Variational Learning Algorithms for Channel Estimation in RIS-Assisted mmWave Systems

Milind Nakul, Anupama Rajoriya, Member, IEEE, and Rohit Budhiraja, Member, IEEE

Abstract—We consider the problem of estimating channel in reconfigurable intelligent surface (RIS) assisted millimeter wave (mmWave) systems. We propose two variational expectation maximization (VEM) based channel estimation algorithms, which exploit the angular domain sparsity of RIS-assisted mmWave channel. To fully capture this sparsity, both within and across UEs, we construct a novel column-wise coupled Gaussian prior. The first proposed structured-mean-field-based VEM (SMF-VEM) algorithm uses the proposed prior, and calculates the posterior distribution of the unknown channel by assuming that it belongs to a set of multivariate distributions. This algorithm inverts a high-dimensional matrix in its posterior update, and consequently does not scale well for a large number of RIS elements and base station antennas, which are commonly used in practical systems. The second proposed fast mean field-based VEM (FMF-VEM) algorithm reduces complexity by assuming a fully-factorized posterior. It also bounds the variational objective to remove the residue coupling between the channel and phase matrices. Using extensive numerical investigations for a practical RIS mmWave system, and by using multiple metrics, we show that the proposed i) SMF- and FMF-VEM algorithms outperform several of their state-of-the-art counterparts; and ii) FMF-VEM has a much lower time complexity than SMF-VEM.

Index Terms—Variational inference, expectation maximization, sparse Bayesian learning (SBL).

I. INTRODUCTION

RECONFIGURABLE intelligent surface (RIS) is a promising technology to assist millimeter wave (mmWave) systems in achieving high spectral and energy efficiency [1], [2], [3], [4]. An RIS consists of a large number of passive reflecting elements, which shift the phase of the incident signal, and reflect it passively. The phase shift of each RIS element can be configured to improve the received signal strength [1], [2], [3], [4]. This is especially useful when the direct path between a user equipment (UE) and base station (BS) is blocked by obstructions [1], [2], [3], [4]. There has been extensive research for optimizing the BS beamformer and RIS reflection coefficient matrix [1], [2], [3]. These works assume availability of accurate channel information at the BS for their designs, which, however, is difficult to obtain in RIS systems. This is due to its large number of passive reflecting elements [2].

He et al. in [2] proposed a two-step channel estimator for RIS-aided massive multiple-input multiple-output (MIMO) systems. This design assumes to know the binary RIS reflection coefficient matrix, and requires it to be sparse, for accurately estimating the channel. The ON/OFF switching of passive RIS elements in [2] is costly, as this requires a separate amplitude control for each RIS element [1]. Also, since the ON/OFF mode needs to be implemented for each individual passive RIS element, it leads to a sub-optimal solution [3]. Taha et al. in [3] estimated the channel by including active elements, with signal processing capability in the RIS. This design used compressive sensing tools to construct channels of all RIS elements by using the channels of only the active elements. The installation of active elements with signal processing capability in RIS is, however, costly, and unsuitable in many applications [4].

The cascaded UE—RIS—BS channel in the virtual angular domain (VAD), due to a limited number of scatterers around the BS and RIS, exhibits sparsity, which [2], [3] do not fully exploit. They, hence, require large training overhead for estimating the channel [4]. Specifically, in the cascaded VAD channel, the (i) common scatterers between the BS and RIS for all UEs cause common-column sparsity; (ii) partially-shared scatterers between some of the UEs cause partially-common row sparsity; and (iii) scatterers individual to the UE cause UE-specific sparsity [4], [5], [6], [7], [8], [9], [10]. These sparsities allow the use of fewer number of pilots, and cast channel estimation as a compressed sensing problem [1]. Several recent works [4], [5], [6], [7], [8], [9], [10] have proposed cascaded channel estimation algorithms for RIS-assisted mmWave systems by exploiting these sparsities.

Wang et al. in [4] employed orthogonal matching pursuit (OMP) and generalized approximate message passing (GAMP) algorithms for sparse cascaded channel in a single-user RIS-assisted mmWave system. They did not consider any particular sparsity structure. Lin et al. in [5] proposed alternating minimization and manifold optimization estimation algorithms for RIS-assisted mmWave MIMO systems. The algorithms used $\ell_1$-minimization and OMP techniques. They also, similar to [4], did not exploit the structured sparsity of the cascaded channel. The authors in [6] and [7] proposed an OMP-based algorithm for RIS-assisted mmWave system with

0090-6778 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
uniform linear and planar arrays, respectively. They exploited the correlation among UEs, which occurs due to the common BS-RIS channel between them. He et al. in [8] proposed a two-stage channel estimation approach in RIS-assisted MIMO systems. The first stage estimates the angle of arrival (AoAs) and angle of departures (AoDs) of the cascaded channel, and the second stage estimates the channel using atomic norm minimization. These works in [6], [7], and [8] only exploited the common-column sparsity, but not the partially-common row and UE-specific sparsities. Wei et al. in [9] exploited both common-column and partially-common row sparsities, jointly referred to as the doubly-structured sparsity of the cascaded VAD channel. They proposed a doubly-structured orthogonal matching pursuit (DS-OMP) based algorithm for RIS-assisted mmWave systems. This design separately estimates the row and column supports of all the UEs. Ju et al. in [10] showed that, in addition to the doubly-structured sparsity, the cascaded VAD channel also has an additional column index shift, which is common for all UEs. After compensating for this shift, all non-zero rows of each UE have the same column support. Their design first calculated the shift, and then applied an OMP-based algorithm to estimate the compensated channel. The OMP-based algorithms in [4], [6], [7], [9], and [10] assume that the number of scatterers i.e., the sparsity is known a priori. However, acquiring sparsity information itself has a high complexity for a large number of BS antennas or RIS elements [1]. The optimization-based algorithms in [8] and [5] outperform their OMP-based counterparts. They, however, require fine-tuning of the regularizing parameters [11]. Also, the performance of OMP- and optimization-based algorithms depend on the pilot matrix structure [11]. The above works in [4], [5], [6], [7], [8], [9], and [10], thus, either fail to exploit the complete cascaded channel sparsity, or require extra parameters to be known a priori.

Sparse Bayesian learning (SBL) is another robust and powerful compressed sensing framework [12]. The RIS-assisted channel estimation literature has also made some progress in exploiting Bayesian learning [13], [14]. The SBL techniques, without requiring any sparsity information and condition on the pilot matrix, also provide an uncertainty measure of the estimate [12]. The choice of prior distribution, which captures the sparsity structure of the unknown channel, is a crucial step in designing SBL-based algorithms [12]. Ruan et al. in [13] developed an approximate message passing (AMP) based Bayesian learning algorithm for VAD channel estimation in a multi-user RIS-assisted mmWave system. To capture the VAD channel sparsity structure, they assumed Bernoulli-Gaussian and Gaussian mixture priors, and used a vector AMP algorithm. They deployed a testing site site T1 to aid the channel estimation. In the off-line phase, T1 transmits pilot signals to the BS for estimating the RIS-BS channel. In the on-line phase, T1 transmits pilot signals to UEs for estimating the RIS-UE channel. The AMP-based algorithm in [13] does not exhaustively exploit the complete sparsity of the cascaded VAD channel. Dong et al. in [14] proposed a variational Bayesian learning-based algorithm for RIS-assisted orthogonal frequency division multiplexing (OFDM) system. They clustered the adjacent RIS elements, and used a Dirichlet prior to capture the clustering. Their algorithm captured the delay-domain OFDM channel sparsity. However, for angular domain channel, it fails to exploit the UE-specific sparsity, since it a-priori clusters the RIS elements.

The SBL algorithms [12], [14] are, however, iterative in nature, and have a high complexity. The complexity is due to the inversion of high-dimensional matrices in each iteration, and the coupled sparse channel elements and the dictionary matrix [15], [16], [17]. Several variational inference (VI) based sparse signal recovery algorithms have been developed to design computationally-efficient, and scalable Bayesian algorithms [15], [16], [17]. The VI framework treats the inference in signal estimation as an optimization problem. This allows us to approximate even intractable posterior distributions, and then use simpler forms of the approximating distribution, for a faster convergence and reduced complexity. Thomas et al. in [15] proposed the space-alternating variational estimation (SAVE) algorithm for sparse vector recovery. It avoided inverting the high-dimensional matrix using a variational approximation from [18]. The SAVE algorithm, however, only partially decouples the signal elements and the dictionary matrix [16], [17]. The remaining coupling does not allow the SAVE updates to run in parallel, which makes its complexity grow with $O(N^2)$, where $N$ is the sparse signal dimension [17]. Duan et al. in [16] proposed an inverse-free SBL (IF-SBL) algorithm to further reduce the SAVE algorithm complexity. They removed the coupling between the dictionary matrix and the unknown channel elements by bounding the objective function, which is obtained using the variational approximation. The approximation further reduces the IF-SBL complexity, when compared with SAVE and SBL. It, however, results in an approximated posterior distribution, which is very different from that of the SAVE and SBL, and is also invalid [17]. To achieve a good posterior approximation, and reduced complexity, Worley in [17] proposed a fast-mean field (FMF) algorithm by bounding the objective function in SAVE using minorization maximization (MM) technique, which is guaranteed to converge to a locally optimal solution. The bound in [17], unlike the one in IF-SBL algorithm [16], ensures that the obtained approximate posterior is valid. The FMF-SBL algorithm proposed in [17] requires only two matrix-vector products in each SBL iteration, and thus readily scales to large dimensional problems. All these low-complexity algorithms [15], [16], [17], which are designed for a fundamental problem of sparse signal vector recovery, do not consider any structured sparsity within the sparse vector. These methods have also, so far, not been applied for estimating channel in RIS-aided mmWave systems.

To summarize, the existing literature for RIS-assisted wireless communication systems [4], [5], [6], [7], [8], [9], [10] has not yet designed a Bayesian learning based channel estimation technique, which exhaustively exploits the entire sparsity structure of the cascaded VAD channel. The Bayesian works in [13] and [14] neither captured the complete sparsity of the cascaded VAD channel, nor designed a low-complexity algorithm for high-dimensional RIS-assisted mmWave systems. The current work proposes two Bayesian learning channel estimation algorithms for RIS-assisted mmWave system by exploiting the
that the bound used in the FMF-VEM algorithm is valid. The FMF-VEM algorithm is shown to have a much lower runtime than the SMF-VEM algorithm.

Notations: The \( m \)th element of vector \( y \) is denoted as \( y_m \). The symbols \( x_{m,n}, x_{n}^T \), and \( x_n \) denote the \((m,n)\)th element, \(m\)th row and \(n\)th column of matrix \( X \), respectively. The symbol \( I_M, I_M \) and \( 0_M \) denote \( M \times M \) identity matrix, \( M \times 1 \) all-one and all-zero vectors, respectively. The notations \( \mathbb{I}() \) denote the indicator function. The notation \( \|\cdot\|_2 \) denotes the \( \ell_2 \) norm of a vector. The statistical expectation with respect to a distribution \( p() \) is denoted as \( \mathbb{E}_{p()} \). The notation \( \mathcal{CN}(x|a,b) \) represents a complex Gaussian random variable \( x \) with mean \( a \) and variance \( b \).

II. SYSTEM MODEL

We consider an uplink RIS-assisted mmWave system with an \( M \)-antenna BS and \( K \) single-antenna UEs. The RIS is designed as an \( N \)-antenna uniform planar array (UPA) with \( N_1 \) horizontal and \( N_2 \) vertical reconfigurable elements, such that \( N = N_1 \times N_2 \). Each RIS element is passive, has no signal processing capability, and only shifts the phase of the incident signal [9]. The uplink UE→RIS channel is denoted as \( h_{R_k} \in C^{N_1 \times 1} \). While the uplink BS→RIS channel is denoted as \( H^B \in C^{M \times N} \). This work estimates cascaded UE→RIS→BS uplink channel.

We consider, similar to [1] and [20], that the channel estimation period is divided into \( G \) sub-phases. Each sub-phase consists of \( T \) time slots wherein the \( k \)th UE transmits a \( T \)-length orthogonal pilot sequence \( x_k = [x_{k1}, \ldots, x_{kT}]^T \in C^{T \times 1} \). Here \( \|x_k\|^2 = p_k \), and \( x_{k}^T x_{k'} = 0 \) for all \( k' \neq k \) UEs. The scalar \( p_k \) is the pilot power of the \( k \)th UE. The UEs transmit same pilot sequences in each sub-phase. The RIS reflection vectors, which are denoted as \( \phi_g \in \{\phi_g,1, \ldots, \phi_g,\overline{N}\}^T \in C^{N_1 \times 1} \) are, however, different in different sub-phases [1], [20]. Here, \( \phi_g,n \) denotes the reflection coefficient of the \( n \)th RIS element, in the \( g \)th sub-phase.

Now, the sum pilot signal received at the BS in \( T \) time slots of \( G \)th sub-phase, which is denoted as \( Y_g \in C^{M \times T} \), is given as follows

\[
Y_g = \sum_{k=1}^{K} (h_{Dk} \ast H_{R_k}) \cdot x_k^T + w_g
\]

The vectors \( h_{Dk} \in C^{N \times 1} \) and \( h_{R_k} \in C^{N_1 \times 1} \) denote the channels of the \( k \)th UE to the BS and the RIS, respectively. The channel matrix \( H^B \in C^{M \times N} \) denotes the BS→RIS channel.

Equality in (a) is obtained by interchanging \( \phi_g \) and \( h_{R_k} \). The vector \( w_g \), with density \( \mathcal{CN}(0, \sigma^2 I_M) \), is the additive white Gaussian noise at the BS. By post-multiplying both sides in (1) with \( x_k^T \), we get

\[
Y_g x_k = (a) (h_{Dk} \ast H_{R_k})^T x_k + \sum_{k' \neq k} (h_{Dk'} \ast H_{R_k'}) x_k' x_k + w_g x_k
\]

\[
y_{kg} = (b) (h_{Dk} \ast H_{R_k}) p_k + w_{kg}.
\]
In equality (a), \( \mathbf{H}^k \triangleq \mathbf{H}^B \text{diag}(\mathbf{h}^R_k) \in \mathbb{C}^{M \times N} \) is the cascaded channel of the \( k \)-th UE. In equality (b), the vector \( \mathbf{y}_k = \mathbf{y}_{k} \mathbf{x}_k \in \mathbb{C}^{M \times 1} \) is the receive pilot of the \( k \)-th UE. This equality exploits the orthogonality of pilot sequence \( \mathbf{x}_k \).

In this work, similar to the existing literature [21, references therein], we do not estimate the direct UE–BS link channel. We assume that direct link is either absent due to blockages, or its channel is first estimated and then removed from the received pilot signal. The direct link channel, for example, can be estimated by switching off the cascaded UE–RIS–BS link channel by switching off of all the RIS elements [22]. The direct channel estimation, now, reduces to conventional mmWave channel estimation, which can be performed using estimators such as least squares, minimum mean squared error (MMSE), orthogonal matching pursuit and sparse Bayesian learning [23, references therein]. The last two techniques exploit sparsity in the direct channel, and are shown to outperform the first two techniques, which do not exploit sparsity.

The direct channel after estimation is removed from the receive signal in (2). The equivalent received pilot signal after removing direct link contribution is expressed as follows

\[
\mathbf{y}_k = \mathbf{H}^k \mathbf{\Phi}_g \mathbf{p}_k + \mathbf{w}_k. \tag{3}
\]

We now concatenate \( \mathbf{y}_k \) for \( G \) sub-phases by assuming, similar to [9], that pilot power \( p_k = 1 \):

\[
\mathbf{Y}^k = \mathbf{H}^k \mathbf{\Phi} + \mathbf{W}^k. \tag{4}
\]

Here \( \mathbf{Y}^k = [\mathbf{y}_{k1}, \ldots, \mathbf{y}_{kG}] \in \mathbb{C}^{M \times G} \) is the concatenated \( G \) sub-phases receive pilot, \( \mathbf{\Phi} = [\mathbf{\phi}_1, \ldots, \mathbf{\phi}_G] \in \mathbb{C}^{G \times G} \) and \( \mathbf{W}^k = [\mathbf{w}_{k1}, \ldots, \mathbf{w}_{kG}] \in \mathbb{C}^{M \times G} \). The \( G \)-phase pilot transmission model, as explained in detail in [1], Section II-A2), [20], helps in estimating the decoupled end-to-end channel of UE \( k \). The channel estimation phase now requires \( GT \) time slots, where \( T = K \) ensures orthogonality across UE pilots. The number of sub-phases \( G \) can, however, be reduced by using the structured sparsity of the cascaded channel, as shown numerically later in Section VI. This overhead is negligible for pedestrian/low-speed UEs considered herein, which have a coherence interval of few lakhs of symbols [24]. We now estimate the cascaded UE–RIS–BS channel, for given RIS reflection coefficients. Before doing that, we digress to discuss its VAD representation, and its sparsity structure in the next two sub-sections.

### A. Virtual Angular Domain (VAD) Representation of the Cascaded UE–RIS–BS Channel

We begin by providing the Saleh-Valenzuela model for the RIS–BS channel \( \mathbf{H}^B \) [9]:

\[
\mathbf{H}^B = \sqrt{\frac{MN}{P^B}} \sum_{p=1}^{P^R} \alpha_p^B \mathbf{b}(\vartheta_p^R, \psi_p^R) \mathbf{a}(\vartheta_p^R, \psi_p^R)^T = \mathbf{U}^B \mathbf{A}^B (\mathbf{U}^R)^H \mathbf{b}(\vartheta_p^R, \psi_p^R) \mathbf{a}(\vartheta_p^R, \psi_p^R)^T \tag{5}
\]

The scalar \( P^B \) is the number of scatterers between the BS and RIS. For the \( p \)-th channel path, (i) \( \alpha_p^B \) is the complex gain; (ii) \( \vartheta_p^R \) (resp. \( \psi_p^R \)) is the azimuth (resp. elevation) angle of arrival (AoA) at the BS; and (iii) \( \vartheta_p^R \) (resp. \( \psi_p^R \)) is the azimuth (resp. elevation) angle of departure (AoD) at the RIS. The vectors \( \mathbf{a}(\vartheta, \psi) \in \mathbb{C}^{N \times 1} \) and \( \mathbf{b}(\vartheta, \psi) \in \mathbb{C}^{M \times 1} \) represent the normalized array steering vectors at the RIS and the BS, respectively. The scalars \( \vartheta \) and \( \psi \) represent the azimuth and elevation angles, respectively. For a UPA of size \( N = N_1 \times N_2 \), the normalized array steering vector at the RIS, \( \mathbf{a}(\vartheta, \psi) \) can be represented as follows [9]:

\[
\mathbf{a}(\vartheta, \psi) = \frac{1}{\sqrt{N}} \left[ e^{-j2\pi d \sin(\vartheta) \cos(\psi) m_1/\lambda}, e^{-j2\pi d \sin(\vartheta) \cos(\psi) m_2/\lambda} \right]. \tag{6}
\]

The normalized array steering vector at the BS, \( \mathbf{b}(\vartheta, \psi) \) can similarly be represented as [10]:

\[
\mathbf{b}(\vartheta, \psi) = \frac{1}{\sqrt{M}} \left[ e^{-j2\pi d \sin(\vartheta) \cos(\psi) m_1/\lambda}, e^{-j2\pi d \sin(\vartheta) \cos(\psi) m_2/\lambda} \right]. \tag{7}
\]

Here \( n_i = [0, 1, \ldots, N_i - 1]^T \) and \( m_i = [0, 1, \ldots, M_i - 1]^T \), for \( i = 1, 2 \). The term \( \lambda \) is the carrier wavelength, and \( d \) is the BS antenna spacing. In equality (a) in (5), matrices \( \mathbf{U}^B \in \mathbb{C}^{M \times P^B} \) and \( \mathbf{U}^R \in \mathbb{C}^{N \times P^R} \) are the true array response matrices at the BS and RIS side, respectively [1]. The \( p \)-th column of \( \mathbf{U}^B \) represents the array steering vector \( \mathbf{b}(\vartheta_p^R, \psi_p^R) \) for the AoA pair \( (\vartheta_p^R, \psi_p^R) \) at the BS. Similarly, the \( p \)-th column of \( \mathbf{U}^R \) represents the array steering vectors for the \( p \)-th AoD pair at the RIS. The diagonal matrix \( \mathbf{A}^B = \text{diag}(\alpha_1^B, \ldots, \alpha_{P^B}^B) \in \mathbb{C}^{P^B \times P^B} \) denotes the path gains of the scatterers between the BS and the RIS. In equality (b), we have defined the over-complete array response matrices \( \mathbf{U}^B \in \mathbb{C}^{M \times M} \) and \( \mathbf{U}^R \in \mathbb{C}^{N \times M} \), which contain \( M \gg P^B \) AoA pairs at the BS and \( N \gg P^B \) AoD pairs at the RIS, discretized equally in the spatial grid [1], [4].

We assume, similar to [4], that the true AoA and AoD pairs corresponding to the \( P^B \) scatterers lie on this discretized grid. The channel representation using these over-complete array response matrices is referred to as its VAD representation [1]. The matrix \( \mathbf{A}^B \in \mathbb{C}^{M \times N} \) now becomes sparse, whose \( (m, n) \)-th non-zero entry represents a scatterer corresponding to the \( m \)-th AoA pair at the BS and \( n \)-th AoD pair at the RIS. The VAD representation enables us to exploit the channel sparsity, and improve the estimation performance.

The \( \mathbf{U}^k \)–RIS channel, denoted as \( \mathbf{h}^R_k \), is similarly represented as follows [9]:

\[
\mathbf{h}^R_k = \sqrt{\frac{N}{Q_k^{Rk}}} \sum_{q=1}^{Q_k^{Rk}} \alpha_k^{Rq} \mathbf{a}(\vartheta_k^{Rq}, \psi_k^{Rq}) = \mathbf{U}^R \mathbf{\alpha}^{Rk} \mathbf{a}(\vartheta_k^{Rk}, \psi_k^{Rk}) = \mathbf{U}^R \mathbf{\alpha}^{Rk}. \tag{8}
\]

The term \( Q_k^{Rk} \) is the number of scatterers between RIS and the \( k \)-th UE. For the \( q \)-th channel path, \( \alpha_k^{Rq} \) is the complex gain, and \( \vartheta_k^{Rq} \) and \( \psi_k^{Rq} \) are its azimuth and elevation AoAs respectively, at the RIS. The matrices \( \mathbf{U}^{Rk} \in \mathbb{C}^{N \times Q_k^{Rk}} \) and \( \mathbf{\alpha}^{Rk} \in \mathbb{C}^{N \times 1} \) are true and over-complete array steering matrices at the RIS [1]. The vectors \( \mathbf{e}_k \in \mathbb{C}^{Q_k^{Rk} \times 1} \) and \( \mathbf{\alpha}^{Rk} \in \mathbb{C}^{N \times 1} \) represent the non-sparse and sparse path gains, respectively [1].

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
The cascaded channel $\mathbf{H}^k = \mathbf{H}^B \text{diag}(\mathbf{h}^{Rk})$ can now be expressed as follows:

$$
\mathbf{H}^k \stackrel{(a)}{=} \mathbf{H}^B \circ (\mathbf{h}^{Rk})^T \stackrel{(b)}{=} (\mathbf{U}^B \mathbf{A}^B (\mathbf{U}^R)^H) \circ (\mathbf{U}^{Rk} \mathbf{\alpha}^{Rk})^T \stackrel{(c)}{=} \mathbf{U}^B \left( \mathbf{A}^B \otimes (\mathbf{\alpha}^{Rk})^T \right) (\mathbf{U}^R)^H \circ (\mathbf{U}^{Rk})^T.
$$

(9)

The notation $\circ$ in equality (a) denotes the Khatri-Rao product \cite{1}. Equality (b) is obtained by using (5) and (8). Equality (c) is obtained by using the property $(\mathbf{AB}) \circ (\mathbf{CD}) = (\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D})$. \cite{1}. In equality (c), we also define $\mathbf{H}^{k} \triangleq (\mathbf{A}^{k})^H \otimes (\mathbf{\alpha}^{Rk})^* \in \mathbb{C}^{N \times M}$ and $\mathbf{U}^{Rk} \triangleq (\mathbf{U}^{Rk})^T \circ (\mathbf{U}^R)^* \in \mathbb{C}^{N \times N^2}$, which are the effective VAD channel and array steering matrix between the RIS and UE, respectively \cite{4}. Since, as discussed above, both $\mathbf{A}^B$ and $\mathbf{\alpha}^{Rk}$ are sparse, their Kronecker product is also sparse \cite{4}. This sparse representation in (9) can be further simplified by noticing that the matrix $\mathbf{U}^{Rk}$ contains multiple redundant columns due to the Khatri-Rao product. Specifically, using \cite{4}, Proposition 1, we can write (9) as

$$
\mathbf{H}^k = \mathbf{U}^B (\mathbf{H}^{k})^H (\mathbf{U}^{Rk})^H \stackrel{(a)}{=} \mathbf{U}^B (\mathbf{H}^{k})^H (\mathbf{U}^{Rk})^H.
$$

(10)

Here $\mathbf{H}^{k} \in \mathbb{C}^{N \times M}$ and $\mathbf{U}^{Rk} \in \mathbb{C}^{N \times N}$ are the reduced versions of $\mathbf{H}^k$ and $\mathbf{U}^{Rk}$, respectively \cite{4}. The VAD representation of $\mathbf{H}^{k}$, which is denoted as $\mathbf{H}^{k}$, is now obtained by assuming the array steering matrices $\mathbf{U}^B = \mathbf{U}_M$ and $\mathbf{U}^{Rk} = \mathbf{U}_N$, as discrete Fourier transform (DFT) matrices \cite{1,4}. The columns and rows of $\mathbf{H}^{k}$ represent the effective channel gains corresponding to scatterers between BS and RIS, and RIS and $k_{th}$ UE, respectively. This VAD representation is similar to that of massive MIMO channel. The limited scattering causes different kind of sparsities in $\mathbf{H}^{k}$, which are discussed next.

B. Sparsity Structure of the Cascaded VAD Channel

1) Common-Column Sparsity: The matrix $\mathbf{H}^{k}$ in (10) is the cascaded VAD channel. Each of its $(n,m)$th entry represents the effective channel gain corresponding to the $m_{th}$ AoA pair at the BS and effective nth azimuth and elevation pairs at the RIS. The channel $\mathbf{H}^{k}$ has only $P^B$ non-zero columns, which correspond to the $P^B$ scatterers between the BS and RIS. We know from (10) that the VAD sparse channel $\mathbf{A}^B$ between BS and RIS is common for all the $K$ UEs. The AoA pairs for $p = 1, \ldots, P^B$ scatterers between BS and RIS are, thus, same for all the UEs. The non-zero column support of $\mathbf{H}^{k}$, denoted as $\Omega^{k}_{c}$, is, thus, also same for all $K$ UEs i.e., $\Omega^{k}_{c} = \Omega^{k}_{c} = \cdots = \Omega^{k}_{c} = \Omega_{c}$. This is shown for an RIS-aided mmWave system in Fig. 1, which serves two different UEs. Their non-zero column support is given as $\Omega_{c} = \{1, 4\}$, which is due to the $P^B = 2$ scatterers between BS and RIS, denoted in Fig. 1a as Scatterers 1 and 2. These common-column supports are denoted using the red outlines in the channels $\mathbf{H}^{1}$ and $\mathbf{H}^{2}$, in Fig. 1b. Since this support is common for all the UEs, the shared sparsity among them can be exploited by jointly estimating their channel. Also note that, due to the common BS-RIS scatterers, the column entries within each UE channel, also have a shared sparsity, which is denoted in Fig. 1b using the dashed black box in the fourth column of $\mathbf{H}$.

2) Partially-Common Row Sparsity: In each non-zero column, only a few rows are non-zero. This is due to a limited number of scatterers around the RIS. The indices of these non-zero rows, as seen from (10), is decided by the azimuth and elevation angle at the RIS. In each non-zero column, there are thus $Q^{Rk}$ non-zero rows, which correspond to the number of scatterers between the RIS and UE. In Fig. 1, we consider $Q^{R1} = Q^{R2} = 2$ scatterers between the RIS and UE1 and UE2. Some of these scatterers, e.g., Scatterer $q = 2$, is common for different UEs. They lead to the channels $h^{Rk}$, for $k$ UEs, having partially-common paths, with the same AoAs and AoDs at the RIS. These paths are shown by the green lines in Fig. 1b. If the number of partially-common paths between the UE and RIS is denoted as $Q^{R}$, with $Q^{R} \leq Q^{Rk}$, then for each non-zero column in $\mathbf{H}^{k}$, there are $Q^{R}$ common paths between all UEs’ channels \cite{9}. If we denote the non-zero row entries of the $m_{th}$ column and $k_{th}$ UE as $\Omega^{Rk}_{m}$, then

$$
\bigcap_{k=1}^{K} \Omega^{Rk}_{m} = \Omega^{R}_{m}.
$$

(11)

Here $\Omega^{R}_{m}$ denotes the partially-common row support for the $m_{th}$ column for all UEs. In Fig. 1, we consider scatterers between the i) BS and RIS is $P^B = 2$; and ii) UE and RIS for each UE is $Q^{Rk} = 2$. We see that Scatterer $q = 2$ between the RIS and UEs, is common to both the UEs. Out of $Q^{Rk}$ paths, thus $Q^{R} = 1$ path is common to both the UEs. This implies that for each UE, there are a total of $P^B \times Q^{Rk} = 4$ non-zero elements in the cascaded VAD channel $\mathbf{H}$. Out of which, $P^B \times Q^{R} = 2$ entries are common across all the UEs, due to the common scatterer between the UEs and RIS. The partially-common row support in the first and fourth column is $\Omega^{R}_{m} = \{3\}$ and $\Omega^{R}_{m} = \{1\}$, respectively. These common non-zero rows are shown in green square in Fig. 1. The common-column sparsity, along with the partially-common row sparsity is, henceforth, referred to as the **doubly-structured** sparsity.

3) UE-Specific Sparsity: The scatterers which are specific to each UEs, cause UE-specific sparsity. This sparsity causes UE-specific $Q^{Rk} - Q^{R}$ non-zero elements in each non-zero column. For example, in Fig. 1, Scatterer $q = 1$ for UE 1 causes one non-zero element in each non-zero column of $\mathbf{H}^{1}$, which is not common with $\mathbf{H}^{2}$. This is denoted by the purple square. The UE-specific sparsity of UE2, due to Scatterer $q = 1$ for UE 2, is denoted using orange square.

By substituting the VAD representation from (10) into (4), pre-multiplying with $\mathbf{U}^H$, and by taking the conjugate transpose, we get:

$$
\mathbf{Y}^{k} = \mathbf{\Phi} \mathbf{H}^{k} + \mathbf{W}^{k}.
$$

(12)

Here $\mathbf{Y}^{k} = (\mathbf{U}^H \mathbf{y}^{k})^H \in \mathbb{C}^{G \times M}$ is the effective received pilot matrix, $\mathbf{\Phi} = (\mathbf{U}^H \mathbf{\Phi})^H \in \mathbb{C}^{G \times N}$ is the effective phase matrix, and $\mathbf{W}^{k} = (\mathbf{U}^H \mathbf{w}^{k})^H \in \mathbb{C}^{G \times M}$ is the effective noise matrix.
A joint system model for all the UEs using (12) can be expressed as follows

$$\hat{Y} = \tilde{\Phi} \tilde{H} + \tilde{W}. \quad (13)$$

Here $\hat{Y} = [\hat{Y}_1^T, \ldots, (\hat{Y}_i^K)^T]^T \in \mathbb{C}^{GK \times M}$ is joint receive pilot matrix, $\tilde{\Phi} = I \otimes \tilde{\phi} \in \mathbb{C}^{GK \times NK}$ is the joint phase matrix, and $\tilde{H} = [(\tilde{H}^1)^T, \ldots, (\tilde{H}^K)^T]^T \in \mathbb{C}^{NK \times M}$ is the joint cascaded channel.

Given the above sparsity structure of the cascaded VAD channel $\tilde{H}$, its estimation from the observation matrix $\hat{Y}$ in (13), becomes a compressed sensing problem [1]. The existing algorithms in [4], [5], [6], [7], [8], [9], and [10] have solved it using OMP and optimization-based algorithms. These works, however, either fail to exploit the complete sparsity of the cascaded VAD channel, or require extra parameters to be known a priori. The Bayesian learning algorithm in [13] also fails to capture all the above-mentioned sparsities well. Further, the design in [14] is not scalable for a large number of RIS elements, and BS antennas. To address these issues related to the (i) requirement of VAD channel sparsity knowledge in [9] and [10], (ii) lack of exploitation of the complete VAD channel sparsity in [14]; and (iii) computational complexity in [13], we propose two variational expectation maximization (VEM)-based sparse channel estimation algorithms.

III. REVIEW OF SPARSE BAYESIAN LEARNING (SBL) AND EXISTING PRIORS

For the system model in (13), the Bayesian learning framework treats the received pilot matrix $\hat{Y}$, and the cascaded VAD channel $\tilde{H}$ as the observation and unknown variables, respectively [12]. The framework also assigns a prior distribution $p(H|\theta_p)$ to $\tilde{H}$, with the hyperparameter set $\theta_p$. The prior incorporates a belief about the cascaded VAD channel $\tilde{H}$, which herein is its sparsity structure. The Bayesian framework uses the likelihood distribution $p(Y|H)$, along with the prior, to infer the posterior distribution of $\tilde{H}$ by using the Bayes rule as $p(Y|H, \theta_p) \propto p(Y|H)p(H|\theta_p)$. The choice of prior is critical in the Bayesian learning framework. Recent literature in [12], [13], [14], and [25] has considered different kinds of sparsity-promoting priors. We now discuss these works, and the limitations of priors used therein, in capturing the structured sparsity of $\tilde{H}$.

1) SBL: The authors in [12] recovered a sparse vector $\hat{h}_m$ from the observation $\hat{y}_m = \Phi \tilde{h}_m + \tilde{n}_m$, a problem which can be obtained by setting $M = 1$ in (13). Their design assigned a zero-mean Gaussian prior to $\tilde{h}_m$ i.e., $p(\tilde{h}_m|\alpha) = \prod_{n=1}^{NK} \mathcal{CN}(\tilde{h}_{nm}|0, \alpha_n^{-1})$, with $\alpha = [\alpha_1, \ldots, \alpha_{NK}]^T \in \mathbb{R}_{+}^{NK \times 1}$ being the precision (inverse of variance) hyperparameter. When $\alpha_n$ is high, the unknown variable $\tilde{h}_{nm}$, with a very high probability, takes a value close to its mean, which is zero. The zero-mean Gaussian prior, with independent entries, thus only captures the UE-specific sparsity of the cascaded VAD channel $\tilde{H}$, but not its common-column and partially-common row sparsity.

2) M-SBL: The authors in [25] proposed multiple SBL (M-SBL) algorithm for multiple measurements, similar to the system model in (13), where the channel $\tilde{H}$, which needs to be recovered, is row sparse. This design captured the rows sparsity by using the prior: $p(H|\alpha) = \prod_{m=1}^{M} \prod_{n=1}^{NK} \mathcal{CN}(\tilde{h}_{nm}|0, \alpha_n^{-1})$. It captures row sparsity by assigning the same precision $\alpha_n$ for all $m = 1, \ldots, M$ entries in a row. This prior, however, is not applicable for the current system, as the channel $\tilde{H}$ herein is column sparse, and not row sparse. If used in the current system, it will wrongly model the complete VAD channel row, as either all zeros or all non-zeros.

3) AMP: Ruan et al. in [13] proposed an AMP algorithm to estimate a sparse RIS channel $\tilde{H}$. They considered Bernoulli-Gaussian prior:

$$p(\tilde{h}_{nm}|\alpha_{nm}) = (1 - \gamma_{nm})\delta(\tilde{h}_{nm}) + \gamma_{nm}\mathcal{CN}(\tilde{h}_{nm}|0, \alpha_{nm}^{-1}), \quad (14)$$

where $\delta(\tilde{h}_{nm})$ is the Dirac-delta function. For the Bernoulli-Gaussian prior in (14), the entry $\tilde{h}_{nm}$ takes a: i) zero value with a probability $(1 - \gamma_{nm})$; and ii) non-zero value from the distribution $\mathcal{CN}(\tilde{h}_{nm}|0, \alpha_{nm}^{-1})$ with a probability $\gamma_{nm}$. Since this prior assumes independence among the elements of $\tilde{H}$, it does not capture its shared sparsity. This prior also fails to capture the common-column and partially-common row sparsity of $\tilde{H}$. 

Fig. 1. (a) Scattering environment in RIS-assisted mmWave MIMO system; and (b) Sparsity structure in the virtual angular domain (VAD) cascaded UE—RIS—BS channel.
4) **CP-SBL:** We considered a coupled prior algorithm in [26] for joint active UE detection and channel estimation in massive MIMO systems with massive connectivity. We assumed a generalized prior on the channel $\mathbf{H}$, whose $(n,m)$th element is given as

$$p(\tilde{h}_{nm}|\alpha_n, \beta_{nm}) = \mathcal{CN}\left(\tilde{h}_{nm}|0, \left(\sum_{b=1}^{M} \beta_{n,m,b}\alpha_{n,b}\right)^{-1}\right)$$

The non-negative scalar $\alpha_{nm}$ is the precision hyperparameter for $\tilde{h}_{nm}$. The scalar $\beta_{n,m,b}$ is the shared weight hyperparameter, which denotes the relevance given to $\alpha_{n,b}$ to calculate the variance of $\tilde{h}_{nm}$. The column vector $\alpha_n = [\alpha_{n,1}, \alpha_{n,2}, \ldots, \alpha_{n,M}]^T \in \mathbb{R}_+^{1 \times M}$, and the row vector $\beta_{nm}^T = [\beta_{n,m,1}, \beta_{n,m,2}, \ldots, \beta_{n,m,M}] \in [0,1]^{1 \times M}$ thus denote the precision and shared weight vector for the $n$th column, respectively. In the prior in (15), the precision of $\tilde{h}_{nm}$ is the weighted sum of the precisions of the elements in $n$th row, i.e., $\alpha_{nb}$ for all $b = 1, \ldots, M$. It does not consider the precision of elements of columns of other rows $n' \neq n$. The prior in [26], thus, only captures the shared sparsity within the row, and not across the columns. It, thus, captures the UE-specific sparsity, and models the row-sparsity, which does not exist in the cascaded VAD channel $\tilde{H}$.

5) **AG-SBL:** The authors in [14] proposed an adaptive grouping SBL (AG-SBL)-based channel estimation for an RIS-assisted OFDM system. They assumed that the cascaded channel is sparse in time domain, and also that this sparsity is shared across group of RIS elements, called as sub-surfaces. They exploited common sparsity among different sub-surfaces, and clustered them by using a Dirichlet process (DP)-based clustering. The DP-distribution-based prior is:

$$p(\mathbf{h}_m|\mathbf{z}_m, \mathbf{\alpha}_m^1, \ldots, \mathbf{\alpha}_m^L) = \prod_{i=1}^{L} \left\{ \mathcal{CN}(\mathbf{h}_m|0, \text{diag}(\mathbf{\alpha}_i^{-1})) \right\}_{\{z_{i,m}=1\}}$$

Here, $L$ is the number of clusters and $z_{i,m}$ is a multinomial variable. This prior, when applied to the cascaded VAD channel $\tilde{H}$, captures its the sparsity shared across the rows. It fails to fully capture the common-column sparsity. The number of clusters $L$ also needs to be fixed a priori.

The aforementioned priors capture either UE-specific sparsity [12], [13], common row sparsity [25] or partially-common row sparsity [14], [26]. None of them captures the common-column sparsity along with UE-specific, and partially-common row sparsity.

IV. PROPOSED COLUMN-WISE COUPLED PRIOR

We now propose a column-wise coupled prior which captures the aforementioned sparsities of the cascaded VAD channel $\tilde{H} = [(\mathbf{H}^1)^T, \ldots, (\mathbf{H}^K)^T]^T$. This sparsity is summarized as follows:

- In each non-zero column of $\tilde{H}$, there are $Q^{RK}$ non-zero rows. Out of which $Q^R$, with $Q^R \leq Q^{RK}$, non-zero row entries in each column, are common to all the UEs. Rest $Q^{RK} - Q^R$ non-zero row entries are UE-specific.

To capture these sparsities, we assume a column-wise coupled prior over $\tilde{H}$, wherein its each element, denoted as $\tilde{h}_{nm}$, is generated from the following prior:

$$p(\tilde{h}_{nm}|\mathbf{\alpha}_m, \mathbf{s}_m) = \mathcal{CN}\left(\tilde{h}_{nm}|0, (s_m\alpha_m^{NZ} + (1-s_m)\alpha_m^{Z})^{-1}\right)$$

The scalar $s_m$ denotes the support of $\tilde{h}_{nm}$, which satisfies $s_m = \mathbb{I}(\tilde{h}_{nm} \neq 0)$. The non-negative scalars $\alpha_m^{NZ}$ and $\alpha_m^{Z}$ represent the precision parameter of $\tilde{h}_{nm}$, when it is non-zero and zero, respectively. The scalar $\gamma_m$ is the variance of $\tilde{h}_{nm}$, which is defined as follows:

$$\gamma_m \triangleq (s_m\alpha_m^{NZ} + (1-s_m)\alpha_m^{Z})^{-1}.$$ (18)

For the $m$th column of the joint concatenated channel $\tilde{H}$, the column vector: i) $\mathbf{\alpha}_m = [\alpha_m^{NZ}, \alpha_m^{Z}]^T \in \mathbb{R}_+^{2\times 1}$ denotes its prior precision parameter; and ii) $\mathbf{s}_m = [s_{m1}, s_{m2}, \ldots, s_{MN}]^T \in \{0,1\}^{NK\times 1}$ is its support vector. The prior distribution for $\tilde{H}$ is thus given as $p(\tilde{H}|\mathbf{A}, \mathbf{S}) = 
\prod_{n=1}^{NK} \prod_{m=1}^{M} p(\tilde{h}_{nm}|\mathbf{\alpha}_m, \mathbf{s}_m) = \prod_{m=1}^{M} \mathcal{CN}(\mathbf{h}_m|0_{NK}, \mathbf{\Sigma}_m).$ (19)

The matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M] \in \mathbb{R}_+^{2\times M}$ is the precision parameter matrix, and $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N] \in \{0,1\}^{NK\times 1}$ is the binary support matrix. The diagonal matrix $\mathbf{\Sigma}_m = \text{diag}(\gamma_m) \in \mathbb{R}_+^{NK\times NK}$ denotes the covariance matrix of the $m$th column of $\tilde{H}$, i.e., $\tilde{h}_m$. The $n$th entry of the vector $\gamma_m \in \mathbb{R}_+^{NK\times 1}$, i.e., $\gamma_{nm}$, is the variance of $\tilde{h}_{nm}$, which is defined in (18).

A. Discussion on How Prior Parameters Capture the Structured Sparsity

We note from (17) that if the channel entry $\tilde{h}_{nm}$

- is non-zero, the binary support $s_m = 1$. The channel entry $\tilde{h}_{nm}$ is, thus, drawn from the Gaussian distribution $\mathcal{CN}(0, (\alpha_m^{NZ})^{-1})$.
- is zero, $s_m = 0$. The entry $\tilde{h}_{nm}$ is drawn from the Gaussian distribution $\mathcal{CN}(0, (\alpha_m^{Z})^{-1})$.

The non-zero and zero entries are calculated using the parameter $s_m$, which is the support of $\tilde{h}_{nm}$. Also, since the hyperparameter $\alpha_m^{NZ}$ corresponds to the case when the channel is non-zero $\tilde{h}_{nm} \neq 0$, it takes a value close to zero. Similarly, since the hyperparameter $\alpha_m^{Z}$ corresponds to the case when $\tilde{h}_{nm} = 0$, it takes a high value. The prior proposed in (19) captures the

1) **common-column sparsity** by having a shared hyperparameters $\alpha_m^{NZ}, \alpha_m^{Z}$, which are independent for the column entries. Now, when the $n$th column is completely zero i.e., $\tilde{h}_m = 0_M$, each of its entry $\tilde{h}_{nm}$, for all $n'$ rows,

Note that for all $n$ denote $n = 1, \ldots, NK$ rows and for all $m$ denotes $m = 1, \ldots, M$ antennas. We avoid repeatedly mentioning it for the sake of brevity.
is drawn from $\mathcal{CN}(0, \alpha_m^Z)$, where $\alpha_m^Z$ takes a very high value. This makes all the entries $h_{nm}$ in the $m$th column go to zero.

2) partially-common row and UE-specific sparsity by assigning the same distribution: i) $\mathcal{CN}(0, \alpha_m^{NZ})$ to all non-zero entries in the $m$th column; and ii) $\mathcal{CN}(0, \alpha_m^Z)$ to all zero entries in the $m$th column.

The proposed prior does not distinguish between the partially-common sparsity across different UEs, and the UE-specific sparsity. It, instead, captures the shared sparsity in the complete column $h_m$, which includes both these sparsities. The prior is thus generalized, and can capture any kind of shared sparsity in the non-zero channel columns. We next show using an example as how the proposed prior, which is governed by the support $s_m$ and precision hyperparameter $\Lambda$, captures the above three sparsities in $\hat{H}$, which are also shown in Fig. 1. The BS, in this figure, has $M = 5$ antennas, which serves $K = 2$ UEs, via an RIS with $N = 5$ elements. The joint cascaded VAD channel of the two UEs is thus $\hat{H} = [\hat{H}_1^T, \hat{H}_2^T]^T \in \mathbb{C}^{10 \times 5}$.

- We see from Fig. 1b that the channel of UE1, $\hat{H}_1^1$, has two non-zero columns corresponding to Scatterers $p = 1$ and $p = 2$ between the BS and RIS. Let us consider the fourth column of $\hat{H}_1^1$, in which the first two entries are non-zero, and the rest are zero. This will be reflected in the corresponding support vector elements $s_{14}$ and $s_{24}$, which will take non-zero values. The remaining support vector elements $s_{34}, s_{44}$ and $s_{54}$ will take zero values. Thus $h_{14}$ and $h_{24}$ are drawn from the same prior distribution $\mathcal{CN}(0, \alpha_4^{NZ})$. The same argument applies to the zero elements. The shared sparsity in the zero and non-zero elements for each UE is, thus, captured by assigning the prior $\mathcal{CN}(0, \alpha_4^{NZ})$ to the non-zero elements, and $\mathcal{CN}(0, \alpha_3^Z)$ to the zero elements. The UE-specific sparsity is, hence, modelled based on the value of the support vector.

- To see how the proposed prior captures the shared sparsity across different UEs, we see the fourth column of the cascaded VAD channel $\hat{H}$ in Fig. 1b. The support elements for the 4th column are, $s_{14} = s_{24} = s_{64} = s_{84} = 1$ and $s_{54} = 0$ for $b \in \{3, 4, 5, 7, 9, 10\}$. The precision of non-zero and zero elements is $\alpha_4^{NZ}$ and $\alpha_3^Z$, respectively. The non-zero elements of the column $h_m$ are, thus, drawn from the prior $\mathcal{CN}(0, \alpha_4^{NZ})$, and zero elements are drawn from $\mathcal{CN}(0, \alpha_3^Z)$. Since the precision values are not UE-specific and only depend on the support, the proposed prior does not distinguish between the non-zero sparse channel entries of different UEs. This helps in modelling the common-column and partially-common row sparsities. The proposed prior, thus, captures the doubly-structured and the UE-specific sparsity of $\hat{H}$.

The parameters of the proposed prior, i.e., the precision matrix $\Lambda$ and the binary support matrix $S = [s_1, \ldots, s_M] \in \{0, 1\}^{NK \times M}$, need to be estimated. To facilitate estimation of $\Lambda$, similar to [12], we assign a Gamma hyperprior distribution over its elements: $p(\Lambda) = \prod_{i \in \{NZ, Z\}} M \prod_{m=1} M p(\alpha_m^i) = \prod_{i \in \{NZ, Z\}} M \prod_{m=1} M \Gamma(c)^{-1} d^c(\alpha_m^i)^{-1} e^{-\alpha_m^i}$.

$$p(\Lambda) = \prod_{i \in \{NZ, Z\}} M \prod_{m=1} M \Gamma(c)^{-1} d^c(\alpha_m^i)^{-1} e^{-\alpha_m^i},$$

Here, $\Gamma(c) = \int_0^\infty t^{c-1} e^{-t} dt$ is the Gamma function. The Gamma distribution is a suitable hyperprior for the positive quantity, precision, and makes the Bayesian inference tractable [12]. With $p(\Lambda|S)$ from (19), and $p(\Lambda)$ from (20), the prior $p(H|\Lambda)$ on $\hat{H}$, which is calculated as $p(H|\Lambda) = \int p(H|\Lambda, S)p(\Lambda)d\Lambda$, has a Student-t distribution [12]. This distribution, due to its sharp peak at zero, has a better sparsity-encouraging properties than the Gaussian prior [12].

Using the proposed coupled prior and gamma hyperprior, we develop an iterative variational expectation maximization (VEM) algorithm for estimating the cascaded VAD channel $\hat{H}$ [18]. The VEM approach incorporates variational inference (VI) in the EM algorithm, which is explained next. Before doing that, we summarize the notations used herein in Table I.

V. VARIATIONAL EXPECTATION MAXIMIZATION APPROACH WITH COUPLED PRIOR

The VI approach estimates the posterior distribution over all unknowns in the model [18]. It helps in dealing with intractable posteriors, and casts the inference as an optimization problem [17]. The idea is to consider a simple family of distributions over the unknown variable, and find the member of the family that is closest to the true posterior distribution, in terms of the KL divergence [18]. For the prior $p(H|\Lambda)$ defined in (19), the unknown variable is $\hat{H}$ and the hyperparameter is the prior precision $\Lambda$. The prior also depends on the support matrix $S$, which is assumed to be known here, for the sake of brevity. It is also estimated later in the sequel.

Let us denote the model hyperparameter as $\theta_p = \Lambda$. The VI method infers the posterior distribution over all the unknowns including the hyperparameters, i.e., $p(\hat{H}, \theta_p|\hat{Y})$. This inference problem is difficult to solve due to the complex posterior distribution [17]. We simplify it using the VEM framework which calculates the posterior distribution over the unknown variable $\hat{H}$, and point estimate of the hyperparameter set $\theta_p$. The point estimation of $\theta_p$, suffices, as it is common to the entire model, and the uncertainty in its estimate is quite low [18]. The VEM-algorithm is a two-step iterative procedure
based on the following identity [18]:

\[ p(\tilde{Y}|\theta_p) = \mathcal{L}(q, \theta_p) + \text{KL}(q||p), \]

where,

\[ \mathcal{L}(q, \theta_p) = \int q(\tilde{H}; \theta_q) \ln p(\tilde{Y}, \tilde{H} | \theta_p) d\tilde{H}, \]

\[ \text{KL}(q||p) = -\int q(\tilde{H}; \theta_q) \ln \frac{p(\tilde{H}, \tilde{Y}, \theta_p)}{q(\tilde{H}; \theta_q)} d\tilde{H}. \]

The notation \( q(\tilde{H}; \theta_q) \) denotes the approximating distribution of \( p(\tilde{Y}, \tilde{H}, \theta_p) \), where \( \theta_q \) denotes its variational parameter set [18]. For example, if \( q(\tilde{H}; \theta_q) \) is a Gaussian distribution, then \( \theta_q \) will be the set containing its mean and covariance matrix. The term \( \text{KL}(q||p) \) is the KL-divergence between the approximating distribution \( q(\tilde{H}; \theta_q) \) and the true posterior distribution \( p(\tilde{H}, \tilde{Y}, \theta_p) \). The term \( \mathcal{L}(q, \theta_p) \) is the evidence lower bound (ELBO).

The VEM algorithm solves an optimization over a class of tractable distributions \( q \approx p(\tilde{H}, \tilde{Y}, \theta_p) \), that is most similar to \( p \). The E-step of the VEM algorithm, for a fixed hyperparameter \( \theta_q \), approximates \( q(\tilde{H}; \theta_q) \) by maximizing the ELBO \( \mathcal{L}(q, \theta_p) \). To perform this maximization, a particular form of distribution \( q(\tilde{H}; \theta_q) \) must be assumed. This ensures that its inference can be reduced to the optimization of the variational parameter set \( \theta_q \) [18]. The ELBO \( \mathcal{L}(q, \theta_p) \), thus, becomes a function of these parameters, and is maximized with respect to (i) \( \theta_q \) in the E-step; and (ii) \( \theta_p \) in the M-step [18]. The M-step of the VEM algorithm is shown to be equivalent to maximizing the expected complete data log-likelihood (CLL) of the model, which is defined as \( p(\tilde{Y}, \tilde{H} | \theta_p)q(\tilde{H}; \theta_q) \). For example, if \( q(\tilde{H}; \theta_q) \) is a Gaussian distribution, the ELBO \( \mathcal{L}(q, \theta_p) \) can be written as:

\[
\mathcal{L}(q, \theta_p) = \mathbb{E}[\ln p(\tilde{Y}, \tilde{H} | \theta_p)] - \mathbb{E}[\ln q(\tilde{H}; \theta_q)]
\]

\[
= \mathbb{E}[\ln p(\tilde{Y}, \tilde{H} | \theta_p)] - \mathbb{E}[\ln q(\tilde{H}; \theta_q)]
\]

\[
= \mathbb{E}[\ln p(\tilde{Y}, \tilde{H} | \theta_p)] - \mathbb{E}[\ln q(\tilde{Y}; \theta_q)] - \mathbb{E}[\ln q(\tilde{H}; \theta_q)]
\]

\[
= -\mathbb{E}[\ln q(\tilde{Y}; \theta_q)] - \mathbb{E}[\ln q(\tilde{H}; \theta_q)]
\]

The E-step in (24) solves an optimization for the variational parameter set \( \theta_q \). Based on the structure of \( q(\tilde{H}; \theta_q) \), we next propose two algorithms. The first structured mean field-VEV (SMF-VEV) algorithm assumes a structured form of the approximating distribution \( q(\tilde{H}; \theta_q) \), and optimizes the ELBO to calculate update for the variational parameter set \( \theta_q \). The structured form of \( q(\tilde{H}; \theta_q) \) belongs to a set of multivariate distributions [17]. The second low-complexity fast mean field-VEM (FMF-VEM) algorithm assumes a fully-factorized approximating distribution, and optimizes a bound on the ELBO to obtain faster iterative update for \( \theta_q \). In the FMF-VEM algorithm, the (i) fully-factorized form of \( q(\tilde{H}; \theta_q) \) results in scalable updates of \( \theta_q \); and (ii) bound on ELBO removes the coupling between the channel and the phase matrix, which reduces its complexity [17]. Before discussing the proposed SMF- and FMF-VEV algorithms, we first derive the ELBO expression in (22) for the proposed prior in (19).

A. ELBO Expression for the Proposed Coupled Prior

To calculate the ELBO expression, we note that the posterior distribution of the cascaded VAD channel \( \tilde{H} \), which needs to be approximated, has the following form:

\[
p(\tilde{H}|\tilde{Y}, A, S) = \prod_{m=1}^{M} p(\tilde{h}_m; \tilde{y}_m, A, S).
\]

The approximating distribution for \( p(\tilde{h}_m; \tilde{y}_m, \theta_q) \) is assumed as \( q(\tilde{h}_m; \theta_q) \). The ELBO from (22) is given as:

\[
\mathcal{L}(q(\tilde{h}_m; \theta_q), \theta_p) = \mathbb{E}[\ln p(\tilde{y}_m, \tilde{h}_m)] - \mathbb{E}[\ln q(\tilde{h}_m; \theta_q)]
\]

\[
= \mathbb{E}[\ln p(\tilde{y}_m, \tilde{h}_m)] + \mathbb{E}[\ln p(\tilde{h}_m)] - \mathbb{E}[\ln q(\tilde{h}_m; \theta_q)]
\]

\[
= -\mathbb{E}[\ln q(\tilde{h}_m; \theta_q)] - \mathbb{E}[\ln q(\tilde{y}_m; \theta_q)]
\]

Equality in (a) is obtained by replacing \( p(\tilde{y}_m, \tilde{h}_m) \) with \( p(\tilde{y}_m, \tilde{h}_m)p(\tilde{h}_m) \). Proportionality in (b) is obtained by substituting the likelihood \( p(\tilde{y}_m, \tilde{h}_m) = CN(\tilde{y}_m, \tilde{h}_m, \sigma^2 I_G) \) and prior from (17), and by considering only the terms which depend on the variational parameter set \( \theta_q \). The notations in equality (c) are defined as: \( \lambda(\theta_q) = \mathbb{E}[\ln q(\tilde{y}_m; \theta_q)] \) and \( H(\theta_q) = \mathbb{E}[\ln q(\tilde{h}_m; \theta_q)] \). The expectations are with respect to the approximating distribution \( q(\tilde{h}_m; \theta_q) \). We now develop the SMF-VEV algorithm based on the above ELBO expression.

B. Structured Mean Field Variational Expectation Maximization (SMF-VEV) Algorithm

We know from the discussion after (23) that calculating an approximate posterior distribution \( q \), (known as variational distribution) is equivalent to minimizing the ELBO with respect to \( q \) i.e., \( q^*(\tilde{H}; \theta_q) = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \theta_p) \). Here \( q(\tilde{H}; \theta_q) \) denotes the variational distribution of \( p(\tilde{H}, \tilde{Y}, \theta_p) \), with \( \theta_q \) being its variational parameter set [18]. The set \( \mathcal{Q} \) denotes a family of tractable distributions from which the variational distribution \( q^*(\tilde{H}; \theta_q) \) is chosen [27]. This maximization reduces to solving the following optimization [27]:

\[
\theta_q^{\text{new}} = \arg\max_{\theta_q \in \mathcal{Q}} \mathcal{L}(q(\tilde{H}; \theta_q), \theta_p^{\text{old}}). \tag{27}
\]

In the SMF-VEV algorithm, this set \( \mathcal{Q} \) is restricted to multivariate Gaussian distributions, i.e.,

\[
q(\tilde{h}_m; \theta_q) = CN(\tilde{h}_m | \mu_m, \Sigma_m). \tag{28}
\]

The variational parameter set \( \theta_q \), thus, reduces to the mean and covariance matrices of each vector \( h_m \in \mathcal{H} \), i.e., \( \theta_q = \{\mu_m, \Sigma_m, m\} \). Calculating the variational distribution \( q(\cdot) \)
is equivalent to solving (27) with respect to the variational parameter set \( \theta_q = \{ \mu_m, \Sigma_m, \forall m \} \).

The E-step of the proposed SMF-VEM algorithm calculates the variational parameters \( \mu_m \) and \( \Sigma_m \), by solving (27). Its M-step calculates the parameters \( \theta_p \), i.e., the precision matrix \( \Lambda \), by assuming known support matrix \( S \). The matrix \( S \) is next detected using the multiple log-likelihood ratio test (LLRT) \([26],[28]\), by assuming known variational parameters \( \mu_m \) and \( \Sigma_m \), and the precision matrix \( \Lambda \).

1) E-Step of the SMF-VEM Algorithm: The E-step optimizes ELBO with respect to \( \mu_m \) and \( \Sigma_m \), by assuming a constant \( \theta_p \). The following lemma provides the ELBO expression for the distribution in (28). For notational simplicity, we omit \( \theta_p \) in E-step discussion as it is fixed here.

**Lemma 1:** The ELBO expression for the structured distribution in (28) is given as follows:

\[
\mathcal{L}_S[\mu_m, \Sigma_m] = - \frac{1}{\sigma^2} \lambda(\mu_m, \Sigma_m) - \sum_{n=1}^{NK} \beta_n(\mu_m, \Sigma_m)(\gamma_{nm})^{-1} + H(\Sigma_m).
\]

Here \( \lambda(\mu_m, \Sigma_m) = \|\hat{y}_m - \Phi(\mu_m)\|^2 + \text{tr}(\Phi^H \Phi \Sigma_m) \), \( \beta_n(\mu_m, \Sigma_m) = \|\mu_m\|^2 + \gamma_{nm} \), where \( \gamma_{nm} \) is the \( (n,m) \)th entry of \( \Sigma_m \), and \( H(\Sigma_m) = \ln \det(\Sigma_m) \). The lemma is proved in Appendix B. By using first order optimality on the ELBO expression in (29) with respect to the variational parameters \( \Sigma_m \) and \( \mu_m \), we get the following updates:

\[
\Sigma_m = \left( \sigma^{-2} \Phi^H \Phi + \text{diag}(\gamma_{nm})^{-1} \right)^{-1}
\]

and

\[
\mu_m = \sigma^{-2} \Sigma_m \Phi^H \hat{y}_m.
\]

2) M-Step of the SMF-VEM Algorithm for the Estimation of Precision Matrix \( \Lambda \): The M-step of the SMF-VEM algorithm estimates the precision matrix \( \Lambda \) by assuming a constant support matrix \( S \). As shown in (24), the hyperparameter \( \theta_p = \Lambda \) is estimated by maximizing the expected CLL in (24). The CLL for \( \theta_p = \Lambda \) takes the form \[ \log p(\hat{Y}, \hat{H}|A, S) + \log p(A) \] = \log p(A, \hat{Y}, \hat{H}|S). The M step thus reduces to the following maximization problem: \( \Lambda = \arg \max_{\Lambda} \{ \log p(A, \hat{Y}, \hat{H}|S) \} \). The expectation is with respect to the approximating variational distribution \( q \), calculated in the E-step in (28). The above maximization can next be simplified as follows:

\[
\max_A \{ \log p(A, \hat{Y}, \hat{H}|S) \} = \max_A \{ \log p(A) + \log p(\hat{Y}|\hat{H}) + \log p(\hat{H}|A, S) \}
\]

\[
= \max_A \sum_{i \in \{N, Z\}} \sum_{m=1}^{M} \left( c \log \alpha_m^i - d \alpha_m^i \right) + \sum_{n=1}^{NK} \sum_{m=1}^{M} \log(\gamma_{nm})^{-1} - (\gamma_{nm}^{-1})(|\hat{h}_{nm}|^2)q.
\]

Equality (a) is obtained by i) ignoring the distribution \( p(\hat{Y}|\hat{H}) \) as it is independent of \( A \); and ii) using \( p(A) \) from (20) and \( p(\hat{Y}|\hat{H}) = \prod_m p(\hat{y}_m|\hat{h}_m) \) from the discussion after (26).

It appears that the optimization in (32) cannot be decoupled into \( 1 \times M \) problems, for each hyperparameter \( \alpha_m^i \). This is because \( \alpha_m^i, \forall m \), are coupled together in the logarithm term \( \log(\gamma_{nm})^{-1} \). However, the binary nature of the parameter \( s_{nm} \) makes it possible to decouple the optimization in (32) for \( \alpha_m^{NZ} \) and \( \alpha_m^Z \). This is because when \( s_{nm} = 1 \), the prior variance in (18) becomes \( \gamma_{nm} = \alpha_m^{NZ} \); and when \( s_{nm} = 0 \), the variance becomes \( \gamma_{nm} = \alpha_m^Z \mbox{.} \) The following theorem uses the properties of the proposed coupled prior, and the support matrix \( S \) to derive the closed-form update of the precision hyperparameters \( \alpha_m^{NZ} \) and \( \alpha_m^Z \). The theorem is proved in Appendix C.

**Theorem 1:** The optimal value of \( \alpha_m^{NZ} \) and \( \alpha_m^Z \) parameters of matrix \( \Lambda \) is

\[
(\alpha_m^{NZ})^* = \frac{c + \sum_{n=1}^{NK} s_{nm} + d}{\sum_{n=1}^{NK} s_{nm} e_{nm} + d},
\]

\[
(\alpha_m^Z)^* = \frac{c + \sum_{n=1}^{NK} (1 - s_{nm})}{\sum_{n=1}^{NK} (1 - s_{nm}) e_{nm} + d}.
\]

The scalar \( e_{nm} \) is defined as \( e_{nm} = |\mu_m|^2 + \gamma_{nm} \), where \( \mu_m \) is the \( n \)th entry of the mean vector \( \mu_m \), and \( \gamma_{nm} \) is the \( (n,m) \)th entry of the covariance matrix \( \Sigma_m \).

3) Detection of Binary Support Matrix \( S \): For the prior in (17), the prior precision of elements of \( \hat{h}_m \) depends on its support \( s_m \), whose \( (n,m) \)th entry \( s_{nm} \) satisfies \( s_{nm} = \mathbb{I}(\hat{h}_{nm} \neq 0) \). This support is detected with a high accuracy using the index-wise LLRT as follows \([26],[28]\):

\[
\tilde{s}_{nm} = \frac{|\tilde{\phi}_m^H(\sigma^2 I_G + \Phi(H_{nm})^{H})^{-1}y_m|}{\tilde{\phi}_m^H(\sigma^2 I_G + \Phi(H_{nm})^{H})^{-1}\tilde{\phi}_m} \geq \tilde{\epsilon}.
\]

The term \( \tilde{\phi}_m^H \in \mathbb{C}^{1 \times M} \) is the \( m \)th column of the phase matrix \( \tilde{\phi}_m \), and the term \( \gamma_{nm} = [\gamma_1, \ldots, \gamma_{N-1}, 0, \gamma_{N+1}, \ldots, \gamma_{NK}]^T \). For a desired probability of false alarm \( \epsilon \in [0,1] \), i.e., the probability that an element is zero but declared non-zero, the threshold \( \tilde{\epsilon} = \left( Q^{-1}(\epsilon) \right)^2 \) \([28]\), where \( Q \) is the standard Q-function.

The binary support matrix \( S \in \{0,1\}^{NK \times M} \) is estimated using the index-wise LLRTs in (35) as: \( \tilde{s}_{nm} = 1 \) if (35) is true, and 0 otherwise. The estimate is denoted as \( \tilde{S} \), with its \( (n,m) \)th entry being \( \tilde{s}_{nm} \).

The proposed SMF-VEM algorithm is summarized in Algorithm 1. Its Step 2 performs the E-step to update the posterior parameters of \( \hat{h}_m \), i.e., covariance matrix \( \Sigma_m \) and mean \( \mu_m \) using (30) and (31), respectively. Step 3 and Step 4 perform the M-step. Step 3 updates the precision parameters \( \alpha_m^{NZ} \) and \( \alpha_m^Z \) for all \( m \) antennas, using (33) and (34), respectively. Step 4 updates the binary support matrix \( S \) using multiple LLRTs in (35). The channel estimate is given as the mean of the estimated posterior distribution, i.e., \( \hat{H} = [\mu_1, \ldots, \mu_M]^T \).

C. Fast Mean Field Variational Expectation Maximization (FMF-VEM) Algorithm

The above SMF-VEM algorithm inverts a full covariance matrix of size \( NK \times NK \) in (30) in its E-step, with a
time complexity of $O(N^3K^3)$. For an RIS-aided wireless system, with a large number of RIS elements $N$, this algorithm will have an extremely high computational complexity. The author in [17] used a variational mean-field approximation to reduce the computational complexity of the conventional SBL algorithm [12]. Their design did not consider any structured sparsity in the unknown sparse vector, and estimated it by using only a simple univariate Gaussian prior. The proposed coupled prior, which captures a variety of sparsity structures, is very different from the simple univariate Gaussian prior used in [17]. This makes the subsequent steps of developing a low-complexity VEM algorithm very different from [17].

To reduce the complexity of E-step of SMF-VEM algorithm, we now develop its faster version, referred to as the fast mean field VEM (FMF-VEM) algorithm. The reduced complexity is later shown to be achieved without sacrificing the channel estimation accuracy. The approximating variational distribution $q(\mathbf{H})$ in the FMF-VEM algorithm is assumed to be fully-factorized. This is unlike the earlier SMF approximation, which assumed a multivariate approximating distribution over each channel column $\tilde{h}_m$. This assumption inhibited a fully-factorized distribution. The FMF-based approximating posterior distribution, thus, takes the following form:

$$q(\mathbf{H}) = \prod_{n=1}^{NK} CN(\hat{h}_{nm}|\mu_{nm}, \tau_{nm}) = CN(\hat{h}_{nm}|\mu_{nm}, \text{diag}(\tau_{nm})).$$

(36)

The vectors $\mu_{nm} \in \mathbb{C}^{NK \times 1}$ and $\tau_{nm} \in \mathbb{C}^{NK \times 1}$ respectively, are the mean and variance of the factorized distribution. The above equation is different from (28) due to the diagonal nature of the covariance matrix, i.e., $\text{diag}(\tau_{nm})$. The E-step of the FMF-VEM algorithm updates $\mu_{m}$ and $\tau_{m}$. Its M-step of estimating the hyperparameter matrix $\Lambda$, and the LLRT step of detecting the support matrix $\mathbf{S}$, are same as that of the SMF-VEM algorithm. We now derive the E-step of the FMF-VEM algorithm, which updates the variational parameters $\mu_{m}$ and $\tau_{m}$ of the approximating distribution in (36).

1) E-Step of the FMF-VEM Algorithm: The factorized ELBO expression is next.

Lemma 2: The ELBO for the fully-factorized approximating posterior distribution of (36) is:

$$\mathcal{L}_F[\mu_{m}, \tau_{m}] = -\frac{1}{\sigma^2} \lambda(\mu_{m}, \tau_{m}) - \sum_{n=1}^{NK} \beta_n(\mu_{m}, \tau_{m})(\gamma_{nm})^{-1} + H(\tau_{m}).$$

(37)

Here $\lambda(\mu_{m}, \tau_{m}) = ||\tilde{y}_m - \Phi \mu_{m}||^2 + (\text{diag}(\Phi^H \Phi))^{-1} \tau_{m}$, $\beta_n(\mu_{m}, \tau_{m}) = ||\mu_{nm}||^2 + \tau_{nm}$, and $H(\tau_{m}) = \sum_{n=1}^{NK} \tau_{nm}$. The lemma is proved in Appendix B.

Maximizing the above ELBO with respect to the variational parameters $\mu_{nm}$ and $\tau_{m}$ completes the E-step. Recall that in the SMF-VEM algorithm, the coupling between the variational parameters occurred due to non-diagonal covariance matrix $\Sigma_m$, whose update in (30) has a computational complexity of $O(N^3K^3)$. The FMF ELBO obviates this coupling by assuming $\Sigma_m = \text{diag}(\tau_{m})$. It is worth noting that the term $\Phi \mu_{m} = \sum_{n=1}^{NK} \Phi_{nm} \mu_{nm}$ in the FMF ELBO in (37) still couples the variational parameter $\mu_{m}$ and the phase matrix. We refer to this coupling as residue coupling to make the discussion easy to comprehend. The update of $\mu_{nm}$, which involves computing $\Phi \mu_{m}$ in each iteration, has a complexity of $O(GN^2K^2)$.

We now reduce the complexity due to the residue coupling by bounding the term $||\tilde{y}_m - \Phi \mu_{m}||^2$ in the ELBO by using a minorizing function. The bound is constructed such that it does not include the term $\Phi \mu_{m}$, and is constructed using the Lipschitz inequality. The bound satisfies the desirable properties required for the minorization framework [19], which instead of maximizing the exact function, maximizes its lower bound. We then maximize this bound to calculate the variational parameters. Worley in [17] used a similar bound to reduce complexity. The complex-valued channel estimation in RIS-assisted mmWave systems, complex-valued coupled prior, and the channel sparsity structure, considered herein, makes our work very different from [17], which considered only real-valued vector estimation, and that too without any sparsity structure.

We next apply the following lemma to obtain a minorizing function for $-\lambda(\mu_{m}, \tau_{m}) = -||\tilde{y}_m - \Phi \mu_{m}||^2 - a^T \tau_{m}$ of (37), where $a = \text{diag}(\Phi^H \Phi)$.

Lemma 3: For any continuously differentiable function $h: \mathbb{C}^n \rightarrow \mathbb{C}$ with an $L_h$-Lipschitz gradient, following holds [17]:

$$h(x) \leq g(x; z) \triangleq h(z) + (x - z)^H \nabla_x h(z) + \frac{L}{2} ||x - z||^2.$$  

(38)

Here $L \geq L_h$ and are any $x, z \in \mathbb{C}^n$. Equality holds when $x = z$.

We apply this lemma by taking $h(\mu_{m}) = ||\tilde{y}_m - \Phi \mu_{m}||^2$. This leads to the following minorizing function for $\lambda(\mu_{m}, \tau_{m})$:

$$\Lambda(\mu_{m}, \tau_{m}; \delta) = ||\tilde{y}_m - \Phi \delta||^2 + 2(\mu_{m} - \delta)^H \Phi^H (\Phi \delta - \tilde{y}_m) + \frac{L}{2} ||\mu_{m} - \delta||^2 + a^T \tau_{m}.$$  

(39)
The notation $\Lambda(\mu_m, \tau_m)$ denotes the minorizing function at the auxiliary point $\delta$. By substituting this into ELBO of (26), we get the lower bound, denoted as $\mathcal{L}_F[\mu_m, \tau_m]$, for $\mathcal{L}_F[\mu_m, \tau_m]$: \[
abla^2 \mathcal{L}_F[\mu_m, \tau_m; \delta] = -\frac{1}{\sigma^2} \Lambda(\mu_m, \tau_m; \delta) - \sum_{n=1}^{NK} \beta_n(\mu_m, \tau_m) \\
\times (\gamma_{nm})^{-1} + H(\tau_m), \tag{40}
\]
By construction, the minorizing function $\mathcal{L}_F[\mu_m, \tau_m]$ lower bounds the FMF ELBO, i.e., $\mathcal{L}_F[\mu_m, \tau_m] \geq \mathcal{L}_F[\mu_m, \tau_m; \delta]$, for all $\delta \in \mathcal{C}^{NK \times 1}$. Also, it becomes equal to the ELBO at $\delta = \mu_m$, i.e., $\mathcal{L}_F[\mu_m, \tau_m] = \mathcal{L}_F[\mu_m, \tau_m; \mu_m]$. The value of auxiliary variable $\delta$ which makes the minorizing ELBO function $\mathcal{L}_F[\mu_m, \tau_m; \delta]$ descend in every iteration is $\delta = \mu_m^{old}$, where $\mu_m^{old}$ is the variational mean obtained in the previous VEM iteration. This is shown in Section V-E. The updates of $\tau_m$ and $\mu_m$ in the FMF-VEM algorithm, derived by applying the first order optimality conditions, are given as follows: \[
\tau_m^{(t+1)} = 1 \circ (\sigma^{-2} a + 1 \circ \gamma_m), \mu_m^{(t+1)} = \mathbf{D}(t) \zeta(t), \tag{41}
\]
where, \[
\zeta(t) = \frac{1}{\sigma^2} \left( \frac{L}{2} \mu_m^{(t)} - \Phi H( \Phi \mu_m^{(t)} - \bar{y}_m) \right), \tag{42}
\]
and \[
\mathbf{D}(t) = \left( \frac{L}{2\sigma^2} I + \Sigma_m^{-1} \right)^{-1}. \tag{43}
\]
The notation $\circ$ denotes the element-wise division, and the vector $1$ is $NK$-sized all ones vector. Note that we have substituted $\delta = \mu_m^{old}$ in (41) to update the variational mean $\mu_m$ at the $(t + 1)$th iteration. Here $\Sigma_m = \text{diag}(\gamma_m)$ is the variance matrix, and $a = \text{diag}(\Phi H \Phi)$. The constant $L$ is taken as twice the maximum eigenvalue of $\Phi H \Phi$ [17].

The FMF-VEM algorithm in its E step assumes that the approximating variational distribution $q(\bar{H})$ is fully-factorized, and then removes the coupling between $\Phi$ and $\mu_m$ in the fully-factorized ELBO by using a minorizing function, which is maximized to calculate the variational parameters. We see that due to these two facts, the update of (i) covariance $\text{diag}(\tau_m)$; and, (ii) mean $\mu_m$, given in (41) become very different from that of SMF-VEM algorithm, given in (31). The differences in variational updates of both the algorithms are shown in Table II. We also show in Section VI that the variational approximation, and the proposed bound used to update the variational parameters reduces the KL divergence between the variational distribution and the true posterior distribution to zero over the EM iterations. This result not only shows that the proposed lower bound is valid, but also that the approximation and the bound do not reduce the efficacy of the FMF-SBL algorithm.

2) M-Step of the FMF-VEM Algorithm: The M-step is same as that of SMF-VEM, which is given in (33)-(34).

3) Detection of Binary Support Matrix $S$: The update of $S$ is also same as that of SMF-VEM algorithm, which is given in (35).

The FMF-VEM algorithm is same as that of Algorithm 1, except its E step (Step 2). The E-step now updates the posterior parameters of $\hat{H}_m$ i.e., variance $\tau_m$ and mean $\mu_m$, according to (41). The channel estimate is given as: $\hat{H} = [\mu_1, \ldots, \mu_M]$.

D. Complexity Analysis of the SMF-VEM and FMF-VEM Algorithms

We now calculate the per-iteration time complexity of the proposed algorithms in Table IIIa and Table IIIb. The complexity of calculating $\Sigma_m$ in (30) is reduced from $O(N^3K^3)$ to $O(C^3)$, by using the Woodbury identity [12]: $\Sigma_m = \Phi_m - \Phi_m \Sigma_m^{-1} \Phi_H \Phi_m^{-1}$. The SMF-VEM algorithm in (41), updates the variance of each element independently. It updates the mean by inverting a diagonal matrix with $O(NK)$ complexity. The complexity of its E step is linear, as in AMP algorithm [13]. Also, as these updates do not invert a full matrix, they are faster than the SMF-VEM algorithm. The LLRT step of detecting the support $s_m$, however, is cubic in number of sub-phases. The existing methods in literature, such as using Hessian matrix-based threshold, sequential SBL [27], and SNR-dependent threshold [25] for pruning the close-to-zero entries, can be used to reduce this complexity. Integration of these methods with the proposed algorithms is non-trivial, and can be taken up as future extension of this work.

The OMP-based algorithms in [4], [6], [7], [9], and [10] and the AMP-based algorithm in [13] have lesser complexity than the proposed ones. The complexity of OMP-based algorithms varies linearly with the sparsity level, which is assumed to be known a priori. Acquisition of sparsity information has an extremely high computational complexity [1], which is usually not included while calculating complexity. The AMP-based algorithms requires additional testing site infrastructure, and have divergent behavior for not-so-large dimensions [29]. All these algorithms also fail to exploit the double-structured channel sparsity. The proposed FMF-VEM algorithm exploits the complete sparsity in cascaded channel without knowing the sparsity level, and provide locally optimal solution. Further, as shown later in Section VI, the AMP, and OMP-based algorithms provide a highly degraded performance than the proposed ones.

E. Convergence Analysis of the FMF-VEM Algorithm

We now show that the updates of FMF-VEM, in each iteration, maximize the minorizing function $\mathcal{L}_F$. This guarantees that the approximation in the E step of VEM, achieves a locally-optimal solution of the variational parameters. We prove this in the following lemma.

\textbf{Lemma 4:} The variational parameter updates \( \{ \mu_m^{(t)}, \tau_m^{(t)} \} \) for all \( t = 1, \ldots, T \) iterations, form a non-decreasing sequence of the ELBO objective function \( \mathcal{L}_F[\mu_m, \tau_m] \), i.e., \( \mathcal{L}_F[\mu_m^{(t+1)}, \tau_m^{(t+1)}] \geq \mathcal{L}_F[\mu_m^{(t)}, \tau_m^{(t)}] \). The \( \mu_m^{(t)}, \tau_m^{(t)} \) are obtained by maximizing \( \mathcal{L}_F[\mu_m, \tau_m] \) in the $t$th VEM iteration.

\textbf{Proof:} We have \( \mathcal{L}_F[\mu_m^{(t+1)}, \tau_m^{(t+1)}] \geq \mathcal{L}_F[\mu_m^{(t+1)}, \tau_m^{(t+1)}] \geq \mathcal{L}_F[\mu_m^{(t+1)}, \tau_m^{(t+1)}] \geq \mathcal{L}_F[\mu_m^{(t)}, \tau_m^{(t)}] \). Equality in (a) is because $\delta^{(t+1)} = \mu_m^{(t+1)}$, which is the equality condition for...
the Lipschitz equality. Inequality in (b) is due to (38). The difference between the LHS and RHS terms of the inequality (b) can be shown to be positive as follows:}

\[
\mathcal{L}_F[\mu^{(t+1)}_m, \tau^{(t+1)}_m; \mu^{(t+1)}_m, \mu^{(t)}_m] = -\sigma^2 \left( H_n - \mathbf{F} \mu^{(t+1)}_m \right) H_n^H (\mu^{(t)}_m - \hat{y}_m) + \frac{L}{2} \left\| \mu^{(t+1)}_m - \mu^{(t)}_m \right\|^2.
\]

We apply the Lipschitz inequality by taking, \( h(\mu^{(t+1)}_m) = \| \hat{y}_m - \mathbf{F} \mu^{(t+1)}_m \|^2 \) with auxiliary variable \( z = \mu^{(t)}_m \). This results in the first three terms inside the bracket being negative. The last term in the bracket is also always negative since the \( l_2 \) norm is always positive. The terms inside the bracket is thus negative. Hence, the difference, \( \mathcal{L}_F[\mu^{(t+1)}_m, \tau^{(t+1)}_m; \mu^{(t+1)}_m] - \mathcal{L}_F[\mu^{(t+1)}_m, \tau^{(t)}_m; \mu^{(t)}_m] \), is positive. Inequality in (d) is because \( \delta^{(t)} = \mu^{(t)}_m \), which again is the equality condition for the Lipschitz inequality. This proves that update equations obtained from optimizing the \( \mathcal{L}_F \), ascend on the factorized objective \( \mathcal{L}_F \).}

VI. NUMERICAL RESULTS

We investigate the performance of the proposed SMF-VE and FMF-VE algorithms by considering the following metrics: (i) NMSE, for channel estimation which is defined as \( \mathbb{E} \left( \| \mathbf{H} - \hat{\mathbf{H}} \|_F^2 / \| \mathbf{H} \|_F^2 \right) \); (ii) normalized support error rate (NSER), which measures the error in the estimated support and is defined as \( \mathbb{E}(\sum_{m} \sum_{n} I(s_{nm} \neq \hat{s}_{nm})/NMK) \); and (iii) KL divergence, which measures the difference between the true and the approximated posterior distributions. It is defined as, \( KL(p||q) = -\mathbb{E}_q \left[ \log \frac{p}{q} \right] \), where \( p \) is the true and \( q \) is the approximated posterior distribution; and (iv) sum-spectral efficiency (sum-SE). The BS precodes the UE data \( x \in \mathbb{C}^{K \times 1} \) by maximum ratio combining (MRC), which for the \( k \)th UE is given as \( y^k = \mathbf{H}^H \theta^k \in \mathbb{C}^{M \times 1} \). The BS detects the \( k \)th UE signal as \( y^k = (v^k)^H \mathbf{H}^H \theta^k + (v^k)^H \sum_{n \neq k} \mathbf{H}^H \theta^k + (v^k)^H \mathbf{w}^k \). The sum-SE (in bps/Hz) for the \( K \) UEs is defined as follows: \( \text{SE} = \left( 1 - \frac{T_p}{T_c} \right) \sum_{k=1}^{K} \log_2 \left( 1 + \frac{||v^k||^2 \mathbf{H}^H \theta^k}{\sum_{n \neq k} ||v^k||^2 \mathbf{H}^H \theta^k + \sigma^2} \right) \). Here, \( T_p \) and \( T_c \) denote the pilot length and coherence interval, respectively.

The simulation studies are performed by considering a BS with \( M = 64 \) antennas which serves \( K = 4 \) UEs via a UPA RIS with \( N = 128 \) elements. The number of paths between the i) BS and RIS is \( P^B = 20 \); and ii) RIS and each UE is \( Q^R = 40 \). We assume, similar to [9], that all spatial angles are on the quantized grids of size \( M = (M_1 \times M_2) \) on the BS side and \( N = (N_1 \times N_2) \) on the RIS side. The channel gains are generated as follows: \( |\alpha^B_k| = 10^{-3} d^{-2.2}_B \) and \( |\alpha^R_k| = 10^{-3} d^{-2.8}_R \). Here, the scalars \( d_B \) and \( d_R \) denote the distance between BS and RIS, and RIS and UE, respectively. Their values are fixed as \( d_B = 10m \) and \( d_R = 100m \) [30], respectively. The elements of RIS reflecting matrix \( \mathbf{F} \), similar to [9] and [30], are selected uniformly from \( \left\{ -\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}} \right\} \).

Further, for the proposed SMF-VE and FMF-VE algorithms, we take maximum number of EM iterations as \( T_{\text{max}} = 200 \), and the stopping threshold as \( \eta = 10^{-3} \). The prior variances are taken as \( \gamma^{(0)}_{nm} = 0.01 \) for all \( m \) and \( n \). The proposed algorithms performance is compared with: (i) DS-OMP [9]: uses OMP algorithm to estimate cascaded VAD channel \( \mathbf{H} \) by incorporating its double sparsity structure; ii) AG-SBL [14]: uses the Dirichlet-Process to cluster the RIS into smaller groups, and uses AG-SBL algorithm to estimate the channel; iii) EM-VAMP [7]: uses AMP algorithm, along with Bernoulli-Gaussian prior, to estimate the channel; and iv) Oracle MMSE: lower bounds the performance of all other algorithms as it knows the true support of \( \mathbf{H} \). Performs MMSE estimation for this over-determined problem.

A. NMSE Comparison

We show in Fig. 2a the NMSE of various algorithms by varying the SNR = 1/\( \sigma^2 \), and by fixing number of sub-phases as \( G = 100 \). The proposed SMF-VE and FMF-VE algorithms have a lower NMSE than other algorithms. This is because they exploit the doubly-structured and UE-specific sparsities of the VAD channel \( \mathbf{H} \). Their NMSE is also close to the Oracle MMSE. The proposed algorithms outperform the (i) DS-OMP algorithm, which does not exploit the UE-specific sparsity and is a greedy algorithm. The DS-OMP also requires the complete sparsity information
of $\mathbf{H}$, which is computed with a high complexity [4]; (ii) AG-SBL algorithm, which considers Dirichlet process based Gaussian prior over each element of the channel matrix, and ignores the sparsity shared within the elements of the column. It, hence, exploits the partially-common row sparsity, but not the common-column and UE-specific sparsities; and (iii) EM-VAMP algorithm, which assumes an independent Bernoulli-Gaussian prior. It completely ignores the structured channel sparsities.

We plot in Fig. 2b the NMSE by varying the number of sub-phases $G$. We see that the NMSE of all algorithms reduces with increase in $G$. This is due to increase in observations. The two proposed algorithms have the lowest NMSE for all $G$ values. This is again because, unlike DS-OMP, AG-SBL, and EM-VAMP, they exploit the complete sparsity structure of $\mathbf{H}$. Our algorithms also require a fewer number of sub-phases $G$ to provide an NMSE, which is similar to that of others. For example, they yield an NMSE of $-14$ dB, while EM-VAMP requires $G=108$. The DS-OMP algorithm, due to its greedy nature [11], and AG-SBL due to its wrong sparsity modeling require an even higher number of sub-phases.

We next plot in Fig. 2c the NMSE by varying $Q^{Rk}$ i.e., the partially common row sparsity. The $Q^{Rk}$, we recall, is the number of scatterers between the UE $k$ and RIS. We see that the NMSE increases with increase in $Q^{Rk}$. This is because the increase in $Q^{Rk}$ increases the sparsity level, which increases the number of unknowns to be estimated. The number of observations or sub-phases to estimate them remains constant. Since same number of sub-phases are used to estimate more unknowns ($Q^{Rk} \times P^D$), the performance of all algorithms degrades. The proposed algorithms still outperform the existing ones, since they exploit the structured sparsities well. The SNR for this study is fixed at 10 dB, and $G=150$.

We plot in Fig. 3a, the NMSE of DS-OMP and FMF-VEM algorithms by varying the number of sub-phases $G$, and the number of scatterers between UEs and RIS $Q^{Rk}$. We observe that for low $Q^{Rk}$ values, both algorithms have low NMSE values. This is because there are sufficient number of sub-phases, and exploiting just the doubly-structured sparsity is enough. But for high $Q^{Rk}$ values, the proposed FMF-VEM algorithm has a much lower NMSE than DS-OMP. This is because it exploits the UE-specific and doubly-structured sparsities, which DS-OMP does not. The FMF-VEM algorithm is, thus, more robust to the change in number of scatterers $Q^{Rk}$. It is, therefore, a better choice for RIS channel estimation in different scattering environments.

We next see from Fig. 3b that the NMSE of all algorithms decreases with increase in $M$. For a fixed number of scattering paths, channel sparsity remains constant $P^D = 10$ and $Q^{Rk} = 40$. Increasing $M$, increases the number of observations, and that too without increasing the sparsity, which reduces the NMSE. The proposed algorithms again have the lowest NMSE. This shows their efficacy, even for a large value of $M$, which is typical in mmWave systems [1].

B. KL-Divergence Comparison

We compare in Fig. 3 the KL-divergence to measure the distance between the posterior distribution of the FMF-VEM and SAVE [15] algorithms. The SAVE algorithm uses a fully-factorized posterior from (36), and does not use Lipschitz inequality of (38) to bound the factorized ELBO. This plot helps us in understanding how Lipschitz inequality affects FMF-VEM convergence. We see from Fig. 3c that SAVE algorithm converges slightly faster than FMF-VEM. The KL-divergence of both algorithms, however, becomes close to zero, with increase in iterations. We conclude that the FMF-VEM algorithm also converges to the SMF-VEM posterior, and Lipschitz inequality does not affect the final distribution, to which the proposed FMF-VEM algorithm converges. It, only, marginally affects the convergence rate.

C. SE Comparison

We compare in Fig. 4a the sum-SE of proposed algorithms with the existing ones, to show the practical performance gains realized by them. We plot the SE by varying the uplink power, and by fixing the number of sub-phases at $G=80$. The number of scatterers between the BS and the RIS is set as $P^D = 35$. The two proposed algorithms have a higher sum-SE than other algorithms. This is due to their lower NMSE. For high values of uplink power, the sum-SE of all algorithms saturates. This is due to high multi-UE interference. The Oracle MMSE has the highest SE because its NMSE, as shown in Fig. 2a, is the lowest.

D. NSER and Run-Time Comparison

We compare in Fig. 4b, the support of $\mathbf{H}$ estimated by the proposed algorithms against the true support at SNR= 15 dB. We compare in Fig. 3 the KL-divergence to measure the distance between the posterior distribution of the FMF-VEM and SAVE [15] algorithms. The SAVE algorithm uses a fully-factorized posterior from (36), and does not use Lipschitz inequality of (38) to bound the factorized ELBO. This plot helps us in understanding how Lipschitz inequality affects FMF-VEM convergence. We see from Fig. 3c that SAVE algorithm converges slightly faster than FMF-VEM. The KL-divergence of both algorithms, however, becomes close to zero, with increase in iterations. We conclude that the FMF-VEM algorithm also converges to the SMF-VEM posterior, and Lipschitz inequality does not affect the final distribution, to which the proposed FMF-VEM algorithm converges. It, only, marginally affects the convergence rate.

We compare in Fig. 4a the sum-SE of proposed algorithms with the existing ones, to show the practical performance gains realized by them. We plot the SE by varying the uplink power, and by fixing the number of sub-phases at $G=80$. The number of scatterers between the BS and the RIS is set as $P^D = 35$. The two proposed algorithms have a higher sum-SE than other algorithms. This is due to their lower NMSE. For high values of uplink power, the sum-SE of all algorithms saturates. This is due to high multi-UE interference. The Oracle MMSE has the highest SE because its NMSE, as shown in Fig. 2a, is the lowest.
We see that although the SMF-VEM algorithm converges faster, the convergence value is similar for both algorithms. The NSER is low for both of them, which shows that the support estimated by them is accurate. This validates the fact that Lipschitz inequality does not affect the value at which the FMF-VEM algorithm converges, it just affects the convergence rate. We finally see from Fig. 4c that the FMF-VEM has a much lower runtime than the SMF-VEM. The gap between the two algorithms increases with $G$. The DS-OMP algorithm requires much lesser time. This is expected as its complexity depends on the values of $Q_{Rk}$ and $P_{EI}$ (sparsity level) [9], which, unlike the proposed algorithms, it assumes to know. Its performance, however, as discussed earlier, is much inferior.

VII. CONCLUSION

We proposed variational expectation maximization (VEM) based algorithms for channel estimation in RIS-assisted mmWave MIMO systems. Both these algorithms use the column-wise coupled prior, which exploits the inherent sparsities of the cascaded channel. We analytically showed that the proposed fast mean field (FMF)-VEM algorithm has a much lower complexity than the proposed structured mean field (SMF)-VEM algorithm. We provided the convergence guarantees of the SMF-VEM algorithm. We numerically showed that the i) proposed algorithms outperform the existing state-of-the-art algorithms in terms of NMSE and SE; and ii) FMF-VEM algorithm has a much lower runtime than the SMF-VEM algorithm for different number of pilots.
We observe that due to the binary nature of the support element $s_{nm}$, the third term in the above equation reduces to $\sum_{n=1}^{NK} s_{nm}$. The final optimality condition can thus be written as:

$$\frac{c}{\alpha_m^{NZ}} - d + \sum_{n=1}^{NK} s_{nm} - \sum_{n=1}^{NK} s_{nm}E[\tilde{h}_{hm}] = 0.$$ (48)

We can similarly, calculate $(\alpha_m^{NZ})^*$ by using the first order optimality, and binary nature of support parameter $s_{nm}$. The expression of $(\alpha_m^{NZ})^*$, thus obtained, is given in (34).

References

[1] C. Pan et al., “An overview of signal processing techniques for RIS/RSS-aided wireless systems,” *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 883–917, Aug. 2022.

[2] Z.-Q. He and X. Yuan, “Cascade channel estimation for large intelligent metasurface assisted massive MIMO,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[3] A. Taha, M. Alrabeiah, and A. Alkhateeb, “Enabling large intelligent surfaces with compressive sensing and deep learning,” *IEEE Access*, vol. 9, pp. 44304–44321, 2021.

[4] P. Wang, J. Fang, H. Duan, and H. Li, “Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems,” *IEEE Signal Process. Lett.*, vol. 27, pp. 905–909, 2020.

[5] T. Lin, X. Yu, Y. Zhu, and R. Schober, “Channel estimation for IRS-assisted millimeter-wave MIMO systems: Sparsity-inspired approaches,” *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 4078–4092, Jun. 2022.

[6] G. Zhou, C. Pan, H. Ren, P. Popovski, and A. L. Swindlehurst, “Channel estimation for RIS-aided multiuser millimeter-wave systems,” *IEEE Trans. Signal Process.*, vol. 70, pp. 1478–1492, 2022.

[7] Z. Peng et al., “Channel estimation for RIS-aided multi-user mmWave systems with uniform planar arrays,” *IEEE Trans. Commun.*, vol. 70, no. 12, pp. 8105–8122, Dec. 2022.

[8] J. He, H. Wymeersch, and M. Juntti, “Channel estimation for RIS-aided mmWave MIMO systems via atomic norm minimization,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5786–5797, Sep. 2021.

[9] X. Wei, D. Shen, and L. Dai, “Channel estimation for RIS assisted wireless communications—Part II: An improved solution based on double-structured sparsity,” *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1403–1407, May 2021.

[10] J. Wu, Y. Li, and L. Xin, “Joint channel estimation for RIS-assisted wireless communication system,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 1087–1092.

[11] G. Joseph and C. R. Murthy, “On the convergence of a Bayesian algorithm for joint dictionary learning and sparse recovery,” *IEEE Trans. Signal Process.*, vol. 68, pp. 343–358, 2020.

[12] D. P. Wipf and B. D. Rao, “Sparse Bayesian learning for basis selection,” in *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[13] C. Ruan, Z. Zhang, H. Jiang, J. Ding, L. Wu, and H. Zhang, “Approximate message passing for channel estimation in reconfigurable intelligent surface aided MIMO multiuser systems,” *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5469–5481, Aug. 2022.

[14] H. Dong, C. Ji, L. Zhou, J. Dai, and Z. Ye, “Sparse channel estimation with surface clustering for IRS-assisted OFDM systems,” *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1083–1095, Feb. 2023.

[15] C. K. Thomas and D. Stoleck, “SAVE—Space alternating variational estimation for sparse Bayesian learning,” in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2018, pp. 11–15.

[16] H. Duan, L. Yang, J. Fang, and H. Li, “Fast inverse-free sparse Bayesian learning via relaxed evidence lower bound maximization,” *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 774–778, Jun. 2017.

[17] B. Worley, “Scalable mean-field sparse Bayesian learning,” in *Proc. IEEE Trans. Signal Process.*, vol. 68, pp. 343–358, 2020.

[18] D. P. Wipf and B. D. Rao, “Sparse Bayesian learning for basis selection,” in *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[19] C. Ruan, Z. Zhang, H. Jiang, J. Ding, L. Wu, and H. Zhang, “Approximate message passing for channel estimation in reconfigurable intelligent surface aided MIMO multiuser systems,” *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5469–5481, Aug. 2022.

[20] H. Dong, C. Ji, L. Zhou, J. Dai, and Z. Ye, “Sparse channel estimation with surface clustering for IRS-assisted OFDM systems,” *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1083–1095, Feb. 2023.

[21] C. K. Thomas and D. Stoleck, “SAVE—Space alternating variational estimation for sparse Bayesian learning,” in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2018, pp. 11–15.

[22] H. Duan, L. Yang, J. Fang, and H. Li, “Fast inverse-free sparse Bayesian learning via relaxed evidence lower bound maximization,” *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 774–778, Jun. 2017.

[23] B. Worley, “Scalable mean-field sparse Bayesian learning,” in *Proc. IEEE Trans. Signal Process.*, vol. 68, pp. 343–358, 2020.

[24] D. P. Wipf and B. D. Rao, “Sparse Bayesian learning for basis selection,” in *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[25] C. Ruan, Z. Zhang, H. Jiang, J. Ding, L. Wu, and H. Zhang, “Approximate message passing for channel estimation in reconfigurable intelligent surface aided MIMO multiuser systems,” *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5469–5481, Aug. 2022.

[26] H. Dong, C. Ji, L. Zhou, J. Dai, and Z. Ye, “Sparse channel estimation with surface clustering for IRS-assisted OFDM systems,” *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1083–1095, Feb. 2023.

[27] C. K. Thomas and D. Stoleck, “SAVE—Space alternating variational estimation for sparse Bayesian learning,” in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2018, pp. 11–15.
[23] R. W. Heath Jr., N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” IEEE J. Sel. Topics Signal Process., vol. 10, no. 3, pp. 436–453, Apr. 2016.

[24] Ö. T. Demir, E. Björnson, and L. Sanguinetti, “Foundations of user-centric cell-free massive MIMO,” Found. Trends Signal Process., vol. 14, nos. 3–4, pp. 162–472, 2021.

[25] D. P. Wipf and B. D. Rao, “An empirical Bayesian strategy for solving the simultaneous sparse approximation problem,” IEEE Trans. Signal Process., vol. 55, no. 7, pp. 3704–3716, Jul. 2007.

[26] A. Rajoriya, S. Rukhsana, and R. Budhiraja, “Centralized and decentralized active user detection and channel estimation in mMTC,” IEEE Trans. Commun., vol. 70, no. 3, pp. 1759–1776, Mar. 2022.

[27] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Berlin, Germany: Springer-Verlag, 2006.

[28] S. Khanna and C. R. Murthy, “Communication-efficient decentralized sparse Bayesian learning of joint sparse signals,” IEEE Trans. Signal Inf. Process. Netw., vol. 3, no. 3, pp. 617–630, Sep. 2017.

[29] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, “Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver,” IEEE Trans. Inf. Theory, vol. 67, no. 5, pp. 2925–2951, May 2021.

[30] Q. Wu and R. Zhang, “Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts,” IEEE Trans. Commun., vol. 68, no. 3, pp. 1838–1851, Mar. 2020.

Milind Nakul received the dual B.Tech. and M.Tech. degree in electrical engineering from the Indian Institute of Technology Kanpur, India, in 2023. He is currently pursuing the Ph.D. degree in machine learning with the Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His research interests include reinforcement learning and optimization.

Anupama Rajoriya (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology, Bhopal, in 2015, and the M.Tech. and Ph.D. degrees from the Department of Electrical Engineering, Indian Institute of Technology Kanpur (IIT Kanpur) in 2017 and 2023, respectively. She is currently a Senior Project Executive Officer with IIT Kanpur, where she is designing embedded software algorithms for 5G and beyond systems. Her research interests include 5G wireless systems, compressed sensing, distributed Bayesian learning, and Bayesian neural networks.

Rohit Budhiraja (Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from IIT Madras in 2004 and 2015, respectively. From 2004 to 2011, he worked for two start-ups where he designed both hardware and software algorithms, from scratch, for physical layer processing of WiMAX- and LTE-based cellular systems. He is currently an Assistant Professor with IIT Kanpur, where he is also leading an effort to design a 5G research testbed. His current research interests include design of energy-efficient transceiver algorithms for 5G massive MIMO and full-duplex systems, robust precoder design for wireless relaying, machine learning methods for channel estimation in mm-wave systems, and spatial modulation system design. His paper was shortlisted as one of the finalists for the Best Student Paper Award from the IEEE International Conference on Signal Processing and Communications, Bengaluru, India, in 2014. He also received the IIT Madras Research Award for the quality and quantity of research work done in the Ph.D., Early Career Research Award and the Teaching Excellence Certificate from IIT Kanpur.