be observed that the accuracy of BDCPMs obtained by the proposed algorithm is not less than that obtained by the RM-FOCUSS method, no matter whether orthogonal pilots with $K = 12$ or non-orthogonal pilots with $K = 24$ are used. Meanwhile, it can also be found that the proposed algorithm has strong anti-noise performance, while the accuracy of the RM-FOCUSS algorithm drops sharply when the SNR is less than 0 dB. The reason why the proposed algorithm outperforms the RM-FOCUSS algorithm at low SNR is that the proposed method has accurately established the relationship between the received signal statistics and the noise statistics, i.e., the received signal model in equation (30), while the RM-FOCUSS algorithm does not.

The accuracy of estimated BDCPMs also depends on the number of received pilot signals, which has been set as $T = 80$ in the simulations for Fig. 1. To show the relationship between the NMSE performance of the estimated BDCPMs and the number of samples used in the estimation, we simulate the NMSE of the estimated BDCPMs for different numbers of samples in the considered massive MIMO system with $M_{r,z} = 8$, $M_{r,x} = 16$. The numbers of samples are set as $T = 10, 20, 40, 80$. Simulation results of $K = 12$ and $K = 24$ for all cases are provided in Fig. 2 and Fig. 3, respectively. We observe that the NMSE of estimated BDCPMs for both the RM-FOCUSS algorithm and the proposed algorithm can achieve close to $-10$ dB performance with only 10 samples of received pilot signals. Furthermore, the NMSE performances of the two algorithms in high SNR regime decrease almost linearly as the number of received pilot signals increases. Thus, we do not need too many samples of receive pilot signals in practical massive MIMO systems to obtain the statistical CSI with good accuracy. Finally, the proposed algorithm greatly outperforms the RM-FOCUSS algorithm in the low SNR.

Fig. 1. NMSE of the estimated BDCPM versus SNRs for four scenarios in a massive MIMO system with $M_{r,z} = 8$, $M_{r,x} = 16$, $T = 80$, where the channels are generated from the 2D BSCM, the number of UTs is 12 and 24, and the methods are the RM-FOCUSS algorithm and the proposed algorithm.

Fig. 2. NMSE of the estimated BDCPM versus SNRs for four different numbers of samples in a massive MIMO system with $M_{r,z} = 8$, $M_{r,x} = 16$, $K = 12$, where the channel is generated from the 2D BSCM and the methods are the RM-FOCUSS algorithm and the proposed algorithm.

Fig. 3. NMSE of the estimated BDCPM versus SNRs for different numbers of samples in a massive MIMO system with $M_{r,z} = 8$, $M_{r,x} = 16$, $K = 24$, where the channel is generated from the 2D BSCM and the methods are the RM-FOCUSS algorithm and the proposed algorithm.

Fig. 4. The objective function value under different iteration times in massive MIMO system with $M_{r,z} = 8$, $M_{r,x} = 16$, where the channel is generated from QuaDRiGa.
Fig. 5. Estimated angle delay domain power spectrum in dB of the first user in a massive MIMO system with $M_{r,z} = 8$, $M_{r,x} = 16$ and (a) SNR = 30 dB, $K = 12$, (b) SNR = −10 dB, $K = 12$, (c) SNR = 30 dB, $K = 24$, (d) SNR = −10 dB, $K = 24$. The channel is generated from QuaDRiGa.

regime, and achieves nearly the same performance as that of the latter in the high SNR regime for all cases. It indicates that the proposed algorithm is robust to the noise, whereas the RM-FOCUSS method needs a higher SNR to work.

B. Channels Generated by QuaDRiGa

In this subsection, we use the widely used massive MIMO channel model QuaDRiGa for simulation. In QuaDRiGa, the scenario is set to “3GPP_3D_UMa_NLOS”. The shadow fading and path loss are not considered and the UTs in the cell are randomly uniformly distributed. In most simulations, massive MIMO with $M_{r,z} = 8$, $M_{r,x} = 16$ is used. In channel estimation simulation, the performance in the extra-large scale MIMO scenario with $M_{r,z} = 16$, $M_{r,x} = 64$ is also verified.

In Fig. 4, the convergence performance of the proposed algorithm in four scenarios in massive MIMO system with $M_{r,z} = 8$, $M_{r,x} = 16$ are shown. The SNRs under consideration are −10 dB and 30 dB, and the numbers of UTs are 12 and 24. It can be observed that, the proposed algorithm can approach convergence within 20 iterations in all the scenarios. Furthermore, the algorithm converges faster with fewer UTs than with more UTs.

In Fig. 5, we give the estimated angle delay domain power spectrum, i.e., the BDCPMs of the first UT under different SNRs and numbers of UTs for the considered massive MIMO with $M_{r,z} = 8$, $M_{r,x} = 16$. The first UT is the same UT for $K = 12$ and $K = 24$. It can be found that the BDCPMs obtained are sparse. By comparing Fig. 5(a) with Fig. 5(b) or Fig. 5(c) with Fig. 5(d), it can be found that reducing the SNR has little effect on the obtained BDCPMs. This indicates that the proposed algorithm has a prominent anti-noise effect. Similarly, by comparing Fig. 5(a) with Fig. 5(c) or Fig. 5(b) with Fig. 5(d), we can find that the BDCPM in Fig. 5(c) or Fig. 5(d) has more non-zero beams, e.g., the part indicated by the green arrow in Fig. 5(c) and Fig. 5(d), which is mainly caused by the interference of the pilot signals on another root. Therefore, the pilot interference on other roots will reduce the accuracy of the BDCPMs to some extent.

Finally, we use the estimated BDCPMs for the estimation of instantaneous CSI. We consider both the massive MIMO with $M_{r,z} = 8$, $M_{r,x} = 16$ and extra-large scale MIMO with $M_{r,z} = 16$, $M_{r,x} = 64$. Due to its high complexity, the RM-FOCUSS method is only used in massive MIMO scenarios. The MMSE algorithm is used for instantaneous channel estimation. We define the mean square error of space-frequency domain channel in dB as

$$\text{MSE}(\text{dB}) \triangleq 10 \log_{10} \left( \frac{1}{K T} \sum_{k=1}^{K} \sum_{t=1}^{T} \left\| \hat{H}_{k,t} - H_{k,t} \right\|_F^2 \right),$$

(62)

where $\hat{H}_{k,t}$ and $H_{k,t}$ are the estimated channel and accurate channel, respectively. The simulation results of the channel estimation performance of the considered massive MIMO and extra-large scale MIMO are shown in Fig. 6 and Fig. 7, respectively. It can be observed in Fig. 6 that when the...
SNR is less than 10 dB, the proposed algorithm can bring significant channel estimation performance gain compared to the RM-FOCUSS algorithm. When the SNR is greater than 10 dB, the proposed algorithm can also obtain comparable performance to the RM-FOCUSS algorithm. Since the proposed algorithm has a complexity much lower than the RM-FOCUSS algorithm, it is superior to the RM-FOCUSS algorithm in estimating the statistical channel information. Finally, the effectiveness of the proposed algorithm in an extra-large scale MIMO system is also verified in Fig. 7. The results of the RM-FOCUSS algorithm are not given because its complexity is too high to simulate.

V. CONCLUSION

In this paper, the beam domain statistical CSI estimation for the 2D BSCM in massive MIMO systems was investigated. We considered the problem of estimating the BDCPMs based on multiple received pilot signals. From the 2D BSCM, we derived a received signal model, which shows the relation between the statistical properties of the received pilot signals and the BDCPMs. Then, we formulated an optimization problem based on the Kullback-Leibler (KL) divergence and the received signal model. A novel method that estimates the BDCPMs without estimating instantaneous CSI was proposed by solving the optimization problem. The complexity of the proposed method was further reduced by utilizing the circulant structures. It was shown in simulation that the proposed method has good convergence and can obtain sparse BDCPMs. Besides, compared with the RM-FOCUSS, the proposed algorithm can obtain overall more accurate statistical CSI with much lower complexity, and can bring significant performance gains when used in channel estimation.

APPENDIX A
DERIVATION OF STEP (A) IN EQUATION (16)

From (11) and (9), we can obtain in (63), shown at the bottom of the page. According to the expression of $b_r(\tau_r)$ in (8), we can get that $b_r(\tau_r) \odot b_r(\tau_r) = b_r(\tau_r, (\tau_r,i+j)_{np})$. Then it is obtained that

$$U_f^T B_r(\tau_r, (p-1)N_f) = I_{N_f, N_p} \Pi^{(p-1)N_f}_{N_p} U_f^T,$$  

(64)

where the property of the permutation matrix is used, that is, $\Pi^N_B$ is equal to the upward cyclic shift of matrix $B$ by $N$ rows. Substituting the above equation into (16), we get that

$$Y_t = \sum_{q=1}^{Q} \sum_{p=1}^{P_q} G_{q, p} \{I_{N_f, N_p} \Pi^{(p-1)N_f}_{N_p}\} U_f^T \tilde{X}_q + Z_{t},$$

(65)

APPENDIX B
PROOF OF THEOREM 1

We define the matrix $\Gamma$ as $\Gamma = E\{[C_1 R C_2] \odot [C_1 R C_2]^T\}$ for convenience. Then, the entries $[\Gamma]_{ij}$ can be specifically represented as

$$[\Gamma]_{ij} = E\{[C_1 R C_2]_{ij} \odot [C_1 R C_2]^T_{ij}\}$$

$$= E\{\sum_{p,q} [C_1]_{ip} [R]_{pq} [C_2]_{qj}\} E\{[R]_{pq} [R]_{p'q'} [C_2]_{q'j}\}^T$$

$$= \sum_{p,q,p',q'} [C_1]_{ip} [C_1]_{ip'} E\{[R]_{pq} [R]_{p'q'} [C_2]_{q'j}\} [C_2]_{qj}^T [C_2]_{q'j}^T.$$

(66)

Since each element of $R$ is independent of each other and has zero mean, we can obtain $E\{[R]_{pq} [R]_{p'q'}\} = \delta(p-p') \delta(q-q') E\{||R||^2\}$, where $\delta(x)$ equals 1 when $x = 0$, 0 otherwise.
otherwise it is 0. Then the expression of \( [\Gamma]_{ij} \) can be simplified to
\[
[\Gamma]_{ij} = \sum_{p,q} |[C_1]_{ip}|^2 E \{ |[R]_{pq}|^2 \} |[C_2]_{iq}|^2.
\] (67)
This can be organized in a matrix form as
\[
\Gamma = (C_1 \odot C_1^*) E \{ R \odot R^* \} (C_2 \odot C_2^*).
\] (68)
By defining \( T_1 = C_1 \odot C_1^* \) and \( T_2 = C_2 \odot C_2^* \), we then obtain
\[
E \{ (C_1 G_t C_2) \odot (C_1 G_t C_2)^* \} = T_1 E \{ R \odot R^* \} T_2 \quad \text{and}
\] (69)
Thus, we conclude the proof.

**APPENDIX C**

**PROOF OF THEOREM 2**

We use \( f_1(M) \) and \( f_2(M) \) to represent the two terms of the objective function (35) with respect to \( M \), i.e.,
\[
f_1(M) = \sum_{ij} |T_a(M \odot M)T_f|_{ij},
\] (70)
\[
f_2(M) = \sum_{ij} |\Phi|_{ij} \log |T_a(M \odot M)T_f + N|_{ij}.
\] (71)

Then \( f(M) \) can be written as \( f_1(M) - f_2(M) + C \), where \( C \) is a constant independent of \( M \).

We calculate the gradient \( \frac{\partial[T_a(M \odot M)T_f]}{\partial M} \), first. The entries \( [T_a(M \odot M)T_f]_{ij} \) are written as
\[
[T_a(M \odot M)T_f]_{ij} = \sum_{p,q} [T_a]_{ip} |M|_{pq} [T_f]_{aq}.
\] (72)

We define \( e_i \) as the column vector whose only nonzero entry is its \( i \)-th element, which has a value of 1. From \( [T_a]_{ip} [T_f]_{aq} = [T_a^T e_i e_j^T T_f^T]_{pq} \), we have that
\[
[T_a(M \odot M)T_f]_{ij} = \sum_{p,q} [T_a^T e_i e_j^T T_f^T]_{pq} |M|_{pq}
\] (73)
\[
\sum_{p,q} \left[ (T_a^T e_i e_j^T T_f^T) \odot (M \odot M) \right]_{pq}.
\]

Then, we can obtain
\[
\frac{\partial[T_a(M \odot M)T_f]}{\partial |M|_{pq}} = 2 [T_a^T e_i e_j^T T_f^T]_{pq} |M|_{pq}.
\] (74)

so the gradient \( \frac{\partial[T_a(M \odot M)T_f]}{\partial M} \) is given by
\[
\frac{\partial[T_a(M \odot M)T_f]}{\partial M} = 2 (T_a^T e_i e_j^T T_f^T) \odot M.
\] (75)

Next, the gradient of the first item \( f_1(M) \) is obtained as
\[
\frac{\partial f_1(M)}{\partial M} = 2 \left( \sum_{ij} T_a^T e_i e_j^T T_f^T \right) \odot M
\] (76)
Furthermore, the gradient of the second item \( f_2(M) \) is calculated as
\[
\frac{\partial f_2(M)}{\partial M} = \sum_{ij} \frac{\partial [\Phi]_{ij}}{\partial M} \frac{\partial |T_a(M \odot M)T_f|_{ij}}{\partial M}
= 2 \sum_{ij} \frac{[\Phi]_{ij}}{|T_a(M \odot M)T_f + N|_{ij}} (T_a^T e_i e_j^T T_f^T) \odot M.
\] (77)

By defining \( Q \) as \( |Q|_{ij} = \frac{|[\Phi]_{ij}}{|T_a(M \odot M)T_f + N|_{ij}} \), we then obtain
\[
\frac{\partial f_2(M)}{\partial M} = 2 \left( \sum_{ij} [Q]_{ij} T_a^T e_i e_j^T T_f^T \right) \odot M
= 2 (T_a^T QT_f^T) \odot M.
\] (78)
Since \( T_a \) and \( T_f \) are symmetric matrices, i.e., \( T_a^T = T_a \), \( T_f^T = T_f \), and \( f(M) = f_1(M) - f_2(M) + C \), the gradient of \( f(M) \) is obtained as
\[
\frac{\partial f(M)}{\partial M} = 2 (T_a^T T_f) \odot M - 2 (T_a^T QT_f) \odot M
= 2 (T_a^T (I - Q) T_f) \odot M.
\] (79)

**APPENDIX D**

**PROOF OF THEOREM 3**

From the properties of the circulant matrix [40], we know that \( A^H D A \) is a circulant matrix when \( D = \text{diag}(d) \) is a diagonal matrix and \( A = I_{M \times N} F_N \) is an over-sampled DFT matrix. Therefore, \( A^H D A \) can be represented in the following form
\[
A^H D A = \sum_{i=0}^{N-1} [c]_i \Pi_N^{N-i},
\] (80)
where \( c \) is the first column of the matrix \( A^H D A \). Since the first column of \( A \) is an all-one vector, then \( c \) can be calculated as
\[
c = (A^H D A) e_1 = A^H d = F_N^H d.
\] (81)
The vector \( d \) is defined as \( d^T \Omega_{N-1}^T \). Then, we have
\[
(A^H D A) \odot (A^H D A)^* = \left( \sum_{i=0}^{N-1} [c]_i \Pi_N^{N-i} \right) \odot \left( \sum_{j=0}^{N-1} [c]_j \Pi_N^{N-j} \right)^*
= \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} [c]_i [c]_j (\Pi_N^{N-i} \odot \Pi_N^{N-j})^*
= \sum_{i=0}^{N-1} [c]_i [c]_j \Pi_N^{N-i}.
\] (82)
The step (a) holds because \( \Pi_N^j \odot \Pi_N^j = \delta(i - j) \Pi_N^j \).
Obviously, \( (A^H D A) \odot (A^H D A)^* \) is also a circulant matrix with \( c \odot c^* \) in the first column. Thus, it can be represented as
\[
(A^H D A) \odot (A^H D A)^* = F_N^H A F_N.
\] (83)
where $\Lambda = \text{diag}(\lambda)$ is a diagonal matrix and satisfies
\[ c \otimes c^* = F_N^H \Lambda. \hspace{1cm} (84) \]
Then, we obtain
\[ \Lambda = \frac{1}{N} \text{diag}(F_N(c \otimes c^*)) \]
\[ = \frac{1}{N} \text{diag} \left( F_N \left( F_N^H d \otimes (F_N^H d)^* \right) \right). \hspace{1cm} (85) \]
Thus, we conclude the proof.

**APPENDIX E**

**PROOF OF COROLLARY 1**

Let us review the formulas of $T_a$ and $T_f$ from equations (32) and (33). First, we substitute $V = V_z \otimes V_x$ into (32) as
\[
T_a = (V^H V) \otimes (V^H V)^* = ((V_z \otimes V_x)^H(V_z \otimes V_x)) \otimes ((V_z \otimes V_x)^H(V_z \otimes V_x))^*. \hspace{1cm} (86)
\]
By using the mixed product properties of the Kronecker product $(A \otimes B)(C \otimes D) = AC \otimes BD$ and $(A \otimes B)(C \otimes D) = (A \otimes C)(B \otimes D)$, we then obtain
\[
T_a = \left( (V_z^H V_z) \otimes (V_z^H V_z)^* \right) \otimes \left( (V_x^H V_x) \otimes (V_x^H V_x)^* \right). \hspace{1cm} (87)
\]
From Theorem 3 and the fact that $V_z = I_{M_z,N_z} F_{N_z}$ and $V_x = I_{M_x,N_x} F_{N_x}$, are over-sampled DFT matrices, we obtain that $(V_z^H V_z) \otimes (V_z^H V_z)^*$ and $(V_x^H V_x) \otimes (V_x^H V_x)^*$ are circulant matrices, given by
\[
(V_z^H V_z) \otimes (V_z^H V_z)^* = F_{N_z}^H \Lambda_z F_{N_z}, \hspace{1cm} (88)
\]
\[
(V_x^H V_x) \otimes (V_x^H V_x)^* = F_{N_x}^H \Lambda_x F_{N_x}, \hspace{1cm} (89)
\]
where $\Lambda_z$ and $\Lambda_x$ are diagonal matrices defined in the theorem. Therefore, $T_a$ can be represented as
\[
T_a = (F_{N_z} \otimes F_{N_x})^H(\Lambda_z \otimes \Lambda_x)(F_{N_z} \otimes F_{N_x}). \hspace{1cm} (90)
\]
by using the mixed product property again.

Next, we derive the DFT structure of $T_f$. From (33) and (22), $T_f$ can be expressed in block matrix form as
\[
T_f = \begin{bmatrix} B_{1,1} & \cdots & B_{1,Q} \\ \vdots & \ddots & \vdots \\ B_{Q,1} & \cdots & B_{Q,Q} \end{bmatrix}, \hspace{1cm} (91)
\]
where the sub-matrix $B_{q_1,q_2}$ is defined as
\[
B_{q_1,q_2} = (U^T X_{q_1} X_{q_2}^H U^*) \otimes (U^T X_{q_1} X_{q_2}^H U^*). \hspace{1cm} (92)
\]
Since $U = I_{M_p,N_p} F_{N_p}$ is an over-sampled DFT matrix and $X_{q_1} X_{q_2}^H = \text{diag}(\overline{x}_{q_1}, \overline{x}_{q_2}^H)$ is a diagonal matrix, we can obtain from Theorem 3 that each submatrix can be represented as
\[
B_{q_1,q_2} = F_{N_p}^H \Sigma_{q_1,q_2} F_{N_p}. \hspace{1cm} (93)
\]
where $\Sigma_{q_1,q_2}$ is a diagonal matrix defined in the theorem. By combining (91) and (93), we have
\[
T_f = \begin{bmatrix} F_{N_p}^H \Sigma_{1,1} F_{N_p} & \cdots & F_{N_p}^H \Sigma_{1, Q} F_{N_p} \\ \vdots & \ddots & \vdots \\ F_{N_p}^H \Sigma_{Q,1} F_{N_p} & \cdots & F_{N_p}^H \Sigma_{Q, Q} F_{N_p} \end{bmatrix}, \hspace{1cm} (94)
\]
where $\Sigma$ is defined as
\[
\Sigma = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,Q} \\ \vdots & \ddots & \vdots \\ \Sigma_{Q,1} & \cdots & \Sigma_{Q,Q} \end{bmatrix}. \hspace{1cm} (95)
\]
Thus, we conclude the proof.

**ACKNOWLEDGMENT**

The authors would like to thank the editor and the anonymous reviewers for their helpful comments and suggestions.

**REFERENCES**

[1] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[2] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, “Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits,” *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.

[3] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An overview of massive MIMO: Benefits and challenges,” *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[4] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassoul, “Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution,” *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May 2016.

[5] L. Sanguinetti, E. Björnson, and J. Hoydis, “Toward massive MIMO 2.0: Understanding spatial correlation, interference suppression, and pilot contamination,” *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 232–257, Jan. 2020.

[6] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, “Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?” *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3070, Jun. 2015.

[7] L. Liu, R. Chen, S. Geirhofer, K. Sayana, Z. Shi, and Y. Zhou, “Downlink MIMO in LTE-advanced: SU-MIMO vs. MU-MIMO,” *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 140–147, Feb. 2012.

[8] C. B. Barneto, S. D. Liyanaarachchi, M. Heino, T. Riihonen, and M. Valkama, “Full duplex radio/radar technology: The enabler for advanced joint communication and sensing,” *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 82–88, Feb. 2021.

[9] X. Zheng, A. Liu, and V. Luu, “Joint channel and location estimation of massive MIMO system with phase noise,” *IEEE Trans. Signal Process.*, vol. 68, pp. 2598–2612, 2020.

[10] E. Björnson and L. Sanguinetti, “Making cell-free massive MIMO competitive with MMSE processing and centralized implementation,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[11] A. Lu, X. Gao, W. Zhong, C. Xiao, and X. Meng, “Robust transmission for massive MIMO downlink with imperfect CSI,” *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5362–5376, Aug. 2019.

[12] C. Wang, A. Lu, X. Gao, and Z. Ding, “Robust precoding for 3D massive MIMO configuration with manifold manifold optimization,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3423–3437, May 2022.

[13] A. Lu, Y. Zhang, and X. Gao, “Iterative generalized eigenvector precoder with deterministic equivalents for 3D massive MIMO,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11784–11795, Nov. 2021.

[14] A. Lu, X. Gao, and C. Xiao, “Robust linear precoder design for 3D massive MIMO downlink with a posteriori channel model,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7274–7286, Jul. 2022.

[15] Y. Zhang, A. Lu, and X. Gao, “Sum-Rate-Optimal statistical precoding for FDD massive MIMO downlink with deterministic equivalents,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7359–7370, Jul. 2022.
