Non-Uniform Burst-Sparsity Learning for Massive MIMO Channel Estimation

Jisheng Dai, *Member, IEEE*, An Liu, *Senior Member, IEEE*, and Hing Cheung So, *Fellow, IEEE*

***Abstract*—We address the downlink channel estimation problem for massive multiple-input multiple-output (MIMO) systems in this paper, where the inherit burst-sparsity structure is exploited to improve the channel estimation performance. In the literature, the commonly used burst-sparsity model assumes a uniform burst-sparse structure in which all bursts have similar sizes. However, such assumption is oversimplified to hold in practice. Outliers deviated from such uniform burst structures can significantly degrade the accuracy of the existing burst-sparsity models, which may result in a reduced recovery performance. To capture a more general burst-sparsity structure in practice, we propose a novel non-uniform burst-sparsity model and introduce an improved pattern-coupled prior to account for more realistic non-uniform burst structures. A generic sparse Bayesian learning based framework to exploit the non-uniform burst-sparsity and to enhance massive MIMO channel estimation performance is then developed. We further prove that our solution converges to a stationary point of the associated optimization problem, and our framework includes the state-of-the-art pattern-coupled method as a special case. Simulation results verify the robust performance of the devised method.**

***Index Terms*—Channel estimation, burst-sparsity, massive multiple-input multiple-output (MIMO), sparse Bayesian learning (SBL), off-grid refinement.**

I. Introduction

MASSIVE multiple-input multiple-output (MIMO) is a core candidate technology in the next generation of wireless communications due to its potential high spectrum and power efficiency [1]–[3]. To significantly improve the capacity and reliability of systems with excessive base station (BS) antennas, knowledge of channel state information at the transmitter (CSIT) is essentially required [4], [5]. However, it is challenging to acquire the accurate CSIT, since the training overhead for CSIT acquisition grows proportionally with the number of BS antennas, which can be very large in massive MIMO systems. In conventional works, time-division duplexing (TDD) mode is usually considered and then channel reciprocity can be exploited to obtain CSIT via uplink pilot training. However, due to random radio-frequency (RF) circuit mismatches in the uplink and downlink and limited coherence time, the channel reciprocity performance may degrade substantially [6], [7]. Moreover, channel reciprocity does not hold for massive MIMO systems with a frequency-division duplexing (FDD) model. Hence, CSIT acquisition for massive MIMO systems can be an extremely challenging task.

In practice, elements in the massive MIMO channel are not completely independent because of the limited local scattering effect in the propagation environment. Many studies have shown that the massive MIMO channel actually has a much lower effective dimension than its original dimension [8]–[11]. For example, the channel could have an approximately sparse representation under the discrete Fourier transform (DFT) basis if the BS is equipped with a large uniform linear array (ULA) [10], [12]–[14]. Exploiting such sparsity with the DFT basis, many compressive sensing (CS) algorithms have been proposed for downlink channel estimation and feedback [8], [10], [11], [15]–[19]. However, these DFT-based methods are applicable to ULAs only since applying the DFT basis requires a special structure of ULAs, and they always suffer from inevitable modeling error caused by direction mismatch [20].

Recently, sparse Bayesian learning (SBL) has become a popular method for sparse signal recovery problems [21]–[25], including massive MIMO channel estimation. The SBL-based framework has an inherent learning capability, and hence, no prior knowledge about the sparsity level, noise variance and/or direction mismatch is required. Moreover, it includes the l_1-norm minimization method as a special case when the maximum $a\ posteriori$ (MAP) optimal estimate is adopted with the Laplace signal prior [21], [23]. To overcome the aforementioned challenges of the DFT-based methods, an off-grid SBL-based method for downlink channel estimation with arbitrary 2D-array geometry has been suggested in [20]. Its main idea is to consider the sampled grid points in the representation basis as adjustable parameters, and then iteratively refine the grid points to minimize the modeling error caused by the direction mismatch.

However, [20] has only considered i.i.d. sparsity (i.e., the entries of the sparse channel are assumed to be i.i.d.). In practice, massive MIMO channel has more sophisticated structured sparsity that can be further exploited to enhance the channel estimation performance [19], [26]–[28]. Specifically, due to the physical scattering structure, the significant elements in the angular domain massive MIMO channel will appear in bursts. Burst-sparsity in massive MIMO channel was first exploited

Manuscript received July 3, 2018; revised October 26, 2018 and December 18, 2018; accepted December 19, 2018. Date of publication December 27, 2018; date of current version January 9, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Athanasios A. Rontogiannis. This work was supported by the National Natural Science Foundation of China under Projects 61571211 and 61571383. *(Corresponding authors: An Liu and Jisheng Dai.)*

J. Dai is with the Department of Electronic Engineering, Jiangsu University, Zhenjiang 212013, China (e-mail: jsdai@ujs.edu.cn).

A. Liu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: anliu@zju.edu.cn).

H. C. So is with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: hcso@ee.cityu.edu.hk).

Digital Object Identifier 10.1109/TSP.2018.2889952

in [19], where a burst LASSO algorithm using a block-sparse lifting transformation has been developed to successfully recover the massive MIMO channel with much fewer pilots compared to the original LASSO algorithm. However, the burst LASSO algorithm requires that all bursts have a similar size and its value must be known in advance. Recently, a structured Turbo-CS algorithm is proposed in [27] to recover burst-sparse signals, where a Markov chain prior is adopted to capture the burst-sparsity structure. Since the Turbo-CS algorithm has a similar procedure as approximate message passing [29], [30], the DFT basis is required in its theoretical derivation and state evolution performance analysis. The burst-sparsity structure of wireless channels is also utilized to enhance the channel estimation performance for millimeter wave communications in [31].

The problem of recovering burst-sparse signals from the perspective of SBL has been tackled in [32]–[36]. Note that a similar block-sparse lifting transformation is adopted in [32] to handle unknown block structures, which results in the same shortcomings as in [19]. [33] extends a "spike-and-slab" prior model to impose clustered prior on non-zero entries, but only a numerical Gibbs sampler is provided to carry out the Bayesian inference due to the troublesome modeling. In [34], unknown group sparsity is induced by organizing the corresponding scale parameters in a conditional autoregressive model. The introduced pattern-coupled SBL (PC-SBL) framework [35] (in which the sparsity of each coefficient is controlled not only by its own hyperparameter, but also by its neighbor hyperparameter) has the potential to enforce a burst-sparse solution, and it has been extended to 2D burst-sparse patterns in [37]. Moreover, a pattern-coupled channel estimation method for millimeter wave communications is proposed in [36]. Nevertheless, there are at least three disadvantages of the existing PC-SBL-based approach: (i) it only works well for separable bursts (i.e., the distance between any two adjacent bursts is sufficiently large); (ii) hyperparameter updates are not optimal in each iteration; and (iii) convergence analysis is theoretically intractable. In this paper, we devise a pattern-coupled SBL-based approach for massive MIMO downlink channel estimation, which can overcome the above shortcomings. The following summarizes the contributions of this paper.

- *Non-Uniform Burst-Sparsity Model:* We present a new non-uniform burst-sparsity model to better capture realistic sparsity structure in massive MIMO channels. The commonly used burst-sparse channel model in [19] requires channels being uniform burst-sparse with similar burst sizes in the angular domain. For burst-sparse signals with any outliers deviated from uniform burst structures, accuracy of the existing burst-sparsity models degrades significantly, which, in return, results in a reduced recovery performance. To address this issue, we propose a new non-uniform burst-sparsity model to characterize a more realistic sparse structure in practical massive MIMO channels, and then an improved pattern-coupled prior to account for the non-uniform burst sparsity is introduced.
- *Generic SBL-Based Framework for Non-Uniform Burst-Sparse Channel Estimation:* We develop a more general SBL-based method to autonomously exploit the non-uniform burst-sparsity during channel estimation. In the literature, there are many SBL-based methods that can be applied to the massive MIMO channel estimation, but few of them can handle the non-uniform burst-sparsity structure. For example, the popular pattern-coupled SBL-based method in [35] is designed for separable bursts only. Moreover, it has other drawbacks, e.g., it employs a sub-optimal solution to update the hyperparameters and cannot guarantee the convergence behavior with a theoretical analysis. In this work, we propose a generic SBL-based framework to exploit the non-uniform burst-sparsity to enhance the performance of massive MIMO channel estimation. It will be shown that our algorithm framework converges to a stationary point of the optimization problem and it includes the pattern-coupled solution in [35] as a special case by fixing some variables as sub-optimal solutions. Moreover, the grid-refining procedure used in [20] is further blended with the framework to combat the modeling error caused by direction mismatch.

The rest of the paper is organized as follows. In Section II, we present the system model and non-uniform burst-sparsity model. In Section III, the SBL-based method for recovering the massive MIMO channel with non-uniform burst-sparsity is developed. In Section IV, we introduce the grid-refining procedure to deal with the modeling error caused by direction mismatch. Numerical experiments and conclusion follow in Sections V and VI, respectively.

Notation: $\mathbb{C}$ denotes complex number, $\|\cdot\|_p$ denotes p-norm, $(\cdot)^T$ denotes transpose, $(\cdot)^H$ denotes Hermitian transpose, $\mathbf{I}$ denotes identity matrix, $\mathcal{CN}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\Gamma(\cdot|a, b)$ denotes Gamma distribution with shape parameter a and inverse scale parameter b, $\mathrm{tr}(\cdot)$ denotes trace operator, $\mathrm{diag}(\cdot)$ denotes diagonal operator, $\mathrm{Re}(\cdot)$ denotes real part operator, $\propto$ stands for equality up to a multiplicative constant, $p(\cdot)$ and $q(\cdot)$ are used to represent probability density functions for the variables of their arguments, and $\langle\cdot\rangle_{q(\cdot)}$ stands for expectation with respect to $q(\cdot)$.

II. Data Model

A. Massive MIMO Channel Model

Consider a flat block-fading massive MIMO system. There is one BS with N ($\gg 1$) antennas and K mobile users (MUs) with a single antenna. For each MU to estimate the downlink channel $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$, the BS broadcasts a sequence of T training pilot symbols $\mathbf{X} \in \mathbb{C}^{T\times N}$. Then, the downlink received signal $\mathbf{y}_k \in \mathbb{C}^{T\times 1}$ at the k-th MU is given by

$$\mathbf{y}_k = \mathbf{X}\mathbf{h}_k + \mathbf{n}_k, \tag{1}$$

where $\mathbf{n}_k \in \mathbb{C}^{T\times 1}$ stands for the additive complex i.i.d. Gaussian noise with each element having zero mean and variance σ^2 in the downlink, and $\mathrm{tr}(\mathbf{X}\mathbf{X}^H) = PTN$, with P/σ^2 measuring the training signal-to-noise ratio (SNR). If the BS is equipped

with a linear array, $\mathbf{h}_k$ can be formulated as [26], [38], [39]

$$\mathbf{h}_k = \sum_{c=1}^{N_c} \sum_{s=1}^{N_s} \xi_{c,s}^k \mathbf{a}(\theta_{c,s}^k), \tag{2}$$

where N_c stands for the number of scattering clusters, N_s is the number of sub-paths per scattering cluster, $\xi_{c,s}^k$ is the complex gain of the s-th sub-path in the c-th scattering cluster for the k-th MU, and $\theta_{c,s}^k$ and $\mathbf{a}(\theta_{c,s}^k)$ are the corresponding angle-of-departure (AoD) and steering vector. Although we fix the number of sub-paths per cluster in the channel model (2), this model allows for more general channels (i.e., with potentially different number of paths per cluster) by fixing some $\xi_{c,s}^k$s to zeros.

For ease of exposition, we focus on the downlink channel estimation problem for a reference MU, where we drop the MU's index k and denote the true AoDs as $\{\theta_l, l = 1, 2, \ldots, L\}$ with $L = N_c N_s$. Assuming that the BS is equipped with a ULA,[1] the steering vector $\mathbf{a}(\theta)$ can then be simplified as

$$\mathbf{a}(\theta) = [1, e^{-j2\pi \frac{d}{\lambda} \sin(\theta)}, \ldots, e^{-j2\pi \frac{(N-1)d}{\lambda} \sin(\theta)}]^T, \tag{3}$$

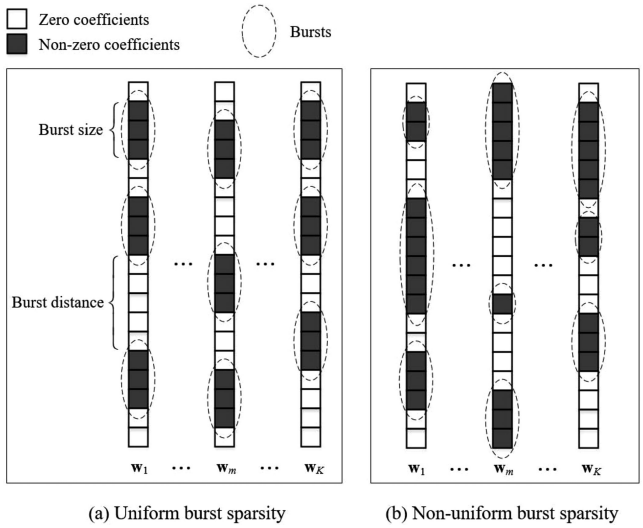
where λ is the wavelength of the downlink propagation and d stands for the distance between adjacent sensors. Let $\hat{\boldsymbol{\vartheta}} = \{\hat{\vartheta}_l\}_{l=1}^{\hat{L}}$ be a fixed sampling grid that covers the angular range $[-\pi/2, \pi/2]$, where $\hat{L}$ denotes the number of grid points. If the grid is fine enough such that the true AoDs θ_l s, $l = 1, 2, \ldots, L$, lie on the grid, we can use the following model for $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}\mathbf{A}\mathbf{w} + \mathbf{n} = \mathbf{\Phi}\mathbf{w} + \mathbf{n}, \tag{4}$$

where $\mathbf{A} = [\mathbf{a}(\hat{\vartheta}_1), \mathbf{a}(\hat{\vartheta}_2), \ldots, \mathbf{a}(\hat{\vartheta}_{\hat{L}})] \in \mathbb{C}^{N \times \hat{L}}$, $\mathbf{\Phi} = \mathbf{X}\mathbf{A}$ and $\mathbf{w} \in \mathbb{C}^{\hat{L} \times 1}$ is a vector with a few non-zero elements corresponding to the true directions at $\{\theta_l, l = 1, 2, \ldots, L\}$. As illustrated in [20], the DFT basis becomes a special case of $\mathbf{A}$ if $\hat{L} = N + 1$ and $\{\sin(\hat{\vartheta}_l)\}_{l=1}^{\hat{L}}$ uniformly covers the range $[-1, 1]$. Note that the assumption that all true AoDs are located on the predefined spatial grid is not always valid in practical implementation [24], [40]. Nevertheless, our solutions are also extended to deal with the direction mismatch, which will be discussed in detail in Section IV.

B. Non-Uniform Burst-Sparsity

According to the geometry-based stochastic channel model (GSCM) [39], the number of scattering clusters N_c is usually small and the sub-paths associated with each scattering cluster are likely to concentrate in a small range due to the limited local scattering effect in the propagation environment [8], [13], [14]. As a result, only a few elements of $\mathbf{w}$ are occupied by non-zero values and these significant elements might appear in bursts [19]. However, it is worth noting that the current burst-sparsity assumption is too strong to hold in practical scenarios, because it requires channels being uniform sparse in the angular domain. What is worse, the grid choice and the direction mismatch can further break the burst-sparsity structure. For example, consider a channel realization consists of $N_c = 2$ random scattering clusters and each cluster contains $N_s = 3$ sub-paths, whose AoDs are $\theta_{1,1} = 0°, \theta_{1,2} = 1°, \theta_{1,3} = 3°, \theta_{2,1} = 10°, \theta_{2,2} = 11°$, and $\theta_{2,3} = 12°$. If the grid is chosen as $\hat{\boldsymbol{\vartheta}} = \{-90°, -89°, -88°, \ldots, 89°, 90°\}$, the sparse representation vector $\mathbf{w}$ is not strictly burst-sparse in the angular domain.



(a) Uniform burst sparsity (b) Non-uniform burst sparsity

Fig. 1. Illustration of the uniform burst-sparsity and non-uniform burst-sparsity, where the significant elements of the sparse representation vector $\mathbf{w}_k$s appear in bursts with uniform and non-uniform burst sizes, respectively.

To capture a more realistic burst-sparsity structure, we adopt a non-uniform burst-sparsity model.

Definition 1. Non-Uniform Burst-Sparsity: The significant elements of $\mathbf{w}$ appear in bursts with possibly non-uniform burst sizes, and the burst distance can be arbitrary.[2]

Fig. 1 illustrates the difference between the uniform burst-sparsity and non-uniform burst-sparsity. Our aim is to automatically detect bursts with non-uniform sizes, and simultaneously obtain the channel estimation. It is expected to obtain more accurate channel estimation performance because we will exclude the harmful effect from outliers, i.e., bursts with very small sizes.

C. Pattern-Coupled Prior

In this subsection, we first review the existing pattern-coupled prior for burst-sparsity, as well as its challenges for the non-uniform burst-sparsity model, and a new pattern-coupled prior to better capture the non-uniform burst-sparsity structure in massive MIMO channels is then developed. In the conventional SBL framework [22], $\mathbf{w}$ is assigned a Gaussian prior distribution:

$$p(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{l=1}^{\hat{L}} \mathcal{CN}(w_l|0, \gamma_l^{-1}), \tag{5}$$

where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_{\hat{L}}]^T$ and $\mathbf{w} = [w_1, w_2, \ldots, w_{\hat{L}}]^T$ with γ_l being the precision of w_l. For tractable inference of $\boldsymbol{\gamma}$, the

[1]For clarity, we focus on the case when BS is equipped with a ULA. However, we may extend the main results to an arbitrary 2D-array geometry as in [20].

[2]The burst size stands for the number of significant elements in the burst, and the burst distance stands for the number of zero elements between two adjacent bursts.

elements of $\boldsymbol{\gamma}$ are usually modeled as independent Gamma distributions, i.e.,

$$p(\boldsymbol{\gamma}) = \prod_{l=1}^{\hat{L}} \Gamma(\gamma_l|a,b), \tag{6}$$

where a and b are some small constants (e.g., $a = b = 10^{-10}$).

It is worth noting that the precisions γ_ls directly indicate the support of $\mathbf{w}$. For example, if γ_l is large, w_l tends to zero; otherwise, the value of w_l is significant. However, the prior model in (5) assumes independence among w_ls and has no freedom to encourage burst-sparse solutions. To handle burst-sparse signals with unknown burst structures, a pattern-coupled model has been suggested in [35], in which $\mathbf{w}$ is modeled as

$$p(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{l=1}^{\hat{L}} \mathcal{CN}(w_l|0, (\beta\gamma_{l-1} + \gamma_l + \beta\gamma_{l+1})^{-1}), \tag{7}$$

where $\gamma_0 = \gamma_{\hat{L}+1} = 0$, and $0 \leq \beta \leq 1$ is a parameter indicating the pattern relevance between the coefficient w_l and its neighboring coefficients.[3] Clearly, when $\beta > 0$, the sparsity of each coefficient is controlled not only by its own hyperparameter, but also by its immediate neighbor hyperparameters; If $\beta = 0$, (7) reduces to (5).

The model (7) has the potential to enforce a burst-sparse solution. If γ_l approaches a significantly large value, w_{l-1}, w_l and w_{l+1} will decrease to zero at the same time, because their precisions are involved by the large γ_l simultaneously. However, it does not work well for sparse signals with nearby bursts. For example, let w_{l-1} and w_{l+1} be sufficiently large and w_l be zero. According to the prior defined in (7), the significant elements (w_{l-1} and w_{l+1}) require the hyperparameters ($\gamma_{(l-2)}$, $\gamma_{(l-1)}$, $\gamma_{(l)}$, $\gamma_{(l+1)}$ and $\gamma_{(l+2)}$) taking some small values. This enforces the value of w_l deviating from zero with a high possibility, because its variance $(\beta\gamma_{l-1} + \gamma_l + \beta\gamma_{l+1})^{-1}$ approaches a large value. The other shortcoming of the pattern-coupled model (7) is that it brings an intractable Bayesian inference, namely, optimal hyperparameter updates cannot be found in each iteration and convergence cannot be theoretically guaranteed. To address these issues, we will present a new pattern-coupled prior to fit the non-uniform burst-sparsity structure as follows.

Definition 2 New Pattern-Coupled Prior: Let $\mathbf{z}_l = [z_{l,1}, z_{l,2}, z_{l,3}]^T$ be an assignment vector that takes values from $\mathbf{e}_1 = [1,0,0]^T$, $\mathbf{e}_2 = [0,1,0]^T$ and $\mathbf{e}_3 = [0,0,1]^T$ with equal probability, and then we model the distribution of $\mathbf{w}$ conditional on $\mathbf{Z} = \{\mathbf{z}_l\}_{l=1}^{\hat{L}}$ and $\boldsymbol{\gamma}$ as

$$p(\mathbf{w}|\mathbf{Z},\boldsymbol{\gamma}) = \prod_{l=1}^{\hat{L}} \underbrace{\left(\left\{\mathcal{CN}(w_l|0,\gamma_{l-1}^{-1})\right\}^{z_{l,1}} \cdot \left\{\mathcal{CN}(w_l|0,\gamma_l^{-1})\right\}^{z_{l,2}} \cdot \left\{\mathcal{CN}(w_l|0,\gamma_{l+1}^{-1})\right\}^{z_{l,3}}\right)}_{\triangleq p(w_l|\mathbf{z}_l,\gamma_{l-1},\gamma_l,\gamma_{l+1})} \tag{8}$$

and the elements of $\boldsymbol{\gamma}$ are similarly modeled as in (6).

[3]How to select β for the pattern-coupled model (7) is still an open problem. According to the empirical evidence provided in [35], we set $\beta = 1$ for PC-SBL in the simulations.

Note that the prior distribution of $\mathbf{z}_l$ can be formulated as a non-informative categorical distribution:[4]

$$p(\mathbf{z}_l) = \left(\frac{1}{3}\right)^{[\mathbf{z}_l=\mathbf{e}_1]} \left(\frac{1}{3}\right)^{[\mathbf{z}_l=\mathbf{e}_2]} \left(\frac{1}{3}\right)^{[\mathbf{z}_l=\mathbf{e}_3]},$$

or, equivalently,

$$p(\mathbf{z}_l) = \left(\frac{1}{3}\right)^{z_{l,1}} \left(\frac{1}{3}\right)^{z_{l,2}} \left(\frac{1}{3}\right)^{z_{l,3}}, \tag{9}$$

where $\mathbf{z}_l \in \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$. Since only one element of $\mathbf{z}_l$ is activated on a single trial, the new pattern-coupled prior (8) can enforce the burst-sparsity of $\mathbf{w}$ and handle outliers simultaneously. For each w_l, if its neighbor w_{l-1} (or w_{l+1}) has a significant value, its value will also be significant with taking $\mathbf{z}_l = \mathbf{e}_1$ (or $\mathbf{z}_l = \mathbf{e}_3$); while for an outlier w_l (i.e., w_{l-1} and w_{l+1} have significant values, but w_l does not), (8) still works, because it can take $\mathbf{z}_l = \mathbf{e}_2$ and adopt its own hyperparameter to account for outliers.

Although the forms of (8) and (7) are quite distinct, we will show that the Bayesian inference for our model includes the one for (7) as a special case, which will be detailed in Section III-D.

III. Non-Uniform Burst-Sparse Channel Estimation

In this section, a general SBL-based method to autonomously exploit the non-uniform burst-sparsity during channel estimation is developed. For ease of exposition, we begin by introducing the SBL formulation for non-uniform burst-sparse signal recovery. Then, we resort to the variational Bayesian inference (VBI) methodology [41] and propose an alternating update algorithm to jointly exploit the non-uniform burst-sparsity and estimate the channel. Finally, a parameterized transformation of the proposed algorithm is provided, together with its convergence analysis. Our solutions can be extended to cope with the modeling error caused by direction mismatch, which will be addressed in the next section.

A. Overview of Proposed Method

Under the assumption of circular symmetric complex Gaussian noise, we have

$$p(\mathbf{y}|\mathbf{w},\alpha) = \mathcal{CN}(\mathbf{y}|\boldsymbol{\Phi}\mathbf{w}, \alpha^{-1}\mathbf{I}), \tag{10}$$

where $\alpha = \sigma^{-2}$ stands for the noise precision, which is usually unknown. Hence, it is modeled as a Gamma hyperprior $p(\alpha) = \Gamma(\alpha|a,b)$, where a and b are defined in (6). Obviously, the hidden variables that need to be estimated are $\boldsymbol{\Theta} \triangleq \{\alpha, \mathbf{w}, \boldsymbol{\gamma}, \mathbf{Z}\}$. Based on the above hierarchical model, the joint distribution of variables is expressed as

$$p(\mathbf{y}, \boldsymbol{\Theta}) = p(\mathbf{y}|\mathbf{w},\alpha)p(\mathbf{w}|\mathbf{Z},\boldsymbol{\gamma})p(\alpha)p(\boldsymbol{\gamma})p(\mathbf{Z}). \tag{11}$$

[4]It is easy to extend our method to more complicated priors for $\mathbf{z}_l$. For example, we may consider a hierarchical prior $p(\mathbf{z}_l|\boldsymbol{\rho}) = \rho_1^{z_{l,1}} \rho_2^{z_{l,2}} \rho_3^{z_{l,3}}$, where $\boldsymbol{\rho} \triangleq [\rho_1, \rho_2, \rho_3]^T$ and ρ_gs are unknown non-negative constants satisfying $\sum_{g=1}^{3} \rho_g = 1$. However, empirical evidence shows that the non-informative prior and hierarchical priors give very similar performance.

Note that the non-uniform burst-sparsity structure and channel estimation are jointly obtained if we can calculate the MAP estimate of $\boldsymbol{\Theta}$ (given $\mathbf{y}$). Specifically, the non-uniform burst-sparsity structure is indicated by the MAP estimator of the assignment vectors $\mathbf{z}_l$ s, and the sparse representation channel vector can be calculated from the MAP estimator of $\mathbf{w}$. Unfortunately, this MAP estimate is intractable.

To make the MAP estimate tractable, we resort to the VBI methodology to find an approximate posterior denoted by $q(\boldsymbol{\Theta})$, instead of calculating the posterior $p(\boldsymbol{\Theta}|\mathbf{y})$ exactly. Let $q(\boldsymbol{\Theta})$ be factorized approximately as

$$q(\boldsymbol{\Theta}) = q(\alpha)q(\mathbf{w})q(\boldsymbol{\gamma})q(\mathbf{Z}) \tag{12}$$

and then the corresponding optimization problem is to find the "best" approximate posterior under the factorized constraint in (12). In other words, the factorization of $q(\boldsymbol{\Theta})$ should be chosen to minimize the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\mathbf{y})) = -\int q(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{\Theta}|\mathbf{y})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta}, \tag{13}$$

or, equivalently,

$$q^\star(\boldsymbol{\Theta}) = \arg\max_{q(\boldsymbol{\Theta})} \underbrace{\int q(\boldsymbol{\Theta}) \ln \frac{p(\mathbf{y},\boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta}}_{\triangleq \mathcal{U}(q(\Theta_1),q(\Theta_2),q(\Theta_3),q(\Theta_4))}, \tag{14}$$

where Θ_n stands for the n-th element in $\boldsymbol{\Theta}$. As shown in [41], the optimal distribution to (14) must satisfy the following equation

$$\ln q^\star(\Theta_n) = \langle \ln p(\mathbf{y},\boldsymbol{\Theta}) \rangle_{\prod_{i\neq n} q^\star(\Theta_i)} + \text{const.}, \; n = 1,2,3,4. \tag{15}$$

Since the solution $q^\star(\Theta_n)$ given in (15) is dependent on other solutions $q^\star(\Theta_j), j \neq n$, it is difficult to find the optimal solution in closed-form. Here, we adopt an alternating update algorithm to find a stationary solution instead. Specifically, $q(\alpha)$, $q(\mathbf{w})$, $q(\boldsymbol{\gamma})$ and $q(\mathbf{Z})$ are iteratively updated as:

$$q^{(i+1)}(\alpha) \propto \langle \ln p(\mathbf{y},\boldsymbol{\Theta}) \rangle_{q^{(i)}(\mathbf{w})q^{(i)}(\boldsymbol{\gamma})q^{(i)}(\mathbf{Z})}, \tag{16}$$

$$q^{(i+1)}(\mathbf{w}) \propto \langle \ln p(\mathbf{y},\boldsymbol{\Theta}) \rangle_{q^{(i+1)}(\alpha)q^{(i)}(\boldsymbol{\gamma})q^{(i)}(\mathbf{Z})}, \tag{17}$$

$$q^{(i+1)}(\boldsymbol{\gamma}) \propto \langle \ln p(\mathbf{y},\boldsymbol{\Theta}) \rangle_{q^{(i+1)}(\alpha)q^{(i+1)}(\mathbf{w})q^{(i)}(\mathbf{Z})}, \tag{18}$$

$$q^{(i+1)}(\mathbf{Z}) \propto \langle \ln p(\mathbf{y},\boldsymbol{\Theta}) \rangle_{q^{(i+1)}(\alpha)q^{(i+1)}(\mathbf{w})q^{(i+1)}(\boldsymbol{\gamma})}, \tag{19}$$

where $(\cdot)^{(i)}$ stands for the i-th iteration. In the following, we first discuss how to solve (16)–(19). It is then revealed that they can be transformed into parameterized problems. Finally, a convergence analysis of the proposed algorithm based on the parameterized transformation is provided.

B. Detailed Updates

In this subsection, we address the updates (16)–(19) in detail.

1) Update of $q(\alpha)$: The update (16) gives a unique solution

$$\begin{aligned}
&\ln q^{(i+1)}(\alpha) \\
&\propto \langle \ln p(\mathbf{y}|\mathbf{w},\alpha) \rangle_{q^{(i)}(\mathbf{w})} + \ln p(\alpha) \\
&\propto (\underbrace{a + T}_{\triangleq a_\alpha^{(i+1)}} - 1) \ln \alpha \\
&\quad - \alpha \cdot \underbrace{\left(b + \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}^{(i)}\|_2^2 + \mathrm{tr}(\boldsymbol{\Phi}\boldsymbol{\Sigma}^{(i)}\boldsymbol{\Phi}^H) \right)}_{b_\alpha^{(i+1)}},
\end{aligned} \tag{20}$$

where $\boldsymbol{\mu}^{(i)} \triangleq \langle \mathbf{w} \rangle_{q^{(i)}(\mathbf{w})}$ and $\boldsymbol{\Sigma}^{(i)} \triangleq \langle (\mathbf{w} - \boldsymbol{\mu}^{(i)})(\mathbf{w} - \boldsymbol{\mu}^{(i)})^H \rangle_{q^{(i)}(\mathbf{w})}$ (whose closed-form expressions will be given latter). Hence, $q^{(i+1)}(\alpha)$ obeys a Gamma distribution:

$$q^{(i+1)}(\alpha) = \Gamma(\alpha | a_\alpha^{(i+1)}, b_\alpha^{(i+1)}) \tag{21}$$

and the mean of α is

$$\hat{\alpha}^{(i+1)} \triangleq \langle \alpha \rangle_{q^{(i+1)}(\alpha)} = \frac{a_\alpha^{(i+1)}}{b_\alpha^{(i+1)}}. \tag{22}$$

2) Update of $q(\mathbf{w})$: The update (17) gives a unique solution

$$\begin{aligned}
&\ln q^{(i+1)}(\mathbf{w}) \\
&\propto \langle \ln p(\mathbf{y},\boldsymbol{\Theta}) \rangle_{q^{(i+1)}(\alpha)q^{(i)}(\boldsymbol{\gamma})q^{(i)}(\mathbf{Z})} && (23) \\
&\propto \langle \ln p(\mathbf{y}|\mathbf{w},\alpha) \rangle_{q^{(i+1)}(\alpha)} + \langle \ln p(\mathbf{w}|\mathbf{Z},\boldsymbol{\gamma}) \rangle_{q^{(i)}(\boldsymbol{\gamma})q^{(i)}(\mathbf{Z})} && (24) \\
&\propto -\hat{\alpha}^{(i+1)} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 \\
&\quad - \mathbf{w}^H (\boldsymbol{\Psi}_1^{(i)} \boldsymbol{\Lambda}_R^{(i)} + \boldsymbol{\Psi}_2^{(i)} \boldsymbol{\Lambda}^{(i)} + \boldsymbol{\Psi}_3^{(i)} \boldsymbol{\Lambda}_L^{(i)}) \mathbf{w}, && (25)
\end{aligned}$$

where $\boldsymbol{\Psi}_g^{(i)} \triangleq \mathrm{diag}\{\phi_{1,g}^{(i)}, \phi_{2,g}^{(i)} \ldots, \phi_{\hat{L},g}^{(i)}\}$ with $\phi_{l,g}^{(i)} \triangleq q^{(i)}(\mathbf{z}_l = \mathbf{e}_g)$, $g = 1,2,3$, $\boldsymbol{\Lambda}^{(i)} \triangleq \mathrm{diag}\{\gamma_1^{(i)}, \gamma_2^{(i)}, \ldots, \gamma_{\hat{L}}^{(i)}\}$, $\boldsymbol{\Lambda}_R^{(i)} \triangleq \mathrm{diag}\{\gamma_{\hat{L}}^{(i)}, \gamma_1^{(i)}, \gamma_2^{(i)}, \ldots, \gamma_{\hat{L}-1}^{(i)}\}$, and $\boldsymbol{\Lambda}_L^{(i)} \triangleq \mathrm{diag}\{\gamma_2^{(i)}, \gamma_3^{(i)}, \ldots, \gamma_{\hat{L}}^{(i)}, \gamma_1^{(i)}\}$. This equality shows that $q^{(i+1)}(\mathbf{w})$ follows a Gaussian distribution:

$$q^{(i+1)}(\mathbf{w}) = \mathcal{CN}(\mathbf{w}|\boldsymbol{\mu}^{(i+1)}, \boldsymbol{\Sigma}^{(i+1)}), \tag{26}$$

where $\boldsymbol{\mu}^{(i+1)} = \hat{\alpha}^{(i+1)} \boldsymbol{\Sigma}^{(i+1)} \boldsymbol{\Phi}^H \mathbf{y}$ and $\boldsymbol{\Sigma}^{(i+1)} = (\hat{\alpha}^{(i+1)} \boldsymbol{\Phi}^H \boldsymbol{\Phi} + \boldsymbol{\Psi}_1^{(i)} \boldsymbol{\Lambda}_R^{(i)} + \boldsymbol{\Psi}_2^{(i)} \boldsymbol{\Lambda}^{(i)} + \boldsymbol{\Psi}_3^{(i)} \boldsymbol{\Lambda}_L^{(i)})^{-1}$.

3) Update of $q(\boldsymbol{\gamma})$: The update (18) gives a unique solution

$$\begin{aligned}
&\ln q^{(i+1)}(\boldsymbol{\gamma}) \\
&\propto \langle \ln p(\mathbf{y},\boldsymbol{\Theta}) \rangle_{q^{(i+1)}(\alpha)q^{(i+1)}(\mathbf{w})q^{(i)}(\mathbf{Z})} && (27) \\
&\propto \langle \ln p(\mathbf{w}|\mathbf{Z},\boldsymbol{\gamma}) \rangle_{q^{(i+1)}(\mathbf{w})q^{(i)}(\mathbf{Z})} + \ln p(\boldsymbol{\gamma}) && (28) \\
&\propto \sum_{l=1}^{\hat{L}} \left\{ \left(\underbrace{a + \phi_{l+1,1}^{(i)} + \phi_{l,2}^{(i)} + \phi_{l-1,3}^{(i)}}_{\triangleq a_l^{(i+1)}} - 1 \right) \ln \gamma_l \right. \\
&\quad \left. - \gamma_l \cdot \underbrace{\left(b + \phi_{l+1,1}^{(i)} \varpi_{l+1}^{(i+1)} + \phi_{l,2}^{(i)} \varpi_l^{(i+1)} + \phi_{l-1,3}^{(i)} \varpi_{l-1}^{(i+1)} \right)}_{\triangleq b_l^{(i+1)}} \right\}, && (29)
\end{aligned}$$

where $\varpi_l^{(i+1)} \triangleq |\mu_l^{(i+1)}|^2 + \Sigma_{l,l}^{(i+1)}$ with $\mu_l^{(i+1)}$ being the l-th element of $\boldsymbol{\mu}^{(i+1)}$ and $\Sigma_{l,l}^{(i+1)}$ being the l-th diagonal element of $\boldsymbol{\Sigma}^{(i+1)}$. Note that we let the undefined μ_0 and $\mu_{\hat{L}+1}$ equal $\mu_{\hat{L}}$ and μ_1, respectively, so are $\Sigma_{0,0}$, $\Sigma_{\hat{L}+1,\hat{L}+1}$, $\phi_{0,g}$ and $\phi_{\hat{L}+1,g}$. Since γ_l is separable for each other, $q^{(i+1)}(\gamma_l)$ obeys a Gamma distribution:

$$q^{(i+1)}(\gamma_l) = \Gamma\left(\gamma_l | a_l^{(i+1)}, b_l^{(i+1)}\right). \tag{30}$$

The means of γ_l and $\ln \gamma_l$ can be calculated as

$$\hat{\gamma}_l^{(i+1)} \triangleq \langle \gamma_l \rangle_{q^{(i+1)}(\gamma_l)} = \frac{a_l^{(i+1)}}{b_l^{(i+1)}} \tag{31}$$

and

$$\begin{aligned}(\widehat{\ln \gamma_l})^{(i+1)} &\triangleq \langle \ln \gamma_l \rangle_{q^{(i+1)}(\gamma_l)} \\ &= \Psi\left(a_l^{(i+1)}\right) - \ln\left(b_l^{(i+1)}\right),\end{aligned} \tag{32}$$

respectively, where $\Psi(\cdot)$ stands for the digamma function.

4) Update of $q(\mathbf{Z})$: The update (19) gives a unique solution

$$\begin{aligned}&\ln q^{(i+1)}(\mathbf{Z}) \\ &\propto \langle \ln p(\mathbf{y}, \boldsymbol{\Theta}) \rangle_{q^{(i+1)}(\alpha) q^{(i+1)}(\mathbf{w}) q^{(i+1)}(\boldsymbol{\gamma})} \qquad (33) \\ &\propto \langle \ln p(\mathbf{w}|\mathbf{Z}, \boldsymbol{\gamma}) \rangle_{q^{(i+1)}(\mathbf{w}) q^{(i+1)}(\boldsymbol{\gamma})}. \qquad (34)\end{aligned}$$

Since each $\mathbf{z}_l$ is a discrete vector, we are able to exhaustively calculate the value of $\ln q^{(i+1)}(\mathbf{z}_l = \mathbf{e}_g), \forall g$, as

$$\ln q^{(i+1)}(\mathbf{z}_l = \mathbf{e}_1) \propto \underbrace{(\widehat{\ln \gamma_{l-1}})^{(i+1)} - \hat{\gamma}_{l-1}^{(i+1)} \varpi_l^{(i+1)}}_{\triangleq \varsigma_{l,1}^{(i+1)}}, \tag{35}$$

$$\ln q^{(i+1)}(\mathbf{z}_l = \mathbf{e}_2) \propto \underbrace{(\widehat{\ln \gamma_l})^{(i+1)} - \hat{\gamma}_l^{(i+1)} \varpi_l^{(i+1)}}_{\triangleq \varsigma_{l,2}^{(i+1)}}, \tag{36}$$

$$\ln q^{(i+1)}(\mathbf{z}_l = \mathbf{e}_3) \propto \underbrace{(\widehat{\ln \gamma_{l+1}})^{(i+1)} - \hat{\gamma}_{l+1}^{(i+1)} \varpi_l^{(i+1)}}_{\triangleq \varsigma_{l,3}^{(i+1)}}. \tag{37}$$

Because $\sum_{g=1}^{3} q^{(i+1)}(\mathbf{z}_l = \mathbf{e}_g) = 1$, we obtain

$$\phi_{l,g}^{(i+1)} = q^{(i+1)}(\mathbf{z}_l = \mathbf{e}_g) = \frac{\exp(\varsigma_{l,g}^{(i+1)})}{\sum_{g=1}^{3} \exp(\varsigma_{l,g}^{(i+1)})}. \tag{38}$$

C. Parameterized Transformation and Convergence Analysis

From Section III-B, it is clear that each factor in $q(\boldsymbol{\Theta}) = q(\alpha) q(\mathbf{w}) q(\boldsymbol{\gamma}) q(\mathbf{Z})$ can be considered as a parameterized function. Specifically, (21) shows that $q(\alpha)$ is in a form of Gamma distribution parameterized by $\Omega_1 \triangleq \{a_\alpha, b_\alpha\}$; (26) indicates that $q(\mathbf{w})$ is Gaussian distributed, parameterized by $\Omega_2 \triangleq \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$; (30) shows that $q(\boldsymbol{\gamma})$ is in a form of Gamma distribution parameterized by $\Omega_3 \triangleq \{a_l, b_l\}_{l=1}^{\hat{L}}$; (38) means that $q(\mathbf{Z})$ is a discrete distribution parameterized by $\Omega_4 \triangleq \{\phi_{l,g}\}_{l=1,G=1}^{\hat{L},3}$. As a result, the optimization problem (14) which is optimized over function spaces can be converted into a conventional parameterized optimization problem:

$$\begin{aligned}&(\Omega_1^\star, \Omega_2^\star, \Omega_3^\star, \Omega_4^\star) \\ &\quad = \arg \max_{\Omega_1, \Omega_2, \Omega_3, \Omega_4} \mathcal{U}(\Omega_1, \Omega_2, \Omega_3, \Omega_4),\end{aligned} \tag{39}$$

where $\mathcal{U}(\Omega_1, \Omega_2, \Omega_3, \Omega_4)$ is short for $\mathcal{U}(q(\alpha|\Omega_1), q(\mathbf{w}|\Omega_2), q(\boldsymbol{\gamma}|\Omega_3), q(\mathbf{Z}|\Omega_4))$ whose definition can be found in (14). Then, the alternating update algorithm is considered as an alternating optimization (AO) approach [42], [43]

$$\Omega_1^{(i+1)} = \arg \max_{\Omega_1} \mathcal{U}\left(\Omega_1, \Omega_2^{(i)}, \Omega_3^{(i)}, \Omega_4^{(i)}\right), \tag{40}$$

$$\Omega_2^{(i+1)} = \arg \max_{\Omega_2} \mathcal{U}\left(\Omega_1^{(i+1)}, \Omega_2, \Omega_3^{(i)}, \Omega_4^{(i)}\right), \tag{41}$$

$$\Omega_3^{(i+1)} = \arg \max_{\Omega_3} \mathcal{U}\left(\Omega_1^{(i+1)}, \Omega_2^{(i+1)}, \Omega_3, \Omega_4^{(i)}\right), \tag{42}$$

$$\Omega_4^{(i+1)} = \arg \max_{\Omega_4} \mathcal{U}\left(\Omega_1^{(i+1)}, \Omega_2^{(i+1)}, \Omega_3^{(i+1)}, \Omega_4\right). \tag{43}$$

Note that the solutions to (40)–(43) almost coincide with the ones to (16)–(19), which can be found in (21), (26), (30) and (38), respectively. The only difference is in that the parameterized method provides the parameters of the posterior distributions, while the standard method gives the whole posterior distributions. Once the algorithm converges, the approximate posteriors $q(\alpha)$, $q(\mathbf{w})$, $q(\boldsymbol{\gamma})$ and $q(\mathbf{Z})$ are obtained. Then, we use the mean of the posterior $q(\mathbf{w})$ as the estimate of $\mathbf{w}$, and the estimated downlink channels $\mathbf{h}^e$ is calculated as

$$\mathbf{h}^e = \mathbf{A}\boldsymbol{\mu}. \tag{44}$$

where $\boldsymbol{\mu}$ is the posterior mean of the angular channel vector $\mathbf{w}$.

It is worth noting that to trigger the AO algorithm (40)–(43), initialization for Ω_2^0, Ω_3^0 and Ω_4^0 is needed. Empirical evidence shows that the proposed method remains very robust to the choice of initial guesses. Therefore, we simply do:

- Ω_2^0 is initialized with $\boldsymbol{\mu}^{(0)} = \boldsymbol{\Sigma}^{(0)} \boldsymbol{\Phi}^H \mathbf{y}$ and $\boldsymbol{\Sigma}_k^{(0)} = (\boldsymbol{\Phi}^H \boldsymbol{\Phi} + \mathbf{I})^{-1}$;
- Ω_3^0 is initialized with $a_l^{(0)} = b_l^{(0)} = 1, \forall l$;
- Ω_4^0 is initialized with random $\phi_{l,g}^{(0)}$s such that $\sum_{g=1}^{3} \phi_{l,g}^{(0)} = 1, \forall l$.

The steps of the proposed algorithm are summarized in Fig. 2.

Finally, we give a convergence analysis for our method. Since the original optimization problem (14) is equal to the parameterized optimization problem (39), we focus on the parameterized optimization problem as follows. The non-decreasing property of the sequence $\mathcal{U}(\Omega_1^{(i)}, \Omega_2^{(i)}, \Omega_3^{(i)}, \Omega_4^{(i)})$, $i = 1, 2, 3, \ldots$, is well guaranteed by the update rules (40)–(43). Since $\mathcal{U}(\Omega_1, \Omega_2, \Omega_3, \Omega_4)$ has an upper bound, the sequence $\mathcal{U}(\Omega_1^{(i)}, \Omega_2^{(i)}, \Omega_3^{(i)}, \Omega_4^{(i)})$, $i = 1, 2, 3, \ldots$, converges to a limit. Moreover, each subproblem in (40)–(43) has a unique solution. Hence, according to Theorem 2-b in [42], we further establish that the limit of the objective solution is an exact stationary point.

Lemma 3: If variables are iteratively updated by solving (40), (41), (42) and (43), the iterates generated by the AO

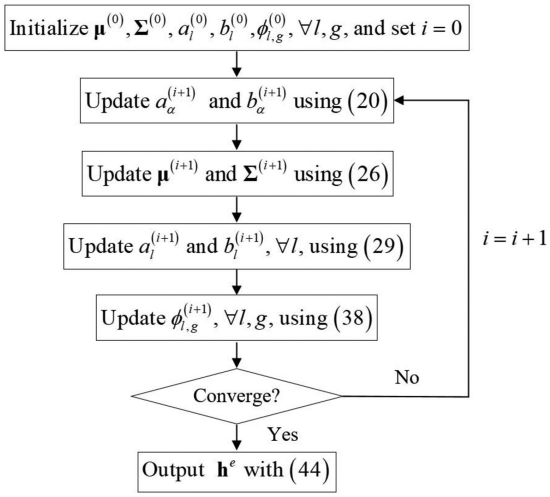


Fig. 2. Procedure of proposed algorithm.

algorithm converge to a stationary solution of the optimization problem (39).

Proof: See Appendix F in [20]. ■

D. Relation With [35]

As mentioned in Section II-C, the method in [35] adopts the pattern-coupled model (7) to exploit burst-sparsity, which brings an intractable Bayesian inference. Since the optimal hyperparameter update for $\hat{\gamma}_l$s cannot be found in each iteration, only a sub-optimal solution is available (see (26) in [35]), i.e.,

$$\hat{\gamma}_l^{(i+1)} = \frac{\bar{a}}{0.5(\beta\varpi_{l-1}^{(i+1)} + \varpi_l^{(i+1)} + \beta\varpi_{l+1}^{(i+1)}) + \bar{b}}, \forall l, \quad (45)$$

where $\bar{a} > 0$ and $\bar{b}$ is a small constant (e.g., $\bar{b} = 10^{-10}$). Note that such sub-optimal choice may degrade the recovery performance, and its convergence cannot be theoretically guaranteed. Different from [35], our method has several advantages: (i) it is able to cope with the non-uniform burst-sparsity model which can capture a more complicated and realistic burst-sparsity structure; and (ii) it overcomes the aforementioned algorithmic shortcomings encountered by the existing pattern-coupled method [35]. Actually, our update for $\hat{\gamma}_l$s in (31) includes the sub-optimal one in (45) as a special case.

Lemma 4: The proposed method includes the existing pattern-coupled method as a special case if we use fixed values for the posterior estimates of $\mathbf{z}_l$s, i.e., $\phi_{l,1}^{(i)} = \phi_{l,3}^{(i)} = \beta, \phi_{l,2}^{(i)} = 1$, $\forall l, i$.

Proof: Let $\phi_{l,1}^{(i)} = \phi_{l,3}^{(i)} = \beta$, and $\phi_{l,2}^{(i)} = 1, \forall l, i$, then (31) can be written as

$$\begin{aligned}\hat{\gamma}_l^{(i+1)} &= \frac{a + 1 + 2\beta}{b + \beta\varpi_{l-1}^{(i+1)} + \varpi_l^{(i+1)} + \beta\varpi_{l+1}^{(i+1)}} \\ &\approx \frac{1 + 2\beta}{\beta\varpi_{l-1}^{(i+1)} + \varpi_l^{(i+1)} + \beta\varpi_{l+1}^{(i+1)}}, \end{aligned} \quad (46)$$

because a and b are sufficiently small. On the other hand, (45) is similarly written as

$$\hat{\gamma}_l^{(i+1)} \approx \frac{2\bar{a}}{\beta\varpi_{l-1}^{(i+1)} + \varpi_l^{(i+1)} + \beta\varpi_{l+1}^{(i+1)}}. \quad (47)$$

The only difference between (46) and (47) in the numerators. However, this gap can be fixed by scaling the dictionary matrix or the measurement directly. For example, if we use the dictionary matrix $\hat{\boldsymbol{\Phi}} \triangleq \sqrt{\frac{2\bar{a}}{1+2\beta}} \cdot \boldsymbol{\Phi}$ instead of $\boldsymbol{\Phi}$ in (4), the corresponding update (46) will coincide with (47). ■

From Lemma 4, the proposed method is expected to outperform [35], because our pattern-coupled coefficients are automatically determined by the Bayesian inference, which will bring enhanced recovery performance. But it should be noted that Lemma 4 is just a byproduct. The core principle behind our solution is the new pattern-coupled prior introduced in Definition 2, which makes the Bayesian inference tractable for the non-uniform burst-sparsity model. On the other hand, since the new model introduces additional assignment vectors $\mathbf{z}_l$s into the SBL framework, the proposed method might suffer from a risk of overfitting. However, empirical evidence shows that the proposed method remains very robust to outliers and works quite well with practical channel models.

IV. Handling Direction Mismatch

In practical scenarios, signals usually come from random directions. Therefore, as mentioned in Section II-A, the assumption of true AoDs being located on the predefined spatial grid may not be valid. To solve the direction mismatch problem, off-grid models have been applied widely to direction-of-arrival estimation in array signal processing [24], [44]. The commonly used first-order linear approximation model does not work well when the grid is not sufficiently fine [40]. The direction mismatch problem for the massive MIMO channel estimation has been investigated in [20], and a dynamic off-grid model is developed to avoid using any approximations so as to significantly alleviate the modeling error. Note that the main idea of the dynamic off-grid model is to consider the sampled grid points as adjustable parameters, which has also been adopted in [45]–[48]. In this section, we blend the off-grid model proposed in [20] with our SBL-based framework to combat the modeling error caused by direction mismatch. Specifically, if $\theta_l \notin \{\hat{\vartheta}_i\}_{i=1}^{\hat{L}}$ and $\hat{\vartheta}_{n_l}, n_l \in \{1, 2, \ldots, \hat{L}\}$, is the nearest grid point to θ_l, θ_l is written as:

$$\theta_l = \hat{\vartheta}_{n_l} + \beta_{n_l}, \quad (48)$$

where β_{n_l} corresponds to the direction mismatch (or off-grid gap). From (48), we have $\mathbf{a}(\theta_l) = \mathbf{a}(\hat{\vartheta}_{n_l} + \beta_{n_l})$, and the received signal $\mathbf{y}$ can be rewritten as

$$\mathbf{y} = \boldsymbol{\Phi}(\boldsymbol{\beta})\mathbf{w} + \mathbf{n}, \quad (49)$$

where $\boldsymbol{\Phi}(\boldsymbol{\beta}) = \mathbf{X}\mathbf{A}(\boldsymbol{\beta})$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{\hat{L}}]^T$, $\mathbf{A}(\boldsymbol{\beta}) = [\mathbf{a}(\hat{\vartheta}_1 + \beta_1), \mathbf{a}(\hat{\vartheta}_2 + \beta_2), \ldots, \mathbf{a}(\hat{\vartheta}_{\hat{L}} + \beta_{\hat{L}})]$, and

$$\beta_{n_l} = \begin{cases} \theta_l - \hat{\vartheta}_{n_l}, & l = 1, 2, \ldots, L \\ 0, & \text{otherwise} \end{cases}.$$

Clearly, the direction mismatch can be significantly alleviated because there always exists some β_{k,n_l} making (48) hold exactly.

With the off-grid model (49), almost all the results in Section III-A remain unchanged, except that (10) is replaced by

$$p(\mathbf{y}|\mathbf{w}, \alpha, \boldsymbol{\beta}) = \mathcal{CN}(\mathbf{y}|\boldsymbol{\Phi}(\boldsymbol{\beta})\mathbf{w}, \alpha^{-1}\mathbf{I}), \tag{50}$$

and the parameterized problem (39) is modified to

$$(\Omega_1^\star, \Omega_2^\star, \Omega_3^\star, \Omega_4^\star, \boldsymbol{\beta}) = \arg \max_{\Omega_1, \Omega_2, \Omega_3, \Omega_4, \boldsymbol{\beta}} \mathcal{U}(\Omega_1, \Omega_2, \Omega_3, \Omega_4, \boldsymbol{\beta}). \tag{51}$$

Then, the corresponding AO algorithm becomes

$$\Omega_1^{(i+1)} = \arg \max_{\Omega_1} \mathcal{U}\left(\Omega_1, \Omega_2^{(i)}, \Omega_3^{(i)}, \Omega_4^{(i)}, \boldsymbol{\beta}^{(i)}\right), \tag{52}$$

$$\Omega_2^{(i+1)} = \arg \max_{\Omega_2} \mathcal{U}\left(\Omega_1^{(i+1)}, \Omega_2, \Omega_3^{(i)}, \Omega_4^{(i)}, \boldsymbol{\beta}^{(i)}\right), \tag{53}$$

$$\Omega_3^{(i+1)} = \arg \max_{\Omega_3} \mathcal{U}\left(\Omega_1^{(i+1)}, \Omega_2^{(i+1)}, \Omega_3, \Omega_4^{(i)}, \boldsymbol{\beta}^{(i)}\right), \tag{54}$$

$$\Omega_4^{(i+1)} = \arg \max_{\Omega_4} \mathcal{U}\left(\Omega_1^{(i+1)}, \Omega_2^{(i+1)}, \Omega_3^{(i+1)}, \Omega_4, \boldsymbol{\beta}^{(i)}\right), \tag{55}$$

$$\boldsymbol{\beta}^{(i+1)} = \arg \max_{\boldsymbol{\beta}} \mathcal{U}\left(\Omega_1^{(i+1)}, \Omega_2^{(i+1)}, \Omega_3^{(i+1)}, \Omega_4^{(i+1)}, \boldsymbol{\beta}\right). \tag{56}$$

The solutions to (52)–(55) can be achieved similarly as in Section III-B, where the only difference is in replacing $\boldsymbol{\Phi}$ by $\boldsymbol{\Phi}(\boldsymbol{\beta})$. What remains is to obtain the solution to (56). Since the same optimization problem has been addressed in (33) of [20], we provide the main update result for $\boldsymbol{\beta}$ here, but without any derivations. The interested reader is referred to Section III-D of [20].

Following the procedure in [20], we apply gradient update on the objective function of (56) and obtain a simple one-step update for $\boldsymbol{\beta}$s, where the derivative of the objective function, with respect to, $\boldsymbol{\beta}$, is calculated as

$$\boldsymbol{\zeta}_{\boldsymbol{\beta}}^{(i+1)} = [\zeta^{(i+1)}(\beta_1), \zeta^{(i+1)}(\beta_2), \ldots, \zeta^{(i+1)}(\beta_{\hat{L}})]^T, \tag{57}$$

with

$$\zeta^{(i+1)}(\beta_l) = 2\mathrm{Re}\left((\mathbf{a}'(\hat{\vartheta}_l + \beta_l))^H \mathbf{X}^H \mathbf{X}(\mathbf{a}(\hat{\vartheta}_l + \beta_l))\right) \cdot c_1^{(i+1)} + 2\mathrm{Re}\left((\mathbf{a}'(\hat{\vartheta}_l + \beta_l))^H \mathbf{X}^H \mathbf{c}_2^{(i+1)}\right). \tag{58}$$

Here, $c_1^{(i+1)} = -\alpha^{(i+1)}(\chi_{ll}^{(i+1)} + |\mu_l^{(i+1)}|^2)$, $\mathbf{c}_2^{(i+1)} = \alpha^{(i+1)}((\mu_l^{(i+1)})^* \mathbf{y}_{-l}^{(i+1)} - \mathbf{X}\sum_{j \neq l} \chi_{jl}^{(i+1)} \mathbf{a}(\hat{\vartheta}_j + \beta_j))$, $\mathbf{y}_{-l}^{(i+1)} = \mathbf{y} - \mathbf{X} \cdot \sum_{j \neq l}(\mu_j^{(i+1)} \cdot \mathbf{a}(\hat{\vartheta}_j + \beta_j))$, $\mathbf{a}'(\hat{\vartheta}_j + \beta_l) = d\mathbf{a}(\hat{\vartheta}_j + \beta_l)/d\beta_l$, $\mu_l^{(i+1)}$ and $\chi_{jl}^{(i+1)}$ denote the l-th element and the (j, l)-th entry of $\boldsymbol{\mu}^{(i+1)}$ and $\boldsymbol{\Sigma}^{(i+1)}$, respectively. With (57), we update $\boldsymbol{\beta}$ in the derivative direction, i.e.,

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + \Delta \cdot \boldsymbol{\zeta}_{\boldsymbol{\beta}^{(i)}}^{(i+1)}, \tag{59}$$

where Δ is the stepsize that can be optimized by backtracking line search [49]. However, choosing the right stepsize can be time-consuming. Alternatively, we adopt a fixed stepsize, i.e.,

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + \frac{r_\theta}{100} \cdot \mathrm{sign}(\boldsymbol{\zeta}_{\boldsymbol{\beta}^{(i)}}^{(i+1)}), \tag{60}$$

where r_θ stands for the grid interval, and $\mathrm{sign}(\cdot)$ is the signum function. As mentioned in [20], the term $r_\theta/100$ guarantees that the final gap is smaller than 1% of r_θ, and the (approximate) true values may be attained within 100 iterations in the worst case. When the algorithm converges, the estimated downlink channels $\mathbf{h}^e$ is calculated as $\mathbf{h}^e = \mathbf{A}(\boldsymbol{\beta})\boldsymbol{\mu}$.

V. Simulation Results

In this section, we conduct numerical simulations to investigate the performance of our proposed method, which is compared with the following baseline schemes:

- *Baseline 1 (LASSO):* $\mathbf{h}$ is recovered using the l_1-norm minimization algorithm [50], [51].
- *Baseline 2 (SBL):* $\mathbf{h}$ is recovered using the standard SBL method [22].
- *Baseline 3 (Off-grid SBL):* $\mathbf{h}$ is recovered using the off-grid SBL method [20].
- *Baseline 4 (Burst LASSO):* $\mathbf{h}$ is recovered using the burst LASSO method [19].
- *Baseline 5 (PC-SBL):* $\mathbf{h}$ is recovered using the pattern-coupled SBL method [35].
- *Baseline 6 (GSVB):* $\mathbf{h}$ is recovered using the group sparse variational Bayes method [34].

For fairness, same size of the grid points is used for all the methods. The 3GPP spatial channel model (SCM) [39] is employed to generate the channel coefficients for an urban microcell, and we assume that the pilot matrix $\mathbf{X}$ has i.i.d. zero-mean circularly symmetric complex Gaussian entries with unit variance. The downlink frequency is set to 2170 MHz and the inter-antenna spacing is $d = c/(2f_0)$, with c being the light speed and $f_0 = 2000$ MHz. The normalized mean square error (NMSE) is defined as

$$\frac{1}{M_c} \sum_{m=1}^{M_c} \frac{\|\mathbf{h}_m^e - \mathbf{h}_m\|_2^2}{\|\mathbf{h}_m\|_2^2}, \tag{61}$$

where $\mathbf{h}_m^e$ is the estimate of $\mathbf{h}_m$ at the m-th Monte Carlo trial and M_c is the number of Monte Carlo trials. Unless stated otherwise, we assume that every channel realization consists of N_c random scattering clusters ranging from -90° to 90°, and each cluster contains N_s sub-paths concentrated in a $\mathcal{A}$ angular spread. In this case, the channels are generated with off-grid (continuous) path angles. Note that MATLAB codes have been made available online at https://sites.google.com/site/jsdaiustc/publication.

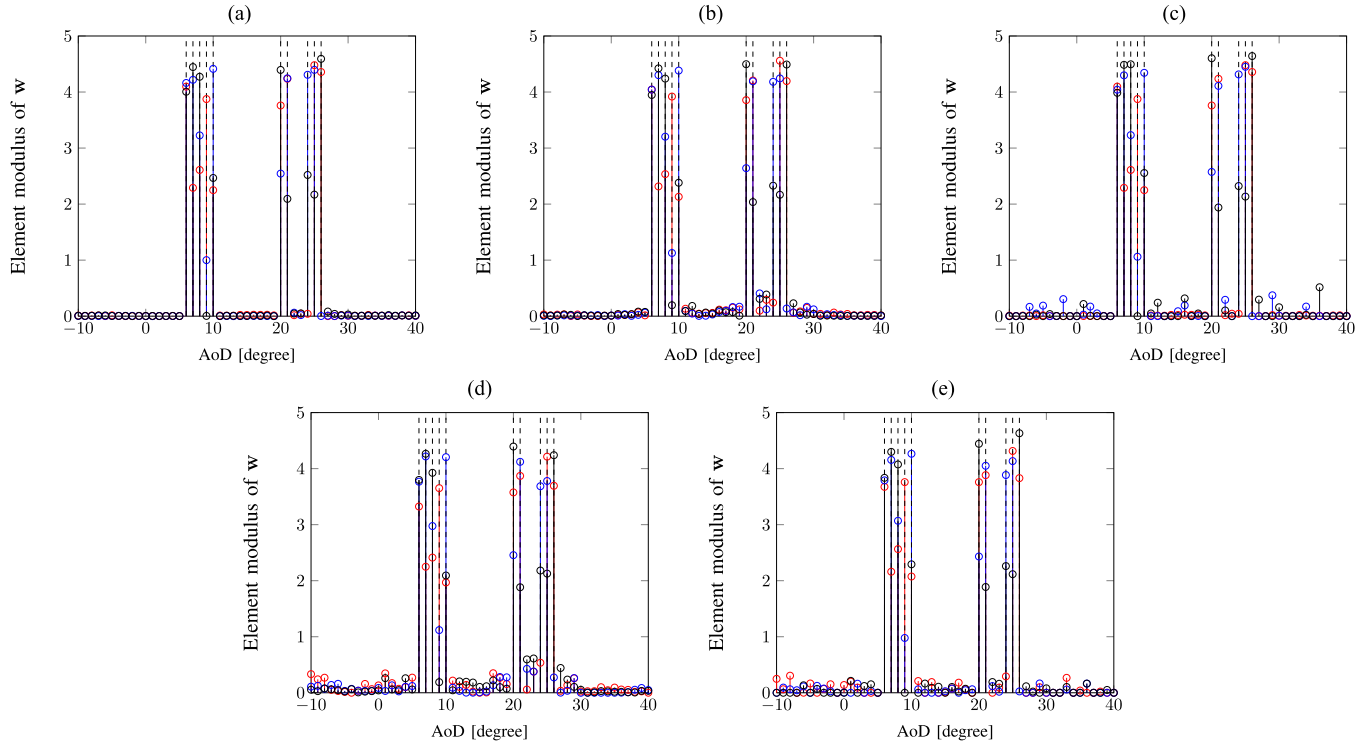


Fig. 3. Element modulus of $\mathbf{w}$ for three independent trials with $N = 100$, $\hat{L} = 180$, $T = 50$ and SNR $= 0$ dB. The true azimuth AoDs are denoted by dotted lines. (a) Proposed. (b) PC-SBL. (c) SBL. (d) Burst LASSO. (e) LASSO.

A. Recovered Channel Sparsity in Angular Domain

We study the effect of non-uniform burst-sparsity on the recovery performance for different channel estimation strategies. Consider a simple on-grid scenario where a ULA with 100 antennas at the BS is used to send the training pilot symbols with $N_c = 2$ scattering clusters and each cluster contains $N_s = 5$ sub-paths. The true AoDs are $\theta_{1,1} = 6°$, $\theta_{1,2} = 7°$, $\theta_{1,3} = 8°$, $\theta_{1,4} = 9°$, $\theta_{1,5} = 10°$, $\theta_{2,1} = 20°$, $\theta_{2,2} = 21°$, $\theta_{2,3} = 24°$, $\theta_{2,4} = 25°$, and $\theta_{2,5} = 26°$. The training pilots are randomly generated with $T = 50$ and the SNR is set to 0 dB. Fig. 3 shows the element modulus of the recovered channel sparse representation $\mathbf{w}$, where the grid is fixed to $[-90°, -89°, -88°, \ldots, 90°]$ for all the methods. It is observed that (i) the methods (PC-SBL and Burst LASSO) designed for burst-sparsity recovery have a significant performance loss due to the leakage of energy over outliers, e.g., at the grid points $22°$ and $23°$; (ii) the methods (SBL and LASSO) designed for individual sparsity recovery are not affected by outliers much, but they have a leakage of energy over some random positions, which will result in a more serious performance loss than the one caused by leakage of energy over outliers, because the random positions are usually far from the true positions; and (iii) our proposed method can greatly improve the sparsity and accuracy of the channel representation, and outliers can almost be eliminated.

B. Channel Estimation Performance Versus T

In Fig. 4, Monte Carlo trials are carried out to investigate the impact of the number of pilot symbols on the channel estimation performance. Assume that a ULA is equipped at the BS with $N = 128$ antennas, the training plots are randomly generated, the number of grid points is fixed at $\hat{L} = 200$, and SNR is chosen as 0 dB. All the results are obtained by averaging over 200 Monte Carlo channel realizations. Every independent run consists of $N_c = 2$ or 3 random scattering clusters ranging from $-90°$ to $90°$, and each cluster contains $N_s = 10$ sub-paths concentrated in a $\mathcal{A} = 10°$ or $30°$ angular spread.[5] Fig. 4 shows the NMSE performance of the downlink channel estimate achieved by the different channel estimation strategies versus the number of training pilot symbols T. It is seen that: (i) the NMSEs of all the methods decrease as the number of training pilot symbols increases, and the LASSO-based methods give the worst performance; (ii) Burst LASSO can get a performance gain by exploiting burst-sparsity with a small angular spread (Fig. 4(a)), because a small angular spread will bring a low possibility of occurring outliers in the angular domain; (iii) the SBL-based methods improve the NMSE performance, especially for off-grid SBL, PC-SBL and GSVB; and (iv) our solution always outperforms the state-of-the-art methods, which verifies that the former can jointly exploit burst-sparsity and exclude the harmful effect from outliers.

C. Channel Estimation Performance Versus SNR

In Fig. 5, Monte Carlo trials are carried out to study the impact of SNR on the channel estimation performance. We consider the same scenario as in Section V-B, except that the

[5]Since path angles are randomly generated, the on-grid assumption is not valid.

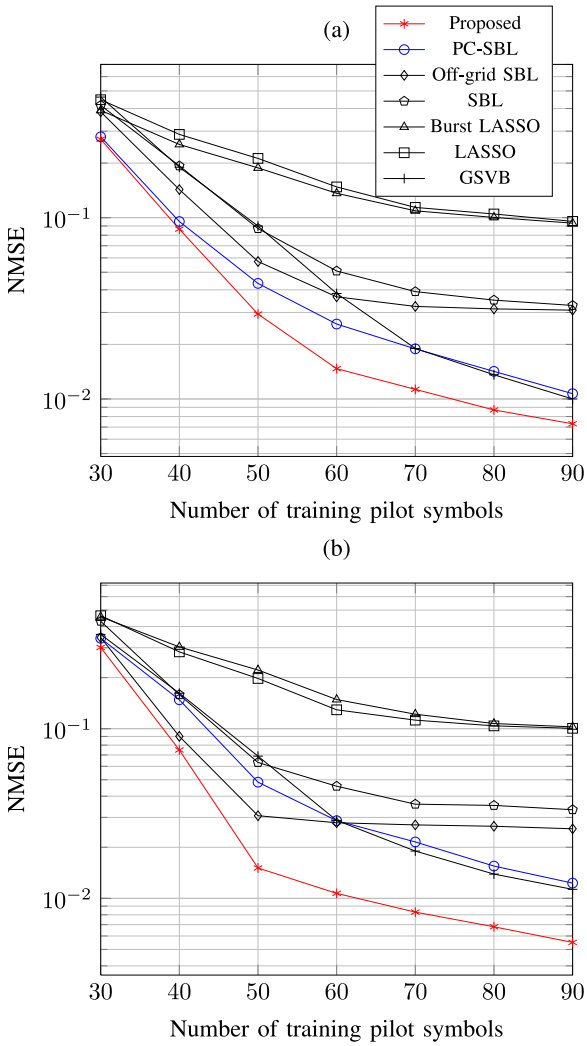


Fig. 4. NMSE of downlink channel estimate versus number of training pilot symbols with $N = 128$, $\hat{L} = 200$ and SNR $= 0$ dB. (a) $N_c = 3$ and $\mathcal{A} = 10^\circ$. (b) $N_c = 2$ and $\mathcal{A} = 30^\circ$.

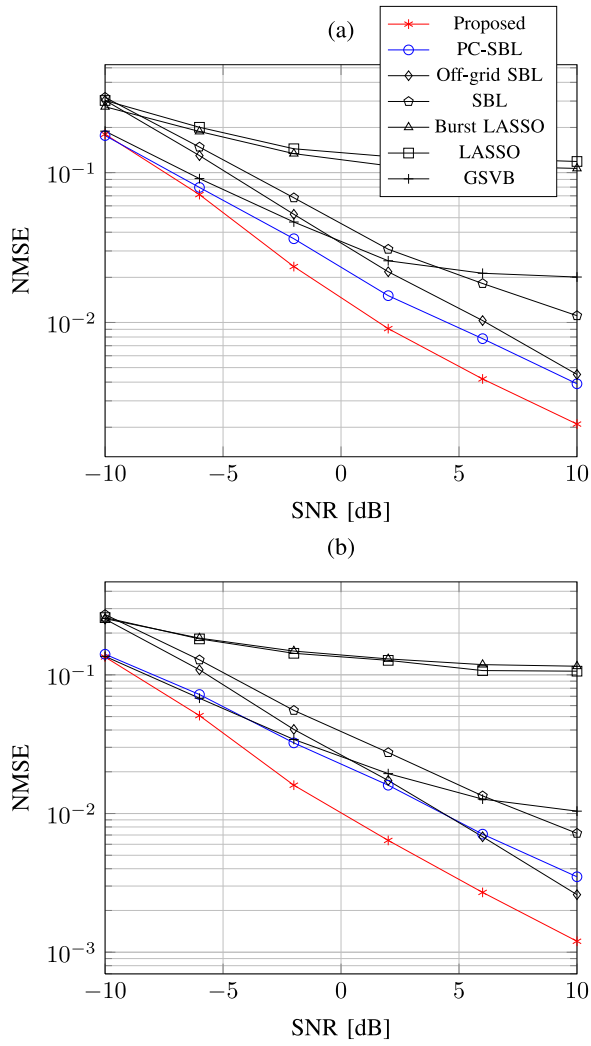


Fig. 5. NMSE of downlink channel estimate versus SNR with $N = 128$, $\hat{L} = 200$ and $T = 60$. (a) $N_c = 3$ and $\mathcal{A} = 10^\circ$. (b) $N_c = 2$ and $\mathcal{A} = 20^\circ$.

number of training pilot symbols is fixed at 60. Fig. 5 shows the NMSE performance of the downlink channel estimate achieved by the different channel estimation strategies versus SNR. All the results are obtained by averaging over 200 Monte Carlo channel realizations. It is observed that: (i) the NMSEs of all the methods decrease as SNR increases, and the LASSO- based methods still give the worst performance; (ii) compared with off-grid SBL, PC-SBL and GSVB obtain very limited gain from the burst-sparsity structure when SNR is sufficiently high, because the outliers deviated from block structures are significant in this case; (iii) our method is always superior to the others, and the performance gap between our method and PC-SBL increases when SNR increases.

D. Channel Estimation Performance Versus $\mathcal{A}$

We now examine the impact of the angular spread on the channel estimation performance, where two scenarios are considered: (i) a ULA is equipped at the BS with $N = 150$ antennas and $N_c = 3$; and (ii) $N = 128$ and $N_c = 2$. Other parameters are set as follows: $T = 60$, $N_s = 10$, $\hat{L} = 200$ and SNR$= 0$ dB. Fig. 6 shows the NMSE performance of the downlink channel estimate achieved by the different channel estimation strategies versus the angular spread. All the results are obtained by averaging over 200 Monte Carlo channel realizations. It is seen that the performance of PC-SBL degrades significantly with the increase of the angular spread; while our method keeps a reasonable performance gain, especially for the second scenario (Fig. 6(b)). Actually, the same observations are found in Figs. 4 and 5. For example, the performance gap between our method and PC-SBL in Fig. 4(b) is larger than the one in Fig. 4(a), where Fig. 4(b) has a larger angular spread. The main reason is that the larger the angular spread is, the higher possibility of occurring outliers is.

E. Channel Estimation Performance Versus $\hat{L}$

Finally, the impact of the number of grid points on the channel estimation performance is examined. Consider a ULA is

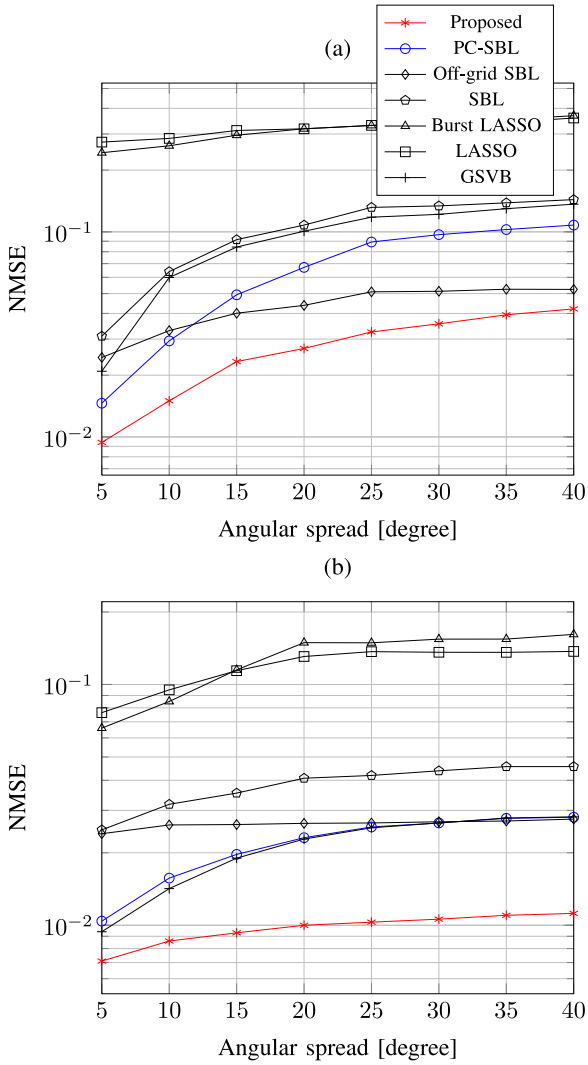


Fig. 6. NMSE of downlink channel estimate versus angular spread with $\hat{L} = 200$ and $T = 60$. (a) $N = 150$, $N_c = 3$ and SNR $= 0$ dB. (b) $N = 128$, $N_c = 2$ and SNR $= 0$ dB.

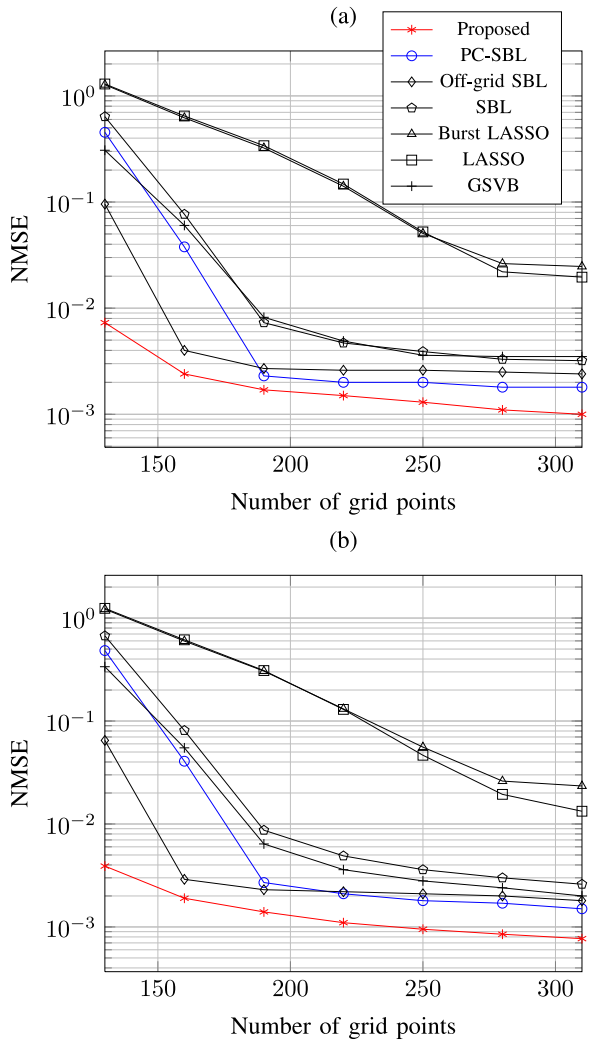


Fig. 7. NMSE of downlink channel estimate versus number of grid points with $N = 150$, $T = 80$ and SNR $= 10$ dB. (a) $N_c = 3$ and $\mathcal{A} = 10°$. (b) $N_c = 2$ and $\mathcal{A} = 20°$.

equipped at the BS with $N = 150$ antennas, the number of training pilot symbols is fixed at $T = 80$, and SNR is chosen as 10 dB. Fig. 7 shows the NMSE performance of the downlink channel estimate achieved by the different channel estimation strategies versus the number of grid points. All the results are obtained by averaging over 200 Monte Carlo channel realizations. We see that the NMSEs of all the methods decrease as the number of grid points increases, and our method always outperforms the state-of-the-art schemes, especially for a small $\hat{L}$. Moreover, it reaffirms that the performance gap between the proposed algorithm and PC-SBL increases when the angular spread increases.

VI. Conclusion

The problem of joint downlink channel estimation and non-uniform burst-sparsity exploiting for massive MIMO systems is tackled in this paper. Firstly, we devise a novel non-uniform burst-sparsity model to capture a more general burst-sparsity structure in practice, and an improved pattern-coupled prior to account for bursts and outliers simultaneously is then introduced. Finally, we propose a generic SBL-based framework to automatically detect unknown outliers and bursts, and simultaneously achieve massive MIMO channel estimation. Simulation results illustrate that our method indeed works for the non-uniform burst-sparsity model, and it can significantly improve the channel estimation performance when the strict block-sparsity assumption is invalid. Compared with the pattern-coupled SBL-based method in [35] that only gives a sub-optimal solution without convergence guarantee, our method provides a stationary solution and can include [35] as a special case.

References

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[3] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

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

[5] J.-C. Shen, J. Zhang, K.-C. Chen, and K. B. Letaief, "High-dimensional CSI acquisition in massive MIMO: Sparsity-inspired approaches," *IEEE Syst. J.*, vol. 11, no. 1, pp. 32–40, Mar. 2017.

[6] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, "Massive MIMO performance with imperfect channel reciprocity and channel estimation error," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3734–3749, Sep. 2017.

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

[8] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[9] J. Hoydis, C. Hoek, T. Wild, and S. Ten Brink, "Channel measurements for large antenna arrays," in *Proc. Int. Symp. Wireless Commun. Syst.*, Paris, France, Aug. 2012, pp. 811–815.

[10] X. Rao and V. K. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Jun. 2014.

[11] A. Liu, F. Zhu, and V. K. Lau, "Closed-loop autonomous pilot and compressive CSIT feedback resource adaptation in multi-user FDD massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 173–183, Jan. 2017.

[12] C.-K. Wen, S. Jin, K.-K. Wong, J.-C. Chen, and P. Ting, "Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1356–1368, Mar. 2015.

[13] Z. Chen and C. Yang, "Pilot decontamination in wideband massive MIMO systems by exploiting channel sparsity," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5087–5100, Jul. 2016.

[14] J.-C. Shen, J. Zhang, E. Alsusa, and K. B. Letaief, "Compressed CSI acquisition in FDD massive MIMO: How much training is needed?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4145–4156, Jun. 2016.

[15] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 802–814, Oct. 2014.

[16] L. You, X. Gao, A. L. Swindlehurst, and W. Zhong, "Channel acquisition for massive MIMO-OFDM with adjustable phase shift pilots," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1461–1476, Mar. 2016.

[17] Z. Gao, L. Dai, W. Dai, B. Shim, and Z. Wang, "Structured compressive sensing-based spatio-temporal joint channel estimation for FDD massive MIMO," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 601–617, Feb. 2016.

[18] Z. Gao, C. Zhang, Z. Wang, and S. Chen, "Priori-Information aided iterative hard threshold: A low-complexity high-accuracy compressive sensing based channel estimation for TDS-OFDM," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 242–251, Jan. 2015.

[19] A. Liu, V. K. Lau, and W. Dai, "Exploiting burst-sparsity in massive MIMO with partial channel support information," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7820–7830, Nov. 2016.

[20] J. Dai, A. Liu, and V. K. Lau, "FDD massive MIMO channel estimation with arbitrary 2D-array geometry," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2584–2599, May 2018.

[21] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.

[22] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. Jun., pp. 211–244, 2001.

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

[24] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse Bayesian inference," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 38–43, Jan. 2013.

[25] J. Dai and H. C. So, "Sparse Bayesian learning approach for outlier-resistant direction-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 744–756, Feb. 2018.

[26] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[27] L. Chen, A. Liu, and X. Yuan, "Structured turbo compressed sensing for massive MIMO channel estimation using a Markov prior," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4635–4639, May 2018.

[28] J. Dai, A. Liu, and V. K. Lau, "Joint channel estimation and user grouping for massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 622–637, Feb. 2019.

[29] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[30] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint-Petersburg, Russia, Jul. 2011, pp. 2168–2172.

[31] X. Li, J. Fang, H. Li, and P. Wang, "Millimeter wave channel estimation via exploiting joint sparse and low-rank structures," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1123–1133, Feb. 2018.

[32] Z. Zhang and B. D. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2009–2015, Apr. 2013.

[33] Q. Wu, Y. D. Zhang, M. G. Amin, and B. Himed, "Multi-task Bayesian compressive sensing exploiting intra-task dependency," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 430–434, Apr. 2015.

[34] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Variational Bayes group sparse time-adaptive parameter estimation with either known or unknown sparsity pattern," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3194–3206, Jun. 2016.

[35] J. Fang, Y. Shen, H. Li, and P. Wang, "Pattern-coupled sparse Bayesian learning for recovery of block-sparse signals," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 360–372, Jan. 2015.

[36] P. Wang, M. Pajovic, P. V. Orlik, T. Koike-Akino, K. J. Kim, and J. Fang, "Sparse channel estimation in millimeter wave communications: Exploiting joint AoD-AoA angular spread," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017, pp. 1–6.

[37] J. Fang, L. Zhang, and H. Li, "Two-dimensional pattern-coupled sparse Bayesian learning via generalized approximate message passing," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2920–2930, Jun. 2016.

[38] 3GPP, "Universal mobile telecommunications system (UMTS); Spatial channel model for multiple input multiple output (MIMO) simulations," 3GPP TR 25.996 version 11.0.0 Release 11, 2012.

[39] A. F. Molisch, A. Kuchar, J. Laurila, K. Hugl, and R. Schmalenberger, "Geometry-based directional model for mobile radio channels-principles and implementation," *Trans. Emerg. Telecommun. Technol.*, vol. 14, no. 4, pp. 351–359, 2003.

[40] J. Dai, X. Bao, W. Xu, and C. Chang, "Root sparse Bayesian learning for off-grid DOA estimation," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 46–50, Jan. 2017.

[41] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[42] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Fac. Grad. School, Ph.D. dissertation, University of Minnesota, Minneapolis, MN, USA, 2014.

[43] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[44] Z. Yang, J. Li, P. Stoica, and L. Xie, "Sparse methods for direction-of-arrival estimation," 2016, arXiv:1609.09596.

[45] D. Shutin and B. H. Fleury, "Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3609–3623, Aug. 2011.

[46] D. Shutin, W. Wang, and T. Jost, "Incremental sparse Bayesian learning for parameter estimation of superimposed signals," in *Proc. 10th Int. Conf. Sampling Theory Appl.*, Bremen, Germany, Jul. 2013, pp. 6–9.

[47] L. Hu, Z. Shi, J. Zhou, and Q. Fu, "Compressed sensing of complex sinusoids: An approach based on dictionary refinement," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3809–3822, Jul. 2012.

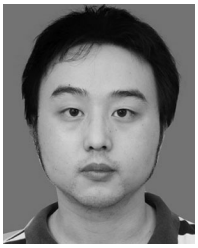
[48] L. Hu, J. Zhou, Z. Shi, and Q. Fu, "A fast and accurate reconstruction algorithm for compressed sensing of complex sinusoids," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5744–5754, Nov. 2013.

[49] S. J. Wright and J. Nocedal, *Numer. Optim.,* Berlin, Germany: Springer Science, 2006.

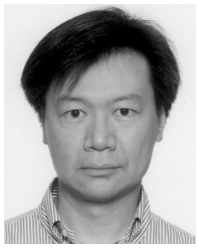
[50] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[51] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

Jisheng Dai (S'08–M'11) received the B.Eng. degree in electronic engineering from Nanjing University of Technology, Nanjing, China, in 2005, and the Ph.D. degree in information and communication engineering from the University of Science and Technology of China, Hefei, China, in 2010. He was a Research Assistant with the Department of Electrical and Electronic Engineering, The University of Hong Kong, in 2009, and a Visiting Scholar with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, during 2017–2018. He is currently a Research Fellow with the School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, China. His research interests include convex optimization theory, wireless communications, machine learning, and bioinformatics.



An Liu (S'07–M'09–SM'17) received the B.S. and Ph.D. degrees in electrical engineering from Peking University, Beijing, China, in 2011 and 2004, respectively. From 2008 to 2010, he was a Visiting Scholar with the Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, CO, USA. He was a Postdoctoral Research Fellow during 2011–2013, a Visiting Assistant Professor in 2014, and a Research Assistant Professor during 2015–2017 with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology. He is currently a Distinguished Research Fellow with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His research interests include wireless communications, stochastic optimization, and compressive sensing.



Hing Cheung So (S'90–M'95–SM'07–F'15) was born in Hong Kong. He received the B.Eng. degree from the City University of Hong Kong, Hong Kong, and the Ph.D. degree from The Chinese University of Hong Kong, both in electronic engineering, in 1990 and 1995, respectively. From 1990 to 1991, he was an Electronic Engineer with the Research and Development Division, Everex Systems Engineering Ltd., Hong Kong. During 1995–1996, he was a Postdoctoral Fellow with The Chinese University of Hong Kong. From 1996 to 1999, he was a Research Assistant Professor with the Department of Electronic Engineering, City University of Hong Kong, where he is currently a Professor. His research interests include detection and estimation, fast and adaptive algorithms, multidimensional harmonic retrieval, robust signal processing, source localization, and sparse approximation.

Dr. So has been on the editorial boards of the IEEE SIGNAL PROCESSING MAGAZINE (2014–2017), IEEE TRANSACTIONS ON SIGNAL PROCESSING (2010–2014), *Signal Processing* (2010–), and *Digital Signal Processing* (2011–). He was also a Lead Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, special issue on "Advances in Time/Frequency Modulated Array Signal Processing" in 2017. In addition, he was an elected member in Signal Processing Theory and Methods Technical Committee (2011–2016) of the IEEE Signal Processing Society where he was the Chair in the awards subcommittee (2015–2016).