Channel estimation using variational Bayesian learning for multi-user mmWave MIMO systems

Bo Xiao1 | Pingmu Huang1 | Zhipeng Lin1 | Jie Zeng2 | Tiejun Lv1

1 School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China
2 Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Correspondence
Tiejun Lv, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China. Email: lvteijun@bupt.edu.cn

Abstract
This paper presents a novel variational Bayesian learning-based channel estimation scheme for hybrid pre-coding-employed wideband multiuser millimetre wave multiple-input multiple-output communication systems. We first propose a frequency variational Bayesian algorithm, which leverages common sparsity of different sub-carriers in the frequency domain. The algorithm shares all the information of the support sets from the measurement matrices, significantly improving channel estimation accuracy. To enhance robustness of the frequency variational Bayesian algorithm, we develop a hierarchical Gaussian prior channel model, which employs an identify-and-reject strategy to deal with random outliers imposed by hardware impairments. A support selection frequency variational Bayesian channel estimation algorithm is also proposed, which adaptively selects support sets from the measurement matrices. As a result, the overall computational complexity can be reduced. Validated by the Bayesian Cramér-Rao bound, simulation results show that, both frequency variational Bayesian and support selection-frequency variational Bayesian algorithms can achieve higher channel estimation accuracy than existing methods. Furthermore, compared with frequency variational Bayesian, support selection-frequency variational Bayesian requires significantly lower computational complexity, and hence, it is more practical for channel estimation applications.

1 INTRODUCTION

In recent years, millimetre wave (mmWave) communication, which can provide gigabit-per-second data rates, has received considerable attention [1]. Combined with hybrid large array techniques, applying mmWave has become crucial for successful development of the forthcoming fifth-generation (5G) mobile networks [2]. Nevertheless, estimating channel state information (CSI) is challenging for hybrid pre-coding-employed wideband multi-user mmWave systems due to large and compressed channel matrices [3]. Random outliers imposed by hardware impairments also need to be considered, because they will degrade the channel estimation (CE) performance of CE algorithms [4]. Therefore, it is necessary to devise innovative CE algorithms to address the challenges in mmWave communication systems.

1.1 Prior work

Most of CE algorithms in mmWave systems concentrate on exploiting the channel sparsity [5]. By using compressed sensing theory, a sparse CE problem can be formulated as a sparse signal recovery problem. Then, the required pilot overhead can be significantly reduced as compared to conventional algorithms, which uses least square (LS) [6] or minimum mean squared error methods [7]. Assuming that the estimated amplitudes of non-zero coefficients are channel gains of each path, applying sparse exploration is to compare the pairs of angle of departure (AoD) and angle of arrival (AoA) of each path in mmWave systems [8]. Most existing CE methods use narrow-band flat channel models [8–10]. However, in practice, mmWave channels generally have wideband and frequency-selective fading characteristics.
There are some publications that propose CE algorithms for wideband mmWave channels [11–16]. By leveraging joint sparsity of angular and delay domains, a time-domain CE algorithm using quantised sparse recovery is proposed in [11]. However, this algorithm is only applicable for single-carrier systems. Based on a multi-resolution codebook, some alternative CE algorithms are presented. A distributed grid matching pursuit-based CE algorithm is proposed in [12] to estimate the channel parameters of frequency-selective fading channel. By exploiting the codebook structure, a combined time-frequency approach is proposed, which can estimate the wideband CSI with low pilot overhead [13]. Authors in [14] propose a support detection (SD)-based CE algorithm. By observing the structural characteristics of mmWave beamspace channel, this algorithm can detect the support of the sparse beamspace [14]. Moreover, a simultaneous weighted orthogonal matching pursuit (SWOMP) algorithm is proposed in [15]. Since this algorithm considers realistic frequency-selective fading channel models, it can exploit spatially common sparsity of mmWave channels. Nevertheless, all these algorithms [12–15] assume that there is no multi-user interference. Therefore, they can only estimate one user CSI at a time. In addition, when applying these algorithms, grid errors would be induced to channel coefficients, because multi-resolution codebook-based CE strategies are employed [17, 18].

In order to improve the overall CE accuracy, sparse Bayesian learning [19–21] is widely used to estimate the channel parameters. This method can avoid the grid errors, and hence, improve the estimation performance. Specifically, an expectation maximisation-based sparse Bayesian learning is proposed in [19]. By using Gaussian mixture prior models, this method can achieve high estimation accuracy with low pilot overhead. [20, 21] use generalised approximate message passing-based methods [20, 21] to estimate CSI of mmWave multiple-input multiple-output (MIMO) systems. However, all these methods do not consider hardware impairments [4], which affects the performance of CE [22]. To solve this problem, [23] proposes a Bayesian compressed sensing-based CE algorithm for mmWave channels with random outliers. However, this algorithm [23] is only applicable for point to point narrow-band flat channels.

1.2 Contributions

Motivated by above, we propose a novel CE scheme based on variational Bayesian inference in the frequency domain. This scheme is designed for wideband multi-user mmWave MIMO communication systems, where hybrid pre-coding architectures and frequency-selective fading channels are considered. A frequency variational Bayesian (FVB) algorithm is presented, which exploits the common sparsity of different subcarriers. The algorithm formulates the CE problem as a sparse Bayesian problem, and thus, it can be solved by using variational Bayesian inference. By simplifying the inverse derivations, a support selection-FVB (SS-FVB) algorithm is also proposed to reduce the computational complexity. In addition, we provide a Bayesian Cramér-Rao bound (BCRB) to evaluate the estimation performance of the proposed algorithms. To the best of our knowledge, our paper is the first publication that considers the effect of the random outliers imposed by hardware impairments for the multi-user wideband CE with hybrid architectures. We summarise the specific contributions as follows.

1. We propose a method that incorporates the machine learning technique into CE of wideband multi-user mmWave MIMO systems. Specifically, the sparse signal recovery problem is formulated as a sparse Bayesian learning problem, which is extensible and robust to inference. Considering the random outliers, a hierarchical Gaussian prior channel model is developed based on the identify-and-reject strategy, and thus, the sparsity of the vectorised CSI is encouraged by hyper-parameters. As a result, the robustness against random outliers can be significantly enhanced.

2. The variational Bayesian inference is employed to solve the sparse Bayesian learning problem. The Kullback-Leibler divergence is also introduced to tackle CSI. We calculate all the posterior distributions of the latent variables from the parameter. Hence, the estimation accuracy of the proposed CE algorithms is significantly improved.

3. The proposed scheme achieves an excellent performance in terms of CE accuracy. Specifically, the FVB algorithm shares all the information of support sets from the measurement matrices, and provides a better estimation performance. For the SS-FVB algorithm, by setting some columns of measurement matrices to zeros, the computational complexity can be reduced.

The work in paper substantially extends our previous work [25], where a robust frequency-distributed variational Bayesian estimation algorithm is presented, and its performance is assessed by comparing the estimation accuracy with some existing algorithms. However, [25] does not provide the analysis of algorithm complexity, and does not solve the problem of high computational complexity caused by using the proposed algorithm. The computation of BCRB is also omitted in [25]. To make our work complete and deep, in this paper, we propose an SS-FVB algorithm to reduce the overall computational complexity. We analyse the computational complexity of the proposed two CE algorithms in detail and summarise it. We also provide the detailed deducing of BCRB in Section IV.

The rest of this paper is organised as follows. Section II presents the wideband multi-user mmWave MIMO system model including a frequency-selective fading geometric channel model and a hybrid pre-coding architecture. In Section III, the CE problem in the frequency domain is further formulated as a sparse Bayesian learning framework. Section IV proposes a CE scheme based on variational Bayesian learning in detail, along with the derivations of the computational complexity and the BCRB. The numerical results in Section V are presented to compare the estimation performance of the two proposed algorithms. Finally, Section VI concludes the main conclusions derived from the simulation results with some ideas for further research work.
1.3 | Notation

The notations used throughout this paper are denoted as follows. \( \mathbf{A} \), \( \mathbf{a} \) and \( a \) denote matrix, column vector and scalar value, respectively. \( \mathbf{A}_i \) is used to denote the \( i \)-th row of \( \mathbf{A} \). \( \mathbb{C}^{M \times N} \) represents the space of \( M \times N \) complex-valued matrices. \( (\cdot)^T \), \( (\cdot)^H \), \( ||\cdot||_2 \), \( ||\cdot||_F \) and \( \langle \cdot \rangle \) denote the conjugate transpose, transpose, \( l_2 \)-norm, Frobenius norm, and expectation, respectively. \( \mathbf{Φ} \) is the Kronecker product and \( \text{vec}(\cdot) \) is the vectorisation operation by the columns of the matrix. \( \langle \mathbf{x} \rangle_{\mathbb{N}} \) is to take the expected value of \( \mathbf{x} \) according to the distribution of \( p(\mathbf{x}) \).

2 | SYSTEM MODEL

As illustrated in Figure 1, we consider an orthogonal frequency division multiplexing (OFDM)-based multi-user mmWave MIMO communication system, which employs hybrid precoding architectures and frequency-selective fading geometric channel models. In this system, we assume that there are \( K \) user equipments (UEs) equipped with \( N_{\text{UE}} \) antennas and \( M_{\text{RF}} \) radio frequency (RF) chains. The UEs transmit pilot sequences to a base station (BS) with \( N_{\text{BS}} \) antennas and \( M_{\text{BS}} \) RF chains.

2.1 | Channel model

We consider a frequency-selective fading geometric channel model according to the IEEE 802.11ad wireless standard [24]. The channel matrix \( \mathbf{H}_{d,k} \in \mathbb{C}^{N_{\text{BS}} \times N_{\text{UE}}} \) of the \( d \)-th delay tap and the \( k \)-th user can be expressed as [26]

\[
\mathbf{H}_{d,k} = \sum_{i=1}^{L_d} \mathbf{\alpha}_{i,k} \beta(\tau_{i,k}) \mathbf{a}_{\text{BS}}(\varphi_{i,k}) \mathbf{a}_{\text{UE}}^H(\psi_{i,k})
\]

(1)

where \( L_d \) represents the number of paths; \( \mathbf{\alpha}_{i,k} \in \mathbb{C}^N \) and \( \beta(\tau_{i,k}) \in \mathbb{C} \) are the complex gain and the time-domain delay of the \( i \)-th path, respectively; and \( \mathbf{a}_{\text{BS}}(\varphi_{i,k}) \) and \( \mathbf{a}_{\text{UE}}(\psi_{i,k}) \) are antenna steering vectors of the BS and the UEs with \( \varphi_{i,k} \in [0, 2\pi] \) and \( \psi_{i,k} \in [0, 2\pi] \) being AoD and AoA, respectively. \( B(\tau_{i,k}) = \rho_{n_d} (\delta T_i - \tau_{i,k}) \), where \( T_i \) is the sampling period and \( \rho_{n_d} \) denote the pulse-shaping filter at time \( \tau \). Typical uniform linear arrays (ULA) with half-wavelength separation are deployed at the BS and the UEs. Hence, the elements of \( \mathbf{a}_{\text{BS}}(\varphi_{i,k}) \) and \( \mathbf{a}_{\text{UE}}(\psi_{i,k}) \) can be expressed as

\[
\begin{align*}
\mathbf{a}_{\text{BS}}(\varphi_{i,k})_{q_{\text{BS}},1} &= \frac{1}{\sqrt{N_{\text{BS}}}} e^{j n_{\text{BS}} (q_{\text{BS}} - 1) \cos(\varphi_{i,k})} \\
\mathbf{a}_{\text{UE}}(\psi_{i,k})_{n_{\text{UE}},1} &= \frac{1}{\sqrt{N_{\text{UE}}}} e^{j n_{\text{UE}} (n_{\text{UE}} - 1) \cos(\psi_{i,k})}.
\end{align*}
\]

(2)

An extended virtual channel model [24] is employed to approximate the channel matrix \( \mathbf{H}_{d,k} \) as

\[
\mathbf{H}_{d,k} \approx \mathbf{a}_{\text{BS}} \mathbf{H}_{d,k} \mathbf{a}_{\text{UE}}^H
\]

(3)

where \( \mathbf{H}_{d,k} \in \mathbb{C}^{G_{\text{BS}} \times G_{\text{UE}}} \) is a sparse matrix of non-zero path gains in the angular domain. \( \mathbf{a}_{\text{BS}} \in \mathbb{C}^{N_{\text{BS}} \times G_{\text{BS}}} \) and \( \mathbf{a}_{\text{UE}} \in \mathbb{C}^{N_{\text{UE}} \times G_{\text{UE}}} \) are the dictionary matrices with sizes \( G_{\text{BS}} \) and \( G_{\text{UE}} \) for AoA and AoD, respectively. To exploit the common sparsity of different subcarriers, the frequency-domain channel matrix with the length of the delay taps \( N_{\text{C}} \) at the subcarrier \( p \) (\( 0 \leq p \leq P - 1 \)) can be written as

\[
\mathbf{H}_{p,k} \approx \sum_{d=1}^{N_{\text{C}}} \mathbf{H}_{d,k} e^{-j \frac{2\pi d}{P}} \approx \mathbf{a}_{\text{BS}} \mathbf{H}_{p,k} \mathbf{a}_{\text{UE}}^H
\]

(4)

where \( \mathbf{H}_{p,k} \in \mathbb{C}^{G_{\text{BS}} \times G_{\text{UE}}} \) denotes a sparse matrix for the \( k \)-th user at the \( p \)-th sub-carrier and \( P \) is the number of the subcarriers.

2.2 | Hybrid pre-coding

Considering that the BS uses a hybrid combining scheme, \( \mathbf{W}_p^{(l)} = \mathbf{W}_{\text{RF}}^{(l)} \mathbf{W}_{\text{BB},p}^{(l)} \in \mathbb{C}^{N_{\text{BS}} \times M_{\text{BS}}} \), where \( \mathbf{W}_{\text{RF}}^{(l)} \in \mathbb{C}^{N_{\text{BS}} \times M_{\text{BS}}} \) and \( \mathbf{W}_{\text{BB},p}^{(l)} \in \mathbb{C}^{M_{\text{BS}} \times M_{\text{BS}}} \) denote the analog and digital combining matrices at the \( l \)-th time frame, respectively. \( \mathbf{W}_p^{(l)} \) is designed as cyclically shifted versions, \( \{\mathbf{W}_p^{(l)}\}^T = \mathbf{I}_{\text{BS}} \) to provide a broad beam direction [11], and we employ zero padding (ZP) [13] instead of the cyclic prefix to avoid corrupting the pilot at the symbol edges. The received signal at the \( p \)-th sub-carrier

\[
\mathbf{r}_p^{(l)} = \mathbf{H}_p^{(l)} \mathbf{a}_p^{(l)} \mathbf{W}_p^{(l)} \mathbf{W}_p^{H}(l) + \mathbf{n}_p^{(l)}
\]
can be written as

$$y_p(t) = \sum_{k=1}^{K} (W_p(t))^H H_{p,k} x_{p,k}^{(t)} + u_p(t) + v_p(t)$$  \hspace{1cm} (5)$$

where $u_p(t) = (W_p(t))^H w_p(t) \in \mathbb{C}^{M_{BS} \times 1}$ and $v_p(t) = (W_p(t))^H n_p(t) \in \mathbb{C}^{M_{BS} \times 1}$ are the effective noises from the hybrid combining procedure; $w_p(t)$ denotes the vector of the random outliers including $G \leq M_{BS}$ non-zero entries; and $n_p(t)$ is the additive multi-variate Gaussian noise vector. In the above equation, $x_{p,k}^{(t)}$ is the effective pilot signal transmitted by the $k$-th user at the $p$-th sub-carrier, which can be expressed as

$$x_{p,k}^{(t)} = F_{p,k}^{(t)} s_{p,k}^{(t)}$$  \hspace{1cm} (6)$$

where $s_{p,k}^{(t)} \in \mathbb{C}^{M_{LE} \times 1}$ is the transmitted training sequence, and $F_{p,k}^{(t)} = F_{RF,k}^{(t)} F_{BB,k}^{(t)} \in \mathbb{C}^{M_{LE} \times M_{LE}}$ is the hybrid pre-coding of the $t$-th OFDM symbol with $F_{RF,k}^{(t)} \in \mathbb{C}^{M_{LE} \times M_{LE}}$ and $F_{BB,k}^{(t)} \in \mathbb{C}^{M_{LE} \times M_{UE}}$ being the analog and digital pre-coding matrices, respectively.

### PROBLEM FORMULATION

In this section, we propose a two-step approach to identify the CE problem. We first formulate the CE process as a sparse signal recovery problem in the frequency domain. Then, we propose a sparse Bayesian learning problem based on a hierarchical Gaussian prior channel model, which employs the identify-and-reject strategy. Using this model for CE, the robustness and estimation accuracy of the proposed algorithm can be improved.

#### 3.1 CE formulation

As the scatters of BS are much smaller than those of the UEs in the single-cell mmWave uplink MIMO system, a small angular spread will appear at the receiver, resulting in channel sparsity [28]. In particular, the number of effective paths generally is less than 30 [28]. Therefore, in this system the vectorised channel vector is sparse, when the sizes of dictionary matrices $G_{BS}$ and $G_{UE}$ are larger than 64 [29]. According to the sparsity of mmWave channels, Equation (5) can be rewritten as

$$y_p(t) = \mathcal{V} p h_p + u_p + v_p$$  \hspace{1cm} (7)$$

where $H_{p,k} = [H_{p,1}, ..., H_{p,K}]$, $x_{p,k}^{(t)} = [(x_{p,1}^{(t)})^T, ..., (x_{p,K}^{(t)})^T]^T$, $\Phi = \{A_{UE,k}^H \times \text{variances} A_{BS}^H, A_{BS}^H = \text{diag}(A_{UE,k1}^H, ..., A_{UE,kK}^H)\}$, and $h_p = \text{vec}(H_{p,k}) \in \mathbb{C}^{K_{GS} \times M_{BS}}$ is a sparse angular vector with $KL$ non-zero values. Since the available signal-to-noise ratio (SNR) of mmWave communication systems is generally very low, $M$ successive OFDM symbols within the channel coherence time are required to estimate the CSI. By stacking the $M$ measurements into a vector, Equation (7) can be expressed as

$$y_p = \Psi h_p + u_p + v_p$$  \hspace{1cm} (8)$$

where $\Psi = \{\mathcal{V} p \}$ is the measurement operator.
where
\[ \mathbf{y}_p = \left[ (\mathbf{y}_p^{(1)})^T , ..., (\mathbf{y}_p^{(M)})^T \right]^T , \quad \mathbf{u}_p = \left[ (\mathbf{u}_p^{(1)})^T , ..., (\mathbf{u}_p^{(M)})^T \right]^T , \quad \text{and} \quad \mathbf{v}_p = \left[ (\mathbf{v}_p^{(1)})^T , ..., (\mathbf{v}_p^{(M)})^T \right]^T . \]

Note that the number of the required pilot dramatically increases when the traditional LS algorithm is used. In order to reduce the required pilot overhead, compressed sensing-based methods, i.e., orthogonal matching pursuit (OMP) [30], compressive sensing matching pursuit (CoSaMP) [31] and the iterative re-weighted \( \ell_1 \) and \( \ell_2 \) algorithms [32], are used to recover the sparse vector \( \mathbf{h}_p \) from \( \mathbf{y}_p \) [see Equation (8)].

On the other hand, we can observe that the frequency-domain channel model in (4) exhibits the same sparse feature for any sub-carrier \( p \) [24]. Therefore, \( \mathbf{h}_p \) can be seen as a vectorised sparse channel vector \( \mathbf{h} = [h_1 , ..., h_N]^T \), where \( N = KG_{BS}G_{UE} \) is the length of the vector common of each sub-carrier. Moreover, the number of required training frames can be substantially reduced due to the common sparse feature of different sub-carriers in the mmWave channels.

### 3.2 Sparse Bayesian learning

Compared to the signal recovery algorithms based on compressed sensing theory, a robust and extensible inference framework based on sparse Bayesian learning, can provide better estimation performance for sparse signal recovery problems [33]. By using this method, the sparsity of the estimated signal is encouraged by employing a sparsity-promoting prior, and all the corresponding components, such as hybrid pre-coding, additive Gaussian noise and random outliers, are presented following a sparse Bayesian learning framework. As a result, a hierarchical Gaussian prior model [34] based on two layers can be constructed.

We assume that non-zero values of channel vector \( \mathbf{h} \) in the first layer are independent and identically distributed (i.i.d.) and the complex Gaussian distribution [34], i.e., \( b_n | \alpha_n \sim \mathcal{CN}(0, f_n^{-1}) \), \( n = 1, 2, ..., N \), where \( \alpha_n \) represents the precision of \( b_n \). We have
\[
\mathbf{h}(\alpha) = \prod_{n=1}^{N} p(b_n | \alpha_n) \quad (10)
\]

where \( \alpha = [\alpha_1 , \alpha_2 , ..., \alpha_N]^T \) denotes the sparsity of the channel \( \mathbf{h} \).

In the second layer, the prior precision is also considered as i.i.d. random variables following the Gamma distribution.

Therefore, the distribution of prior precision is given by
\[
p(\alpha, a, b) = \Gamma(\alpha)^{-1} b^a \alpha^{a-1} e^{-b\alpha} \quad (11)
\]
with \( \alpha \geq 0 \), where
\[
\Gamma(\alpha) = \int_0^\infty \tau^{\alpha-1} e^{-\tau} d\tau \quad (12)
\]
is the Gamma function. \( a \) and \( b \) are set as small constant values to make these priors non-informative over \( \{\alpha_n\} \). By using the same method as Equation (11), the noise variance \( \gamma \) can be expressed as
\[
p(\gamma) = \Gamma(\gamma)^{-1} d^{\gamma-1} e^{-d\gamma} \quad (13)
\]
where \( c \) and \( d \) are small values [34].

Considering the random outliers caused by hardware impairments, a set of indicator variables \( \{\theta_m\} \) is assigned to detect the outliers by employing the identify-and-reject strategy [35]. Therefore, the received signal can be expressed as
\[
y_m = \left\{ \begin{array}{l}
\mathbf{y}_m \mathbf{h} + u_m + v_m, \quad \theta_m = 0 \\
\mathbf{y}_m \mathbf{h} + v_m, \quad \theta_m = 1
\end{array} \right. \quad (14)
\]
where \( \mathbf{y}_m \) denotes the \( m \)-th row of \( \mathbf{y}, u_m, v_m \), and \( \theta_m \) are the \( m \)-th entry of \( u, v \), and \( \theta \), respectively. Thus, the probability density function (PDF) of the received signal can be rewritten as
\[
p(y | h, \theta, \gamma) = \prod_{m=1}^{M} (\mathcal{N}(y_m | \mathbf{y}_m \mathbf{h}, 1/\gamma))^{\theta_m} \quad (15)
\]
This operation ensures that, in the presence of random outliers, the received signal \( y_m \) is independent to the PDF \( p(y | h, \theta, \gamma) \).

Since \( \theta_m \in \{0, 1\} \), a beta-Bernoulli hierarchical prior placed on \( \theta_m \) is
\[
p(\theta_m | \beta_m) = \beta_m^\theta (1 - \beta_m)^{(1 - \theta_m)}, \forall m \quad (16)
\]
where \( \beta_m \) denotes the \( m \)-th entry of \( \beta \). In addition, the distribution of prior precision can be written as
\[
p(\beta_m) = \text{Beta}(\beta_m | r, f), \forall m \quad (17)
\]
where \( r \) and \( f \) follow the Beta distribution. For clarity, Figure 2 gives the graphical model of the hierarchical prior channel model based on the identify-and-reject strategy, where the constants, the hidden variables and the observation are denoted as squares, circles and a shaded circle, respectively.

Now the considered CE problem is formulated as a sparse Bayesian problem, which can be solved by the proposed CE scheme in Section IV.
4 | CE SCHEME DEVELOPMENT

In this section, we propose and analyse the novel CE scheme to solve the sparse Bayesian learning problem. Generally, \( h \) can be estimated by the posterior distribution \( p(h|y) \) \([36]\). As will be shown in Section V, by employing the variational Bayesian inference, the FVB CE algorithm provides an excellent estimation accuracy, because it takes advantage of overall support sets of the measurement matrices. Nevertheless, the computational complexity of the FVB algorithm is relatively high since the inverse of the measurement matrix is complicated. To reduce the computational complexity, the SS-FVB algorithm is proposed by using an adaptive selection operation based on the correlations between different support sets. We also derive the computational complexity and the BCRB in this section to analyse the performance of the proposed algorithms.

4.1 | FVB CE algorithm

We apply variational Bayesian inference to solve the sparse Bayesian learning problem, and thus, the convergence to the local optimum can be guaranteed \([37]\). For simplicity, the set of the unobserved variables \( \{ \theta, h, \beta, \alpha, \gamma \} \) is denoted as \( \{ z \} \). Since the derivation of \( p(z|y) \) is difficult to compute directly \([36]\), we decompose the log-marginal probability of the received signal into two terms

\[
\ln \hat{p}(y) = L(q) + \text{KL}(q||\hat{p}) \tag{18}
\]

where \( L(q) = \int q(z) \ln p(y, z)/q(z) dz \), \( \text{KL}(q||\hat{p}) = -\int q(z) \ln p(z|y)/q(z) dz \) is the Kullback-Leibler divergence which describes the distance between \( p(z|y) \) and \( q(z) \) \([38]\), and \( q(z) \) can be any PDF close to \( p(z|y) \). As \( p(z|y) \leq q(z) \) and \( \text{KL}(q||\hat{p}) \geq 0 \), the nearest analytical expression of the posterior distribution \( p(z|y) \) can be obtained by minimising the Kullback-Leibler divergence. As \( \ln \hat{p}(y) \) is only dependent to \( y \), the problem of minimising the Kullback-Leibler divergence is equivalent to maximising \( L(q) \). Consequently, our objective becomes to approximate the \( p(z|y) \) by maximising \( L(q) \) with respect to \( q(z) \), i.e., \( q(\theta), q(\phi), q(y), q(\beta) \) \([36]\). Details of the derivations are given as follows.

4.1.1 | Derivation of \( q(y) \)

According to Equations (13) and (15), \( \ln q(y) \) can be expressed as

\[
\ln q(y) \propto \left(\ln \hat{p}(y|h, \theta, \beta) + \ln \hat{p}(y|\phi, d)\right)_{q(h), q(\phi)}
\]

Rewriting (19) in the compact form: \( q(y) \propto (\bar{c} - 1) y - d y \), where \( \bar{c} = c + 0.5 \sum_{m=1}^{M} M \langle \theta_m \rangle \) and

\[
d = d + 0.5(y - \Psi h)^T D_{\theta}(y - \Psi h) + \text{trace}(\Psi^T D_{\theta} \Psi) \tag{20}
\]

\( q(y) \) obeys a Gamma distribution as

\[
q(y) = \text{Gamma}(\bar{c}, d).
\]

4.1.2 | Derivations of \( q(h) \) and \( q(\alpha) \)

Based on Equations (11) and (15), \( \ln q(h) \) can be derived as

\[
\ln q(h)
\]

\[
\propto \left(\ln \hat{p}(y|h, \theta, \beta) + \ln \hat{p}(h|\alpha)\right)_{q(\theta), q(h), q(\phi)}
\]

\[
\propto -\frac{1}{2} h^T \Sigma^{-1} h - h^T \Sigma^{-1} \mu_h. \tag{22}
\]

Since Equation (22) is the exponent of Gaussian distribution, we have \( \mu_h = (\bar{c}) \Sigma \Psi^T D_{\theta} \Sigma \Psi^T D_{\phi}^{-1} \), where \( D_{\phi} = \text{diag}(\{ \phi \}) \) and \( D_{\alpha} = \text{diag}(\{ \alpha \}) \). \( q(h) \) follows a Gaussian distribution, i.e., \( q(h) = \text{N}(\mu_h, \Sigma) \).

We can also obtain

\[
\ln q(\alpha)
\]

\[
\propto \left(\ln \hat{p}(h|\alpha) + \ln \hat{p}(\alpha|a, b)\right)_{q(h)}
\]

\[
\propto (a - \frac{1}{2}) \sum_{n=1}^{N} \ln \alpha_n - \sum_{n=1}^{N} \frac{1}{2} \langle h_n^2 \rangle + b \alpha_n. \tag{23}
\]

As Equation (20) is the exponent of Gamma distribution, \( q(\alpha) \) can be expressed as

\[
q(\alpha) = \prod_{n=1}^{N} \text{Gamma}(\alpha_n | a, b_n). \tag{24}
\]

where

\[
\bar{c} = 0.5 + a,
\]

\[
\tilde{b}_n = 0.5 \langle h_n^2 \rangle + \Sigma_{n,n} + b \tag{25}
\]

and \( \Sigma_{n,n} \) is the \( n \)-th diagonal element of \( \Sigma \) in Equation (22).
ALGORITHM 1 Proposed FVB Algorithm

\begin{flushleft}
\textbf{Input:} \( \mathbf{y}, \mathbf{x}, W, A_{BS}, A_{UE, K} \) and tolerance \( \tau \).
\textbf{Output:} The channel estimator \( \mathbf{h}(k) \) of the \( k \)-th user.
\begin{enumerate}
  \item Initialize \( a = b = \epsilon = d = f = 10^{-6}, \hat{c} = 0 \).
  \item Compute the matrix \( \mathbf{W} \) according to (8).
  \item Repeat
    \begin{enumerate}
      \item Compute \( \hat{a} \) and \( \hat{b} \) in Equations (20) and (25) to update the hyper-parameters \( \alpha \) and \( \gamma \), respectively.
      \item Detect the indicator \( \Theta \) relying on Equations (27) and (28), and reject the outliers.
      \item Until \( ||\mathbf{I}^\text{new} - \mathbf{I}^\text{old}||_2 < \tau \).
      \item Estimate the channel vector \( \mathbf{h} \).
    \end{enumerate}
  \item For \( k = 1, \ldots, K \) do
    \begin{enumerate}
      \item Find the channel vector \( \mathbf{h}(k) \) of the \( k \)-th user by detecting the non-zero values.
      \item \( \mathbf{h}(k) = [\hat{h}_k]_{k=1}^{K} \).\( G_{BS,CU,1:K}G_{BS,CU,K} \).
    \end{enumerate}
\end{enumerate}
\end{flushleft}

4.1.3 Derivations of \( q(\Theta) \) and \( q(\beta) \)

According to Equations (15) and (16), the derivation of \( \ln q(\Theta) \) is obtained as

\[
\ln q(\Theta) \propto \left\{ \ln p(\mathbf{y}|\mathbf{h}, \Theta, \gamma) + \ln p(\Theta|\gamma) \right\}_g(h, \gamma; q(\gamma))
\]
\[
\propto \sum_{m=1}^{M} \left( \frac{\gamma y_m - \mathbf{m}^\top \mathbf{h}^2}{2} + \ln \beta_m \right) + (1 - \Theta_m) \times \ln(1 - \beta_m), \quad (26)
\]

Note that \( \Theta_m \) still obeys a Bernoulli distribution, i.e.,

\[
P(\Theta_m = 1) = C e^{(1-\beta_m)\gamma y_m - \mathbf{h}^\top \mathbf{m}} \quad (27)
\]
\[
P(\Theta_m = 0) = C e^{(\ln(1-\beta_m))}
\]

where \( C \) is a normalisation constant. We have

\[
P(\Theta_m = 1) + P(\Theta_m = 0) = 1.
\]

Then, \( q(\beta) \) can be derived as

\[
\ln q(\beta) \propto \left\{ \ln p(\Theta|\beta) + \ln p(\beta|\epsilon, f) \right\}_g(\Theta)
\]
\[
\propto \sum_{m=1}^{M} (\Theta_m \ln \beta_m + (1 - \theta_m) \ln(1 - \beta_m) + (f - 1) \ln \beta_m)
\]
\[
+ (f - 1) \ln(1 - \beta_m) \}
\]
\[
\propto \sum_{m=1}^{M} (\Theta_m + 1) \ln \beta_m + (f - 1) \ln(1 - \beta_m), \quad (29)
\]

Equation (29) can be rewritten as

\[
q(\beta) = \prod_{m=1}^{M} \beta(\beta_m) = \prod_{m=1}^{M} \beta(\beta_m|\theta_m + \epsilon, f + 1 - \theta_m)
\]

which indicates that the posterior \( q(\beta) \) follows a Beta distribution

\[
q(\beta) = \prod_{m=1}^{M} \beta(\beta_m|\tilde{\theta}_m, \tilde{f})
\]

where

\[
\tilde{\theta}_m = \Theta_m + \epsilon, \quad \tilde{f} = 1 + f - \langle \Theta_m \rangle.
\]

Using the previously derived expressions for the variational distributions \( q(\mathbf{h}), q(\gamma) \) and \( q(\Theta) \), the posterior distributions \( p(\mathbf{h}|\gamma) \), \( p(\gamma|\mathbf{y}) \) and \( p(\Theta|\mathbf{y}) \) can be calculated in the same way [39]. Then, the channel vector \( \mathbf{h} \) can be estimated by the posterior mean, i.e.,

\[
\mathbf{h} = \langle \gamma \rangle \mathbf{D}_\mathbf{y}.
\]

The details of the algorithmic implementation of the proposed FVB CE algorithm can be found in Algorithm 1.

4.2 SS-FVB CE algorithm

The FVB CE algorithm requires high computational complexity for the necessary matrix inversion. In addition, the high complexity of the measurement matrix will significantly increase the overall computational complexity of the algorithm.

To reduce the algorithmic complexity, here we propose the reduced complexity CE algorithm, SS-FVB, which is based on the property that a small number of the relevant support set occupies most of the energy of the measurement matrices [18]. This algorithm exploits the correlation between the residuals and the columns of the measurement matrices.

We first propose a convenient reconstruction of the measurement matrix by making a proper adaptive selection operation of the support set. We define the column of the measurement
matrix as a “support set”, which has the strongest correlation with the received signal. The index number of the corresponding support set, $\rho_1$, can be obtained as

$$\rho_1 = \arg \max_{j=1,\ldots,N} \langle y, \Psi_{\rho_1} \rangle.$$  \hfill (31)

Then, a new measurement matrix is constructed as

$$Y_1 = [0, \ldots, 0, \Psi_{\rho_1}, 0, \ldots, 0]_{MMBS \times N}.$$ \hfill (32)

By subtracting the contribution of $Y_1$ from $y$, the residual $r_1$ can be obtained by

$$r_1 = y - Y_1 (Y_1^T Y_1)^{-1} Y_1^T y.$$ \hfill (33)

In the same way, the correlation between the residual and the columns of the original measurement matrix can be exploited to find the new index number as follows:

$$\rho_2 = \arg \max_{j=1,\ldots,N} \bigg| \langle r_1, \Psi_{\rho_2} \rangle \bigg|.$$ \hfill (34)

Without loss of generality, we assume $\rho_1 < \rho_2$, and then, the new measurement matrix $Y_2$ is constructed as

$$Y_2 = [0, \ldots, 0, \Psi_{\rho_1}, 0, \ldots, 0, \Psi_{\rho_2}, 0, \ldots, 0]_{MMBS \times N}.$$ \hfill (35)

The residual $r_2$ can be calculated as Equation (36) by replacing $Y_1$ with $Y_2$. This procedure is repeated until the $K_L$ column vectors of the original measurement matrix $\Psi$ are selected. After the $K_L$ iterations, the corresponding index set $\rho$ can be represented as $\rho = \{\rho_1, \rho_2, \ldots, \rho_{K_L}\}$, where $\rho_{K_L}$ is the $K_L$-th index value. By rearranging the elements of $\rho$ in an ascending order, a new index set can be denoted as $\xi = \{\xi_1, \xi_2, \ldots, \xi_{K_L}\}$, where $\xi_1 < \xi_2 < \ldots < \xi_{K_L}$. Then, the newest measurement matrix, which is expressed as

$$Y_{KL} = [0, \ldots, 0, \Psi_{\xi_1}, 0, \ldots, 0, \Psi_{\xi_2}, 0, \ldots, 0, \Psi_{\xi_{K_L}}, 0, \ldots, 0]_{MMBS \times N}$$ \hfill (36)

can be employed to learn the latent parameters.

By recalling (9), the CE problem can be expressed as

$$y = Y_{KL} h + u + v.$$ \hfill (37)

The same as the FVB CE algorithm, the sparse channel vector $h$ can be estimated by computing the posterior distribution $p(h|y)$. Since the procedure in approximating $p(h|y)$ with $q(h)$ is identical to the one presented for the FVB algorithm, this procedure is omitted. Following Equation (33), the sparse channel vector $h$ can be estimated as

$$\hat{h} = \mu = \langle y \rangle \Sigma_{KL}^{-1} D y.$$ \hfill (38)

Algorithm 2 presents the detailed algorithmic implementation of the proposed SS-FVB algorithm. The following theorem shows that the proposed adaptive selection operation always converges to a local optimum.

**Theorem 1.** After $K_L$ iterations, the selection operation of the support sets converges to a local optimum.

**Proof.** See Appendix A. \hfill \square

### 4.3 Computational complexity

Considering that the inverse operations of $\Psi$ and $Y_{KL}$ require many computational overheads in Algorithms 1 and 2, we derive the computational complexity of the inverses of $\Psi$ and $Y_{KL}$ at every step. More specifically, the measurement matrices are singular matrices, and therefore we discuss the pseudoinverses for $\Psi$ and $Y_{KL}$, i.e.,

$$\begin{align*}
\text{pinv}(\Psi) &= (\Psi^T \Psi)^{-1} \Psi^T, \\
\text{pinv}(Y_{KL}) &= (Y_{KL}^T Y_{KL})^{-1} Y_{KL}^T.
\end{align*}$$

The computational complexity of different mathematical operations are summarised in Table 1. From this table, it can be observed that the computational complexity of the
where E[·] represents the expectation operation and J is a FIM. Assume that the derivatives of the logarithmic function of joint PDF $p(y, h)$ exist. The FIM can be expressed as [41]

$$
J = \mathbb{E} \left[ \left( \frac{\partial \ln p(y, h)}{\partial h} \right) \left( \frac{\partial \ln p(y, h)}{\partial h} \right)^\intercal \right].
$$

(42)

Since Equation (45) is complex to be computed, we propose the following theorem to obtain the expression of BCRB.

**Theorem 2.** It is assumed that $p(y| h)$ satisfies the regularity condition

$$
\mathbb{E} \left[ \frac{\partial \ln p(y|h)}{\partial h} \right] = 0, \text{ for all } h.
$$

Then, the complex FIM is given by the matrix

$$
J = J_{y|h} + J_h
$$

(43)

where

$$
J_{y|h} = -\mathbb{E} \left[ \frac{\partial^2 \ln p(y|h)}{\partial h^2} \right]
$$

and

$$
J_h = \left( \frac{\partial \ln p(h)}{\partial h} \right)^\intercal \left( \frac{\partial \ln p(h)}{\partial h} \right).
$$

Proof. See Appendix B. \qed

According to Equations (11) and (12), it is clear that the true prior distribution $p(h; a, b)$ of the sparse channel vector $h$ reduces to

$$
p(h; a, b) = \int p(h|\alpha)p(\alpha; a, b)d\alpha
$$

$$
= \int \prod_{i=1}^N \mathcal{N}(h_0|0, \alpha^{-1}_s)\Gamma\mathcal{m}(\alpha_s; a, b)d\alpha_s
$$

$$
= \prod_{i=1}^N \mathcal{I}(b_0|\mu, \lambda, \nu)
$$

(44)

It is noted that BCRB, which is a metric to measure the performance of the CE scheme, is the inverse of the Fisher Information Matrix (FIM) [40].

---

**TABLE 1**  Computational complexity of the inverse of $\Psi$ and $\Psi_{KL}$

| Operation                  | $\Psi^T\Psi$                                                                 | $\Psi_{KL}^T\Psi_{KL}$                  |
|----------------------------|-----------------------------------------------------------------------------|-----------------------------------------|
| Addition Operation         | $O(M^2M^2_{BS}N - M^2M^2_{BS})$                                              | $O(M^2M^2_{BS}KL - M^2M^2_{BS})$       |
| Multiplication Operation in Inverse | $O(M^2M^2_{BS}N)$                                                             | $O(M^2M^2_{BS}KL)$                     |
| Inverse Operation          | $O(M^2M^2_{BS})$                                                             | $O(M^2M^2_{BS})$                       |

---

**FIGURE 3**  The running time versus the pilot overhead for the proposed CE algorithms

The running time versus the pilot overhead for the proposed two algorithms in Figure 3, where FVB algorithm at the cost of reduced estimation accuracy. We, therefore, the low computational complexity is achieved by the SS-FVB algorithm. BCRB can be calculated by $\Psi = \frac{1}{N}\sum_{n=1}^N \Psi_n$, where $\Psi_n$ is the inverse of the Fisher Information Matrix (FIM) [40].

3 It is noted that BCRB, which is a metric to measure the performance of the CE scheme, is the inverse of the Fisher Information Matrix (FIM) [40].

4 From now on, $p(h; a, b)$ will be denoted as $p(h)$ for the convenience of the notation.
where \( \text{St}(b_a|\mu, \lambda, \nu) \) is a Student-t distribution with \( \mu = 0, \lambda = a/b \) and \( \nu = 2a \).

Moreover, the prior distribution of \( b_a \) can be obtained as

\[
p(b_a) = \text{St}(b_a|\mu, \lambda, \nu)
\]

where

\[
\text{St}(b_a|\mu, \lambda, \nu) = \frac{\Gamma((\nu + 1)/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\pi \nu} \right)^{1/2} \left( \frac{F}{\nu} \right)^{-\nu + 1/2}
\]

with \( F = \nu + \lambda(b_a - \mu)^2 \).

Taking the log-likelihood function on both sides of Equation (47), we can obtain

\[
\ln p(h) = \sum_{n=1}^{N} \ln p(b_n).
\]

The log-likelihood function of \( p(b_n) \) can be calculated as

\[
L(b_n) = \ln \frac{\nu - 1}{\nu - 2} - \frac{1}{2} \ln \nu + \frac{\nu + 1}{2} \ln \nu + \frac{1}{2} \ln \lambda - \frac{1}{2} \ln \nu + \frac{1}{2} \ln F.
\]

The first-order partial derivative of \( L(b_n) \) is

\[
U(b_n) = \frac{\partial L(b_n)}{\partial b_n} = \left[ \frac{\partial L(b_n)}{\partial \mu}, \frac{\partial L(b_n)}{\partial \lambda}, \frac{\partial L(b_n)}{\partial \nu} \right]^{T}
\]

where

\[
\frac{\partial L(b_n)}{\partial \mu} = -\frac{(\nu + 1)b_a \lambda}{F},
\]

\[
\frac{\partial L(b_n)}{\partial \lambda} = \frac{1}{2 \lambda} - \frac{(\nu + 1)b_n^2}{2F},
\]

\[
\frac{\partial L(b_n)}{\partial \nu} = \frac{1}{2} \ln \left( \frac{\nu}{F} \right) + \frac{\lambda b_n^2 - 1}{2F} - \frac{1}{(\nu - 1)(\nu - 2)}.
\]

Based on Equation (48), the first-order partial derivative of \( \ln p(h) \) is the sum of the \( U(b_n) \). Thus, \( U(h) \) can be calculated as

\[
U(h) = \frac{\partial \ln p(h)}{\partial h} = \sum_{n=1}^{N} U(b_n).
\]

Using Theorem 2, \( J_h \) can be calculated for both FVB and SS-FVB algorithms, as given by

\[
J_h = U(h)^T U(h).
\]

Using the same way as [42], the measurement matrix \( \Psi \) is added as the additional observation, because it will impact the estimation of \( h \). Then, \( J_{y|h} \) for the proposed two algorithms can be expressed as [43]

\[
J_{y|h} = E\left[ -\frac{\partial^2 \ln p(y|\Psi|h)}{\partial h^2} \right].
\]

It can be observed that the derivations of \( J_{y|h} \) and \( J_h \) use the same variables for our proposed CE algorithms, so their lower bounds are equal, denoted as \( J \). The bound \( J \) will be used in the next section for analysing the CE performances of algorithms.

5 | NUMERICAL RESULTS

In this paper, the wideband multi-user mmWave communication system has been implemented in software and its performance has been obtained by means of computer simulations. The simulation parameters in this paper are set according to references [3], [23], [24]. There is \( K = 4 \) UEs, each with \( N_{\text{UE}} = 32 \) antennas, and a BS with \( N_{\text{BS}} = 128 \) antennas. We set \( M_{\text{BS}} = 4 \) at the receiver and \( M_{\text{UE}} = 1 \) at the transmitter. The dictionary matrices are constructed with \( G_{\text{BS}} = 128 \) and \( G_{\text{UE}} = 128 \) for the receiver and transmitter, respectively.

Both FVB and SS-FVB have been implemented, and their performance is compared with some other well-known CE algorithms, including OMP [30], CoSaMP [31], SW-OMP [15] and the SD-based CE algorithms [14]. In addition, the performance evaluation results for all these algorithms have been obtained by using Monte Carlo error counting techniques. The BRCB is also provided.

In accordance with the IEEE 802.11ad wireless standard, we also generate the channel (4) with the following parameters.

1. The number of the i.i.d. channel paths is \( L = 4 \).
2. The number of the delay taps is \( N_C = 4 \), and the delay of the \( l \)-path \( \tau_l \) is uniformly chosen from \( [0, (N_C - 1)T]\) with \( T_i = \frac{1}{1760} \) ms.
3. The roll-off factor of the pulse-shaping filter is \( \rho_p(\tau) = 0.8 \).

To compare the performance of various CE algorithms, the Normalised Mean Squared Error (NMSE) of \( \hat{h}(k) \) is introduced, which is expressed as

\[
\text{NMSE} = \sum_{k=1}^{K} \frac{||\hat{h}(k) - h(k)||^2_F}{||h(k)||^2_F}.
\]

The performance of different CE algorithms versus the required pilot overhead is shown in Figure 4. We can see the superiority of the proposed FVB and SS-FVB CE algorithms. The poor performance of the OMP algorithm is caused by its inability to exploit the joint sparsity shared by the delay and frequency domains. The CoSaMP method, which uses OMP, exhibits a slightly lower error when \( M \) takes small values. Although the SD-based algorithm takes advantage of the structural characteristic of mmWave beamspace channel, estimating only one path at each iteration degrades its estimation performance. The FVB and SS-FVB algorithms achieve the best CE performance because both of them make full use of the common sparsity of different sub-carriers in the frequency domain.

Figure 5 shows the NMSE performance versus SNR for various CE algorithms. The SNR is ranked from -15 to 5 dB in accordance with the realistic mmWave systems. It is noted that
the OMP and CoSaMP algorithms have a poor performance for the whole SNR range as they ignore the joint sparse vector basis. In particular, the SD-based algorithm is the worst, and it cannot work when the SNR reduces to -10 dB. Compared with other four CE schemes, both of the proposed algorithms have much higher estimation accuracy even at low SNR range, because they can effectively exploit the variational Bayesian inference.

Figure 6 shows the performance of the different competing algorithms versus the number of the random outliers. It can be observed that the compared algorithms are very sensitive to random outliers and their performance has a large degradation. More specifically, the OMP, CoSaMP and SD-based algorithms cannot guarantee the estimation accuracy when the number of the random outliers increases to 10. According to the identify-and-reject strategy, the NMSE of the proposed algorithms increases very slowly as the number of the random outliers increases. Therefore, the variational Bayesian inference and the identify-and-reject strategy are the proper means to improve the estimation accuracy and overcome the random outliers in wideband mmWave systems.

In Figure 7, we compare the NMSE versus the number of the pilot for the proposed CE algorithms as well as the BCRB. It can be seen that the NMSE of FVB algorithm is closer to
the BCRB than that of SS-FVB algorithm. This is because the SS-FVB algorithm does not exploit all the information from the support sets to estimate the common sparse channel vectors. Actually, employing a larger number of pilot can increase the estimation performance at the expense of both higher training overhead and computational complexity. Therefore, future research should consider ways of narrowing this gap and optimize the trade-offs between computational complexity and estimation performance by exploiting better recovery algorithms.

Figures 8 and 9 show the NMSE of the proposed CE algorithms versus the SNR and the random outliers, respectively. Figure 7 shows that the NMSE values obtained by FVB and SS-FVB algorithms approximate closely to the BCRB within the whole SNR range. As shown in Figure 8, the identify-and-reject strategy enhances the robustness against hardware impairments for the proposed scheme, so that their performance has a slight degradation with the variation of the random outliers.

6 | CONCLUSION

In this paper, we have proposed a novel CE scheme for wideband multi-user mmWave MIMO-OFDM systems. We have first proposed the FVB CE algorithm by exploiting the common sparsity of different sub-carriers. Considering the random outliers imposed by hardware impairments, a hierarchical channel model based on sparse Bayesian learning has been developed to improve the estimation accuracy. Besides, the power of the identify-and-reject scheme has enhanced the robustness of the FVB CE algorithm significantly. Then, the SS-FVB CE algorithm, which has developed the adaptive selection operation of the measurement matrices, has been proposed to reduce the computational complexity. This algorithm has provided the trade-off between computational complexity and estimation accuracy. Simulation results have verified that the proposed CE scheme is capable of achieving high estimation accuracy even at low SNR and pilot overhead. Future research can focus on balance the trade-offs between the computational complexity and estimation performance. It is also interesting to develop an inverse-free algorithm to reduce the computation complexity imposed by sparse Bayesian learning.

REFERENCES

1. Larsson, E.G., et al.: Massive MIMO for next generation wireless systems. IEEE Commun. Mag. 52(2), 186–195 (2014)
2. Ahmed, I., et al.: A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives. Commun. Surveys Tuts. 20(4), 3060–3097 (2018)
3. Gao, X., et al.: Energy efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays. IEEE J. Sel. Areas Commun. 34(4), 998–1009 (2016)
4. Garcia-Rodriguez, A., et al.: Hybrid analog-digital precoding revisited under realistic RF modeling. IEEE Wireless Commun. Lett. 5(5), 528–531 (2016)
5. Lin, Z., et al.: 3-D indoor positioning for millimetre-Wave massive MIMO systems. IEEE Trans. Commun. 66(6), 2472–2486 (2018)
6. Gao, H., et al.: Compressive sensing techniques for next-generation wireless communications. IEEE Wireless Commun. 25(3), 144–153 (2018)
7. Bustin, R., et al.: On MMSE crossing properties and implications in parallel vector Gaussian channels. IEEE Trans. Inf. Theory 59(2), 818–844 (2013)
8. Alkhateeb, A., et al.: Channel estimation and hybrid precoding for millimeter wave cellular systems. IEEE J. Sel. Topics Signal Process. 8(5), 831–846 (2014)
9. Xu, W., et al.: Compressive channel estimation exploiting block sparsity in multi-user massive MIMO systems. In: Proceedings IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, 1–5 (2017)
10. Cheng, X., et al.: Channel estimation for FDD multi-user massive MIMO: A variational Bayesian inference based approach. IEEE Trans. Wireless Commun. 16(11), 7590–7602 (2017)
11. Venkatesh, K., et al.: Bayesian channel estimation algorithms for massive MIMO systems with hybrid analog-digital processing and low resolution ADCs. IEEE J. Sel. Topics Signal Process. 12(3), 499–513 (2018)

12. Gao, Z., et al.: Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels. IEEE Commun. Lett. 20(6), 1259–1262 (2016)

13. Venkatesh, K., et al.: Channel estimation for hybrid architecture-based wideband millimeter wave systems. IEEE J. Sel. Areas Commun. 35(9), 1996–2009 (2017)

14. Gao, X., et al.: Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array. IEEE Trans. Wireless Commun., 16(9), 6010–6021 (2017)

15. Rodriguez-Fernandez, J., et al.: Frequency-domain compressive channel estimation for frequency-selective hybrid mmWave MIMO systems. IEEE Trans. Wireless Commun. 17(5), 2946–2960 (2018)

16. Lin, Z., et al.: 3D wideband mmWave localization for 5G massive MIMO systems. In: Proc. IEEE International Global Communications (GLOBECOM), Waikoloa, HI, USA, 1–7 (2019)

17. Zhu, D., et al.: Auxiliary beam pair enabled AoD and AoA estimation in closed-loop large-scale millimeter-wave MIMO systems. IEEE Trans. Wireless Commun. 16(7), 4770–4785 (2017)

18. Hu, C., et al.: Super-resolution channel estimation for mmWave massive MIMO with hybrid precoding. IEEE Trans. Veh. Technol. 67(9), 8954–8958 (2018)

19. Wei, C., et al.: Expectation-propagation based low-complexity channel estimation for massive MIMO systems. In: Proceedings IEEE International Conference on Communications (ICC), Paris, FR, 1–5 (2017)

20. Kamilov, US., et al.: Message-passing de-quantization with applications to compressed sensing. IEEE Trans. Signal Process. 60(12), 6270–6281 (2012)

21. Wang, P., et al.: Sparse channel estimation in millimeter wave communications: Exploiting joint AoD-AoA angular spread. In: Proceedings IEEE International Conference on Communications (ICC), Paris, FR, 1–6 (2017)

22. Lin, Y., et al.: Robust state estimation against measurement and network parameter errors. IEEE Trans. Power Syst. 33(5), 4751–4759 (2018)

23. Zhang, Z., et al.: The variational approximation for Bayesian inference. IEEE Signal Process. Mag. 25(6), 131–146 (2008)

24. Shi, L., et al.: Multipitch estimation using block sparse Bayesian learning and intra-block clustering. In: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 666–670 (2018)

25. Tropp, J.A., Gilbert, A.C.: Signal processing with random matrices. Cambridge University Press (2012)

26. Rodriguez-Fernandez, J., et al.: Variational Bayesian Channel Estimation for Wideband Millimeter Wave Systems. In: Proceedings IEEE International Conference on Communications (ICC), Paris, FR, 1–6 (2017)

27. Wen, C., et al.: Bayes-optimal joint channel-and-data estimation for massive MIMO systems with hybrid analog-digital processing. IEEE Trans. Signal Process. 60(11), 7112–7139 (2014)

28. Zhang, L., et al.: Auxiliary beam pair enabled AoD and AoA estimation in mixed antenna array. IEEE Trans. Signal Process. 66(7), 2207–2225 (2018)

29. Zeng, J., Lv, T.: Channel estimation using variational Bayesian Cramer-Rao bound for 5G massive MIMO systems. IET Commun. 15:566–579, 2021;15:566–579.

https://doi.org/10.1049/cm2.12088

APPENDIX A

A.1 Proof of Theorem 1

In this appendix, the selection operation of the measurement matrix is theoretically analysed to show its convergence to a local optimum. According to [44], a sufficient metric of convergence to a local optimum is that the energy of the residual at the (Λ−1)-th iteration is strictly larger than that of the Λ-th iteration, i.e.,

\[ \| r_{\Lambda-1} \| ^2 > \| r_{\Lambda} \| ^2, \quad \Lambda = 1, 2, \ldots, KL \]  

(A.1)

In order to select the suitable support sets [45], we first pick up one of the most correlated column from the measurement matrix to satisfy

\[ \rho_1 = \arg \max_{j=1,\ldots,N} \left| \langle y, \Psi_j \rangle \right| \]  

(A.2)
where \( \rho_1 \) represents a column index of the measurement matrix. In this way, the received signal \( \mathbf{y} \), which can be represented as

\[
\mathbf{y} = \langle \mathbf{y}, \mathbf{\Psi}_{\cdot \cdot \rho_1} \rangle \mathbf{\Psi}_{\cdot \cdot \rho_1} + \mathbf{r}_1
\]

is decomposed into the vertical projection component \( \langle \mathbf{y}, \mathbf{\Psi}_{\cdot \cdot \rho_1} \rangle \mathbf{\Psi}_{\cdot \cdot \rho_1} \) and the residual value of the closest matching atom \( \mathbf{r}_1 \). Similarly, the residual \( \mathbf{r}_\Lambda \) can also be decomposed at the \( \Lambda \)-th iteration as

\[
\mathbf{r}_{\Lambda - 1} = \langle \mathbf{r}_{\Lambda - 1}, \mathbf{\Psi}_{\cdot \cdot \rho_\Lambda} \rangle \mathbf{\Psi}_{\cdot \cdot \rho_\Lambda} + \mathbf{r}_\Lambda.
\]

After the iterations, the residuals can be negligible, and \( \mathbf{y} \) can be represented by the linear sum of the correlated columns. By the property that \( \mathbf{r}_\Lambda \) and \( \mathbf{\Psi}_{\cdot \cdot \rho_\Lambda} \) are orthogonal, one can obtain

\[
\| \mathbf{r}_{\Lambda - 1} \|^2 = \| \mathbf{r}_\Lambda \|^2 + \left| \langle \mathbf{r}_{\Lambda - 1}, \mathbf{\Psi}_{\cdot \cdot \rho_\Lambda} \rangle \right|^2.
\]

It is apparent that \( \left| \langle \mathbf{r}_{\Lambda - 1}, \mathbf{\Psi}_{\cdot \cdot \rho_\Lambda} \rangle \right|^2 \geq 0 \) and the energy of the residuals computed at the \( \Lambda \)-th iteration is strictly bigger than that of the \( \left( \Lambda - 1 \right) \)-th iteration. Thereby, the condition in (51) is satisfied and the Theorem 1 is proved.

**APPENDIX B**

**B.1 | Proof of Theorem 2**

Consider the transformations of

\[
\frac{\partial \rho(y, h)}{\partial h} = \frac{\partial \ln \rho(y, h)}{\partial h} \rho(y, h)
\]

and

\[
\frac{\partial \rho(h)}{\partial h} = \frac{\partial \ln \rho(h)}{\partial h} \rho(h).
\]

The widely accepted trivial equality

\[
\int \rho(y, h) dy = \rho(h).
\]

Then, by differentiating both sides of (58) over \( h \), it can be derived that

\[
\int \frac{\partial \ln \rho(y, h)}{\partial h} \rho(y, h) dy = \frac{\partial \ln \rho(h)}{\partial h} \rho(h).
\]

In view of Bayesian law \( \rho(y, h) = \rho(y|h) \rho(h) \), the equation above can be further written as

\[
\int \frac{\partial \ln \rho(y, h)}{\partial h} \rho(y|h) dy = \frac{\partial \ln \rho(h)}{\partial h}.
\]

Again by differentiating (60) over \( h \) again, we have

\[
\int \frac{\partial^2 \ln \rho(y, h)}{\partial h^2} \rho(y|h) dy + \int \frac{\partial^2 \ln \rho(y, h)}{\partial h^2} \rho(y|h) dy = \frac{\partial^2 \ln \rho(h)}{\partial h^2}.
\]

Moreover, it can be written as

\[
\int \frac{\partial^2 \ln \rho(y|h)}{\partial h^2} \rho(y|h) dy + \int \frac{\partial^2 \ln \rho(h)}{\partial h^2} \rho(y|h) dy + \int \frac{\partial \ln \rho(h)}{\partial h} \frac{\partial \ln \rho(y, h)}{\partial h} \rho(y|h) dy
\]

\[
= \frac{\partial^2 \ln \rho(h)}{\partial h^2}.
\]

By substituting (60) into (62), we obtain

\[
\mathbf{E} \left[ \left( \frac{\partial \ln \rho(y, h)}{\partial h} \right)^{\dagger} \left( \frac{\partial \ln \rho(h)}{\partial h} \right) \right]
\]

\[
= -\mathbf{E} \left[ \frac{\partial^2 \ln \rho(y|h)}{\partial h^2} \right] + \left( \frac{\partial \ln \rho(y, h)}{\partial h} \right)^{\dagger} \left( \frac{\partial \ln \rho(h)}{\partial h} \right).
\]

Finally, the expression shown in (46) is obtained and the Theorem 2 is proved.