A Unified Activity Detection Framework for Massive Access: Beyond the Block-Fading Paradigm

Jianan Bai and Erik G. Larsson

***Abstract*—The wireless channel changes continuously with time and frequency and the block-fading assumption, which is popular in many theoretical analyses, never holds true in practical scenarios. This discrepancy is critical for user activity detection in grant-free random access, where joint processing across multiple coherence blocks is undesirable, especially when the environment becomes more dynamic. In this paper, we develop a framework for low-dimensional approximation of the channel to capture its variations over time and frequency, and use this framework to implement robust activity detection algorithms. Furthermore, we investigate how to efficiently estimate the principal subspace that defines the low-dimensional approximation. We also examine pilot hopping as a way of exploiting time and frequency diversity in scenarios with limited channel coherence, and extend our algorithms to this case. Through numerical examples, we demonstrate a substantial performance improvement achieved by our proposed framework.**

***Index Terms*—Massive access, activity detection, continuously varying channel, low-dimensional approximation, pilot hopping**

I. Introduction

Massive access features a scenario where a wireless network serves a very large number of users simultaneously [2]. Due to the limited radio resources, a considerable number of users need to communicate over the same time-frequency block of a size constrained by the channel *coherence* length (i.e., the number of samples over which the channel remains relatively constant to allow for coherent signal processing). Many applications in massive access, e.g., intelligent transportation systems and tactile internet, also require low access latency. Balancing these intertwined needs – massive connectivity, efficient resource utilization, and minimal latency, is challenging.

In this paper, we focus on the initial stage in massive access, during which the active users inform the base station (BS) of their needs for communication by sending a pilot sequence. Due to the limited channel coherence, users either need to reuse a set of mutually orthogonal pilots with contention or use unique, yet non-orthogonal pilots. We restrict our attention to the use of non-orthogonal pilots since they have shown superior detection performance compared to orthogonal pilots and are particularly suitable for short-packet transmission, which is generally the case in grant-free random access (GFRA) [3]. This process, commonly known as user *activity detection*, has received much attention in recent years.

State-of-the-art activity detection schemes exploit the traffic sporadicity [4], [5]. That is, although the number of potential users can be exceedingly large, the number of simultaneously active users is comparable to the channel coherence length. The majority of activity detection methods fall into two primary categories: compressed sensing (CS)-based and covariance-based approaches. Consider a network with a multi-antenna BS and single-antenna users, the CS-based approaches are motivated by the linear measurement model $\mathbf{Y} = \mathbf{\Phi}\mathbf{X} + \mathbf{W}$ that appears in pilot transmission, where each column of $\mathbf{Y}$ is the received signals at one BS antenna, each column of $\mathbf{\Phi}$ represents the pilot sequence of a user, each row of $\mathbf{X}$ represents the effective channel gains (assumed to be *constant*) from a user to all BS antennas, and $\mathbf{W}$ represents additive noise. Due to the sporadic traffic, $\mathbf{X}$ is row-sparse and can be accurately recovered under certain conditions, and activity detection can be performed by thresholding the norm of each row in the recovered $\mathbf{X}$. The CS-based algorithms range from conventional optimization-based methods [6], [7] to statistical methods using, for example, approximate message passing (AMP) [4]. There are several advantages of the CS-based approaches. First, they directly provide the channel estimates as a byproduct. Second, additional side information (e.g., temporal correlation [8]) can be incorporated, especially for the statistical methods. Third, although the increase in the number of BS antennas can improve detection performance in multiple-input multiple-output (MIMO) systems, the CS-based approaches can work for a moderately small number of antennas, enabling flexible deployments. However, the detection capabilities of CS-based approaches are fundamentally limited by the need to recover all non-zero elements in $\mathbf{X}$ – we usually cannot accurately detect more active users than the pilot length without additional side information. Furthermore, as observed in [5], the AMP algorithm may exhibit random, non-convergent behaviors in certain application regimes.

The covariance-based approach [5], which originates from the support recovery problem in sparse Bayesian learning [9], takes another perspective when solving the activity detection problem. Specifically, when the non-zero elements in $\mathbf{X}$ are statistically independent (let them also be zero-mean and unit-variance for simplicity), as the number of BS antennas M increases, the sample covariance matrix $\mathbf{Y}\mathbf{Y}^{\mathsf{H}}/M$ approaches the true covariance matrix $\mathbf{\Phi}\mathbf{D}_{\mathbf{a}}\mathbf{\Phi}^{\mathsf{H}} + \sigma^2\mathbf{I}$, which is parameterized by the binary vector of user activities $\mathbf{a}$, with noise variance σ^2. An efficient coordinate descent algorithm is developed in [5] to recover $\mathbf{a}$ by matching these two covariance matrices with respect to (w.r.t.) the log-determinant divergence

This paper was presented in part at the Asilomar SSC 2023 conference [1].

The authors are with the Department of Electrical Engineering (ISY), Linköping University, 58183 Linköping, Sweden (email: jianan.bai@liu.se, erik.g.larsson@liu.se). This work was supported in part by Excellence Center at Linköping-Lund in Information Technology (ELLIIT), and by the Knut and Alice Wallenberg (KAW) foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

– a direct consequence of a maximum-likelihood (ML) estimation of $\mathbf{a}$. Different from the CS approaches, the covariance-based one does not directly provide channel estimates (which may be acquired after activity detection, using, for example, linear minimum mean-square error (MMSE) estimation), and it becomes less straightforward to incorporate statistical side information (although not impossible, as showcased in [10]). However, the covariance-based approach can identify a considerably larger number of active users than the CS-based ones in large, albeit finite, antenna regimes; the fundamental scaling law is formally substantiated in [5].

Despite the promising detection performance showcased in various scenarios, current activity detection schemes (e.g., [4], [5], [6], [8], [10]) are predominantly restricted to an ideal block-fading model, where the channel is assumed to be constant during pilot transmission to make the linear measurement model $\mathbf{Y} = \mathbf{\Phi X} + \mathbf{W}$ valid. This block-fading assumption can be formally justified by the concept of coherence block (a typical definition is a time-frequency block with up to π phase shifts [11, Ch. 2.1]) and the sampling theorem, only when joint processing or coding across a sufficiently large number of blocks is feasible. Nevertheless, interpolation may not be possible in the context of activity detection, wherein instantaneous processing is anticipated within a block or a few blocks. The situation is most challenging in highly dynamic IoT environments associated with emerging 5G scenarios, where higher operating frequencies, larger sub-carrier spacings, and potentially high mobility need to be considered.

Some efforts have been made to address the channel variations in activity detection in orthogonal frequency division multiplexing (OFDM) systems. In [12], sub-carriers are partitioned into sub-blocks, and the channel gain is assumed to change linearly within each sub-block; consequently, an AMP-based detection algorithm was developed. The method in [12] was further refined in [13] by approximating the channel variations in each sub-block as a low-order polynomial. The covariance-based approach was extended for OFDM systems in [14] by assuming independent and identically distributed (i.i.d.) channel taps in the delay domain and capitalizing on the discrete Fourier transform (DFT) structure present in the OFDM symbols. A similar idea was explored in [15] by using the discrete cosine transform (DCT) representations. All these approximation models, including the simplest block-fading model, as we will show later, can be unified from a dimensionality reduction perspective. Another direction to combat the channel variations in activity detection is to use pilot hopping to explore extra time and/or frequency diversities, as demonstrated in [16], [17].

A. Contributions and Organization of the Paper

1) Robust Activity Detection:

We consider a general channel model for activity detection in Section II, which allows the channel coefficients to change symbol-by-symbol, departing from the commonly assumed block-fading model in the literature. We present a dimensionality-reduction framework for channel variations that generalizes various approaches in, for example, [12], [13], [14], [15]. By leveraging the low-dimensional structure, we implement a modified covariance-based activity detection algorithm that is robust under highly varying channels.

2) Low-Dimensional Structure Learning:

The proposed framework relies on the knowledge of a low-dimensional structure of the channel, which is generally not known a priori. We investigate the learning of the low-dimensional structure in Section IV, by allowing the users to transmit additional, *dedicated* all-one pilots. In this case, the low-dimensional structure can be accurately learned by solving a standard low-rank matrix approximation problem, without the need for knowing the user activities.

We also explore *joint* estimation of the low-dimensional structure and detection of the user activities, by reusing the same received signals. For this joint problem, we consider an alternating procedure, where we iteratively update the estimates of user activities and the low-dimensional structure. Particularly, the structure learning is formulated as a weighted low-rank matrix approximation (WLRMA) problem, where the weight matrix is determined by the (estimated) user activities. However, we observe that this approach performs poorly due to the inherent ill-conditioning of the joint problem. Details on this approach are provided in the Appendix.

3) Pilot Hopping:

We examine the pilot hopping scheme and combine it with the activity detection algorithm in Section V. We develop a new algorithm for hopping pattern generation, inspired by the configuration model for random graphs [18, Ch. 5.3]. The proposed hopping pattern generation algorithm significantly outperforms the random hopping pattern generation methods in [16], [17]. Furthermore, we use pilot hopping to reduce the overhead associated with learning the low-dimensional structure of the channels using dedicated pilots.

Remark: We presented the dimensionality-reduction framework for activity detection in the conference paper [1]. The learning of the low-dimensional structure and the pilot hopping scheme are new contributions introduced in this paper.

B. Notation

Vectors are denoted by boldface lowercase letters, $\mathbf{x}$, matrices by boldface uppercase letters, $\mathbf{X}$, with determinant $|\mathbf{X}|$, and sets by calligraphic letters, $\mathcal{X}$, with cardinality $|\mathcal{X}|$. $[N]$ denotes the set $\{1, \cdots, N\}$. $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\mathsf{H}}$, $(\cdot)^{*}$, and $(\cdot)^{-1}$ denote transpose, conjugate transpose, complex conjugate, and inverse, respectively. $\mathbf{I}_N$, $\mathbf{1}_N$, and $\mathbf{0}_N$ denote respectively the identity matrix, the all-one vector, and the all-zero vector of size N (omitted when the size is obvious). $\mathbb{E}[\cdot]$ denotes the statistical expectation. $\mathbb{1}\{\cdot\}$ is the indicator function, which equals to 1 for true propositions and 0 otherwise. The multivariate circularly symmetric complex Gaussian distribution with covariance $\mathbf{R}$ is denoted by $\mathcal{CN}(\mathbf{0}, \mathbf{R})$. $\mathbf{D}_{\mathbf{x}}$ denotes a diagonal matrix with $\mathbf{x}$ on its diagonal. $\mathbb{R}_{+}$, $\mathbb{C}$, and $\mathbb{S}_{+}$ denote the spaces of non-negative numbers, complex numbers, and positive semidefinite (p.s.d.) matrices. $\|\cdot\|$ denotes the norms.

II. System Model

We consider a single-cell system with an M-antenna BS and K single-antenna users. Each user, $k \in [K]$, is pre-assigned a unique pilot sequence of length L, denoted by $\boldsymbol{\phi}_k \triangleq [\phi_{k1}, \cdots, \phi_{kL}]^\mathsf{T}$, which is normalized to have unit average energy per symbol, i.e., $\|\boldsymbol{\phi}_k\|_2^2 = L$. Our focus lies primarily in the regime where $L \ll K$, and therefore, pilot sequences allocated to distinct users are mutually non-orthogonal in general, i.e., $\boldsymbol{\phi}_i^\mathsf{H}\boldsymbol{\phi}_j \neq 0$ when $i \neq j$.

In a given communication round, the active users, denoted by $\mathcal{K}_{\text{act}}$, with $|\mathcal{K}_{\text{act}}| \ll K$ due to the sporadic traffic, transmit their pre-assigned pilots synchronously on a time-frequency block of size $T \times F$ (i.e., a resource block with T OFDM symbols and F sub-carriers). We assume $L = TF$ for simplicity. At the mth antenna, denoting by h_{lkm} the small-scale fading coefficient experienced by the lth pilot symbol sent from user k, the received pilot signal is given by

$$\mathbf{y}_m = \sum_{k \in [K]} a_k \sqrt{\beta_k} \mathbf{D}_{\mathbf{h}_{km}} \boldsymbol{\phi}_k + \mathbf{w}_m, \tag{1}$$

where $a_k \in \{0, 1\}$ represents the activity of user k, β_k is the received signal strength (i.e., the product of the large-scale fading coefficient (LSFC) and the transmit power), $\mathbf{h}_{km} \triangleq [h_{1km}, \cdots, h_{Lkm}]^\mathsf{T}$ is the channel vector from user k to antenna m, and $\mathbf{w}_m \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ is additive noise that is independent across antennas.

Different from the block-fading model where the channel coefficients are assumed to be constant during pilot transmission, i.e., $\mathbf{h}_{km} = \overline{h}_{km}\mathbf{1}_L$ for some scalar-valued $\overline{h}_{km}$, the received signal model (1) allows the transmitted symbols to experience different, static channels. Such a model can be obtained by prepending sufficiently long cyclic prefixes to the OFDM symbols, because the cyclic convolution operation is diagonalized in the coordinate system defined by columns of the DFT matrix, as elucidated in, for example, [19, Ch. 7.1].

Remark 1: Our model assumes that the channel is static within a symbol, such that the effect of the channel for each symbol is to multiply the symbol with a complex number; in particular, there is no inter-symbol or inter-carrier interference. In practice, the channel is time-varying, and this model is, strictly speaking, an approximation. However, the channel generally varies slowly compared to the OFDM symbol duration, and this approximation is usually sufficiently good and used in real-world algorithm implementations of data transmission with OFDM.

III. Activity Detection with Dimensionality-Reduced Channel Variations

The underlying physical propagation model generally permits a low-dimensional structure in the channel vectors that we can exploit. Mathematically, a low-dimensional approximation of $\mathbf{h}_{km}$ can be expressed as

$$\mathbf{h}_{km} \approx \mathbf{G}\boldsymbol{\theta}_{km} = \sum_{n \in [N]} \theta_{nkm}\mathbf{g}_n, \tag{2}$$

where $\boldsymbol{\theta}_{km} = [\theta_{1km}, \cdots, \theta_{Nkm}]^\mathsf{T}$ is a random vector with i.i.d. entries, $\mathbf{G} \triangleq [\mathbf{g}_1, \cdots, \mathbf{g}_N] \in \mathbb{C}^{L \times N}$ is a deterministic basis matrix (the basis vectors $\{\mathbf{g}_n\}$ may have different lengths), and we refer to N as the approximation *order*. The intuition behind (2) is that all the channel vectors $\{\mathbf{h}_{km}\}$ approximately lie in an N-dimensional subspace, with $N \ll L$, spanned by $\{\mathbf{g}_n\}$, and the stochastic nature of $\mathbf{h}_{km}$ is encapsulated within a significantly reduced representation $\boldsymbol{\theta}_{km}$.

The assumed low-dimensional structure is motivated by the fact that even if the channel varies substantially within a block, the fading experienced by adjacent pilot symbols can still exhibit a strong statistical correlation. This can be seen as a generalization of the coherence block, nominally defined in a deterministic sense (one typical definition is a time-frequency block with up to π phase shifts), to a statistical viewpoint.

Remark 2: The concept of low-dimensional approximation by exploring statistical correlation as such is not new; it originates from the idea of principal component analysis (PCA) [20]. The existing works [12], [13], [14], [15] can be seen as special cases of the model in (2). Previously, the channel subspace was mainly exploited in the spatial domain (e.g., [21]), instead of *over time and frequency* as in our work.

A. Existing Low-Dimensional Models

The block-fading model, for which $\mathbf{G} = \mathbf{1}_L$, is the simplest example of (2). The block-fading assumption can be justified by the sampling theorem, which allows us to interpolate back to the continuous channel variations by taking one sample from each coherence block when the number of samples is sufficiently large. However, typical applications of GFRA require low latency, which limits the number of blocks that can be jointly processed. The problem becomes more challenging when substantial channel variations are present due to a large excess delay and high mobility in dynamic IoT environments.

Recently, research efforts have been made for activity detection in wideband systems, especially in OFDM systems. We can classify these approaches into two primary categories, and both can be seen as special cases of (2):

1) The first category of approaches approximates the channel variations in a block-wise manner. Specifically, the OFDM sub-carriers are divided into multiple sub-blocks. The variations within each sub-block are approximated as a linear function in [12], or as a low-degree polynomial in [13]. These models correspond to a block-diagonal basis matrix $\mathbf{G}$. As an example, the block-wise linear (BWL) model has $\mathbf{G} = \text{bdiag}(\mathbf{G}_0, \cdots, \mathbf{G}_0)$ with $N/2$ diagonal blocks (assume $N/2$ is an integer). Each diagonal block $\mathbf{G}_0$ is a $(L/N) \times 2$ matrix, with the first column being an all-one vector accounting for the mean value, and the entries in the second column are equally spaced and centered at zero, representing the linear variations. The block-wise polynomial (BWP) model can be constructed similarly by incorporating additional columns in the diagonal blocks accounting for the higher degrees in the polynomial.

2) Another direction of work utilizes the transform-domain representations. Particularly, in OFDM systems, the channel frequency responses are Fourier transforms of channel impulse responses (CIRs). Since the CIRs are generally dominated by a few strong channel taps, the basis matrix can be selected as the corresponding columns in the DFT matrix [14] when assuming that these dominant channel taps are independently distributed. A similar method in [15] is to use the columns in the DCT matrix instead.

B. A Viewpoint from Low-Rank Covariance Approximation

To further motivate the dimensionality reduction techniques and to unify those low-dimensional channel models, we provide another viewpoint here. We assume that the channel between each user-antenna pair is modeled by the correlated Rayleigh fading that is stationary across different user-antenna pairs. In this case, $\mathbf{h}_{km} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R})$ with the same $L \times L$ covariance matrix[1] $\mathbf{R}$ for all $k \in [K]$ and $m \in [M]$. The channel covariance $\mathbf{R}$ is assumed to have rank N, or to have a good rank-N approximation. Then, we can obtain a Gramian factorization $\mathbf{R} = \mathbf{G}\mathbf{G}^{\mathsf{H}}$, where the $L \times N$ matrix $\mathbf{G}$ serves as the basis matrix in (2). Additionally, under the Rayleigh fading assumption, the random vector $\boldsymbol{\theta}_{km}$ has the distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$. The aforementioned low-dimensional channel approximation approaches, by imposing the particular structures in $\mathbf{G}$, equivalently assume corresponding special structures in the channel covariance matrix $\mathbf{R}$.

Remark 3: For simplicity, we assume that the small-scale fading is spatially stationary to reduce the number of unknowns in the channel covariance estimation. This assumption can be justified for example if the users are in the far-field of the array, and move around the area over a sufficiently long time window. However, in practice, users located in different areas may certainly experience non-stationarities. To model this, one could for example group the users based on their locations, with location-dependent channel covariances. We have to leave this as a topic for possible future work.

C. Covariance-Based Activity Detection

We assume that the channel covariance $\mathbf{R}$ is known, has rank N, and can be factorized into the Gramian form $\mathbf{R} = \mathbf{G}\mathbf{G}^{\mathsf{H}}$, as discussed in Section III-B. (If $\mathbf{R}$ is not perfectly known, we replace it with an estimate. A discussion on estimating a low-rank $\mathbf{R}$ will be provided in Section IV.) We aim to detect the user activities $\{a_k\}$ by exploiting this low-rank structure. Although there are various algorithms for activity detection based on CS, we will focus on the covariance approach proposed in [5]. This approach does not directly produce channel estimates but can potentially detect many more active users than the CS-based approaches when using a large antenna array in massive MIMO systems.

The covariance approach for activity detection is developed from a (relaxed) ML estimation of $\boldsymbol{\gamma} \triangleq [\gamma_1, \cdots, \gamma_K]^{\mathsf{T}}$, with $\gamma_k \triangleq a_k \beta_k$. By assuming i.i.d. $\{\mathbf{h}_{km}\}$ with distribution $\mathcal{CN}(\mathbf{0}, \mathbf{R})$, the received signals $\{\mathbf{y}_m\}$ become i.i.d. with distribution $\mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ and the covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \triangleq \mathbb{E}[\mathbf{y}_m \mathbf{y}_m^{\mathsf{H}}] = \sum_{k\in[K]} \gamma_k \mathbf{S}_k \mathbf{S}_k^{\mathsf{H}} + \sigma^2 \mathbf{I}, \tag{3}$$

where $\mathbf{S}_k \triangleq \mathbf{D}_{\boldsymbol{\phi}_k} \mathbf{G}$ is the effective pilot (matrix) of user k. In the block-fading case where $\mathbf{G} = \mathbf{1}$, we have $\mathbf{S}_k = \boldsymbol{\phi}_k$, and each active user contributes a rank-one component in the signal covariance. However, in general cases with varying channels, the contribution from each active user becomes a rank-N component that complicates the problem.

After some rescaling and removal of constant terms, the negative log-likelihood function of $\boldsymbol{\gamma}$ given $\{\mathbf{y}_m\}$ is

$$f(\boldsymbol{\gamma}) \triangleq \log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}| + \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \widehat{\boldsymbol{\Sigma}}), \tag{4}$$

where

$$\widehat{\boldsymbol{\Sigma}} \triangleq \frac{1}{M} \sum_{m\in[M]} \mathbf{y}_m \mathbf{y}_m^{\mathsf{H}} \tag{5}$$

is the sample covariance matrix of the received signals. Notice that $\widehat{\boldsymbol{\Sigma}}$ is a sufficient statistic for the estimation of $\boldsymbol{\gamma}$, and the ML formulation can also be interpreted as a covariance matching problem by minimizing the log-determinant divergence between $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$. The binary constraint $a_k \in \{0,1\}, \forall k$, renders the original ML problem combinatorial, causing the complexity to grow exponentially with K. The search space can be relaxed to $\boldsymbol{\gamma} \in \Gamma$, where Γ represents the box constraint $\{\boldsymbol{\gamma} \in \mathbb{R}^K : 0 \le \gamma_k \le \beta_k, \forall k\}$ when the LSFCs $\{\beta_k\}$ are known, and represents the non-negative orthant $\mathbb{R}_+^K$ otherwise. The relaxed ML estimation of $\boldsymbol{\gamma}$ is then given by

$$\widehat{\boldsymbol{\gamma}}_{\mathrm{ML}} = \operatorname*{argmin}_{\boldsymbol{\gamma}\in\Gamma} f(\boldsymbol{\gamma}). \tag{P0}$$

An efficient coordinate descent algorithm is developed in [5] for block-fading channels. In each (inner) iteration of the coordinate descent algorithm, we pick a coordinate (user) k based on some pre-determined schedule and make the update

$$\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} + d^* \mathbf{e}_k, \tag{6}$$

where $\mathbf{e}_k$ is the kth column of $\mathbf{I}_K$, and

$$d^* = \operatorname*{argmin}_{d\in[-\gamma_k,\infty)} f(\boldsymbol{\gamma} + d\mathbf{e}_k). \tag{7}$$

When $N = 1$, changing d results in a rank-one update in the covariance matrix, and the optimal d^* can be obtained in closed form [5]. An extension of this approach was proposed in [14], and we can apply it to the general case when $N > 1$. For completeness, we briefly present the development of the approach in [14] in the following.

By applying Sylvester's determinant identity we obtain

$$\log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} + d\mathbf{S}_k \mathbf{S}_k^{\mathsf{H}}| = \log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}| + \log|\mathbf{I} + d\mathbf{A}_k|,$$

where $\mathbf{A}_k = \mathbf{S}_k^{\mathsf{H}} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{S}_k$, and Woodbury's matrix identity gives

$$(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} + d\mathbf{S}_k \mathbf{S}_k^{\mathsf{H}})^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} - d\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{S}_k (\mathbf{I} + d\mathbf{A}_k)^{-1} \mathbf{S}_k^{\mathsf{H}} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}.$$

After the eigenvalue decomposition (EVD) $\mathbf{A}_k = \mathbf{V}_k \mathbf{D}_{\boldsymbol{\lambda}_k} \mathbf{V}_k^{\mathsf{H}}$ with $\boldsymbol{\lambda}_k = [\lambda_{k1}, \cdots, \lambda_{kN}]^{\mathsf{T}}$, the cost function in (7) becomes

$$\begin{aligned} \mathcal{L}_k(d) &\triangleq f(\boldsymbol{\gamma} + d\mathbf{e}_k) - f(\boldsymbol{\gamma}) \\ &= \log|\mathbf{I} + d\mathbf{D}_{\boldsymbol{\lambda}_k}| - d\operatorname{tr}\left((\mathbf{I} + d\mathbf{D}_{\boldsymbol{\lambda}_k})^{-1} \boldsymbol{\Xi}_k\right) \\ &= \sum_{n\in[N]} \left(\log(1 + d\lambda_{kn}) - \frac{d\xi_{kn}}{1 + d\lambda_{kn}} \right) \end{aligned} \tag{8}$$

with $\boldsymbol{\Xi}_k = \mathbf{V}_k^{\mathsf{H}} \mathbf{S}_k^{\mathsf{H}} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{S}_k \mathbf{V}_k$ and $\xi_{kn} = [\boldsymbol{\Xi}_k]_{n,n}$.

The optimal d^* can be obtained by comparing the cost value for all feasible stationary points of (8) and boundary points.

[1] We have two different types of covariance matrix in this paper: the *channel* covariance matrix $\mathbf{R} \triangleq \mathbb{E}[\mathbf{h}\mathbf{h}^{\mathsf{H}}]$ and the *signal* covariance $\boldsymbol{\Sigma} \triangleq \mathbb{E}[\mathbf{y}\mathbf{y}^{\mathsf{H}}]$. They are not to be confused with each other or with the *spatial* covariance matrix.

Algorithm 1 Activity Detection

```
Input: sample covariance $\widehat{\mathbf{\Sigma}}$, and effective pilots $\{\mathbf{S}_k\}$
Initialize: $\gamma_k \leftarrow 0, \forall k$ and $\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{-1} \leftarrow \sigma^{-2}\mathbf{I}$
 1: for $i = 1, 2, \cdots, I$ do
 2:     Generate $\mathcal{K}$ by randomly permuting $[K]$
 3:     for $k$ taken from $\mathcal{K}$ by order do
 4:         $\mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{V}^{\mathsf{H}} \xleftarrow{\text{EVD}} \mathbf{S}_k^{\mathsf{H}}\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{-1}\mathbf{S}_k$
 5:         $\widetilde{\mathbf{V}} \leftarrow \mathbf{\Sigma}_{\boldsymbol{\gamma}}^{-1}\mathbf{S}_k\mathbf{V}$
 6:         $\forall n$: $\xi_n \leftarrow [\widetilde{\mathbf{V}}^{\mathsf{H}}\widehat{\mathbf{\Sigma}}\widetilde{\mathbf{V}}]_{n,n}$
 7:         $d^* = \operatorname{argmin} \mathcal{L}_k(d)$ for $d \in [-\gamma_k, \gamma_{\max} - \gamma_k]$
            ($\gamma_{\max}$ can be selected as an estimate of $\max\{\beta_k\}$)
 8:         $\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{-1} \leftarrow \mathbf{\Sigma}_{\boldsymbol{\gamma}}^{-1} - d^*\widetilde{\mathbf{V}}(\mathbf{I} + d^*\mathbf{D}_{\boldsymbol{\lambda}})^{-1}\widetilde{\mathbf{V}}^{\mathsf{H}}$
 9:         $\gamma_k \leftarrow \gamma_k + d^*$
10:     end for
11: end for
Output: activity estimate $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_K]^{\mathsf{T}}$
```

To find the stationary points of (8), where the derivative $\mathcal{L}_k'(d)$ equals zero, one needs to solve

$$\mathcal{L}_k'(d) = \sum_{n\in[N]} \left(\frac{\lambda_{kn}}{1 + d\lambda_{kn}} + \frac{\xi_{kn}}{(1 + d\lambda_{kn})^2} \right) = 0. \tag{9}$$

Multiplying both sides of the second equality in (9) by $\prod_{n\in[N]}(1 + d\lambda_{kn})^2$, the stationary points can also be represented as the solutions to

$$\sum_{n=1}^{N} (\lambda_{kn}(1 + d\lambda_{kn}) + \xi_{kn}) \prod_{j\neq n} (1 + d\lambda_{kj})^2 = 0, \tag{10}$$

which is a polynomial equation with order $2N - 1$. No explicit formula exists for $N \geq 3$, while efficient algorithms for the real-root isolation of high-order polynomials were developed [22], [23]. Alternatively, as suggested in [24], a one-dimensional search can be employed to minimize (8) directly.

The procedure above is repeated for I (outer) iterations. We summarize the overall procedure in Algorithm 1.

Complexity: The runtime complexity of Algorithm 1 is dominated by: 1) the matrix multiplications in steps 4, 5, 6, and 8, which require $\mathcal{O}(L^2N)$ arithmetic operations; 2) the EVD in step 4, which requires $\mathcal{O}(N^3)$ arithmetic operations; and 3) finding the minimizer of $\mathcal{L}_k(d)$ in step 7. Finding the roots in (9) has complexity $\mathcal{O}(N^3)$. Alternatively, we can use golden section search and parabolic interpolation to perform a one-dimensional search whose complexity does not scale with N [25]. Overall the algorithm has a computational complexity of order $\mathcal{O}(IKN(L^2 + N^2))$.

IV. Learning the Low-Dimensional Structure

In Section III, we assumed that the channel vectors lie in a low-dimensional linear subspace as the result of a low-rank channel covariance matrix. When the channel covariance matrix $\mathbf{R}$ is known, the optimal approximation that minimizes the approximation order can be obtained by projecting the channel vectors onto their principal subspace. However, in practical wireless environments, the channel covariance will not be perfectly known and may continuously change.

A naive approach is to first estimate the channels and use these channel estimates to form a channel covariance estimate. However, this approach fails when we allow symbol-to-symbol channel variations, as we cannot estimate LKM unknowns in $\{\mathbf{h}_{km}\}$ using the LM observations from $\{\mathbf{y}_m\}$. While it may seem appealing to use the low-dimensional channel model in (2) to facilitate channel estimation by reducing the number of unknowns to NKM, it is crucial to recognize that assuming a low-dimensional structure for channel estimation already confines the channel estimates within that pre-determined subspace, making it impossible to update this subspace.

In what follows, we consider learning the low-dimensional channel structure directly from the received signals.

Notice that the covariance matrix of the received signal $\mathbf{y}_m$ in (3) can be re-written as

$$\mathbf{\Sigma}_{\boldsymbol{\gamma},\mathbf{R}} = \mathbf{C}(\boldsymbol{\gamma}) \odot \mathbf{R} + \sigma^2\mathbf{I}, \tag{11}$$

where
$$\mathbf{C}(\boldsymbol{\gamma}) \triangleq \sum_{k\in[K]} \gamma_k \boldsymbol{\phi}_k \boldsymbol{\phi}_k^{\mathsf{H}}, \tag{12}$$

and $\odot$ represents the Hadamard (element-wise) matrix product. It might appear attempting to jointly estimate $\boldsymbol{\gamma}$ and $\mathbf{R}$. For instance, the joint ML estimation can be formulated as

$$\begin{aligned} \min_{\boldsymbol{\gamma}\in\Gamma, \mathbf{R}\in\mathbb{S}_+^L} \quad & \log|\mathbf{\Sigma}_{\boldsymbol{\gamma},\mathbf{R}}| + \operatorname{tr}(\mathbf{\Sigma}_{\boldsymbol{\gamma},\mathbf{R}}^{-1}\widehat{\mathbf{\Sigma}}) \\ \text{s.t.} \quad & \operatorname{rank}(\mathbf{R}) \leq N. \end{aligned} \tag{13}$$

Assuming an estimate of $\mathbf{R}$ from previous transmissions, denoted by $\widehat{\mathbf{R}}$, we may consider the following alternating procedure to estimate $\boldsymbol{\gamma}$ and update $\widehat{\mathbf{R}}$: 1) find an activity estimate $\widehat{\boldsymbol{\gamma}}$ while fixing $\widehat{\mathbf{R}}$ by running the coordinate descent algorithm in Section III-C, then 2) re-estimate the channel covariance while keeping $\widehat{\boldsymbol{\gamma}}$ fixed (to be discussed in the Appendix). These two steps can be repeated for several iterations.

However, the joint estimation problem appears to be ill-conditioned, and our experiments have not yielded much success. We suspect this is because the estimation of $\boldsymbol{\gamma}$ and $\mathbf{R}$ are two conflicting objectives that require pilots with contradicting properties. When estimating $\boldsymbol{\gamma}$, one generally prefers pilot sequences with low cross-correlation to better distinguish users by their unique pilots. However, sending pilots with low cross-correlation typically results in a diagonally dominant $\mathbf{C}(\boldsymbol{\gamma})$ in (12), while the off-diagonal elements in $\mathbf{\Sigma}_{\boldsymbol{\gamma},\mathbf{R}}$ have low magnitudes and become sensitive to noise and estimation error of $\boldsymbol{\gamma}$. On the other hand, when the focus shifts to the estimation of $\mathbf{R}$, using the all-one sequence, i.e., $\boldsymbol{\phi}_k = \mathbf{1}$, becomes an obvious choice, as the weight matrix becomes $\mathbf{C}(\boldsymbol{\gamma}) = (\sum_{k\in[K]} \gamma_k)\mathbf{1}\mathbf{1}^{\mathsf{T}}$, reducing the Hadamard product with $\mathbf{C}(\boldsymbol{\gamma})$ to scalar multiplication with $\sum_{k\in[K]} \gamma_k$. However, by forcing all users to send the same pilot sequence, user identification becomes impossible.

Exploring an optimal tradeoff between activity detection and low-dimensional structure learning through pilot design presents a formidable challenge. We will not investigate further in this direction and will instead restrict our focus to the use of dedicated all-one pilot sequences for channel covariance estimation. When using the all-one pilots, the learning of low-dimensional channel structure is straightforward. We first

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3486200

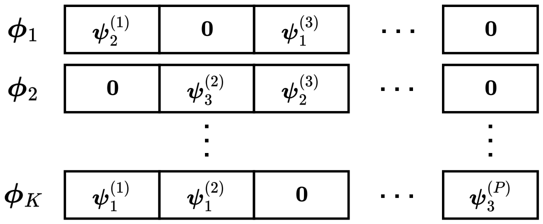

$\boldsymbol{\phi}_1$
$\boldsymbol{\psi}_2^{(1)}$
$\mathbf{0}$
$\boldsymbol{\psi}_1^{(3)}$
$\cdots$
$\mathbf{0}$
$\boldsymbol{\phi}_2$
$\mathbf{0}$
$\boldsymbol{\psi}_3^{(2)}$
$\boldsymbol{\psi}_2^{(3)}$
$\cdots$
$\mathbf{0}$
$\vdots$
$\vdots$
$\boldsymbol{\phi}_K$
$\boldsymbol{\psi}_1^{(1)}$
$\boldsymbol{\psi}_1^{(2)}$
$\mathbf{0}$
$\cdots$
$\boldsymbol{\psi}_3^{(P)}$

Fig. 1: Pilot hopping as integrating a special structure in pilots.

perform the EVD on $\widehat{\boldsymbol{\Sigma}} - \sigma^2\mathbf{I}$ and take the N dominant eigenvalues $\{\varsigma_n\}$ and the corresponding eigenvectors $\{\boldsymbol{\mu}_n\}$. The nth basis vector of the low-dimensional space is taken as

$$\mathbf{g}_n = \sqrt{\frac{L\varsigma_n}{\sum_{n'\in[N]}\varsigma_{n'}}}\boldsymbol{\mu}_n, \tag{14}$$

where the scaling factor makes the small-scale fading coefficients unit-variance, i.e., $h_{nkm} \sim \mathcal{CN}(0,1)$. The corresponding channel covariance is $\mathbf{R} = \sum_{n\in[N]}\mathbf{g}_n\mathbf{g}_n^\mathsf{H}$. Notice that we do not need to know the user activities when estimating $\{\mathbf{g}_n\}$ since the user activities introduce only a scaling factor which will be removed by the rescaling in (14).

The drawback of using dedicated all-one pilots for low-dimensional structure learning is that the users have to make additional transmissions, as those pilots cannot be reused for activity detection. However, as will be elaborated in Section V-C, this overhead can be substantially reduced when combined with pilot hopping.

V. Extension: Pilot Hopping

Another scheme to combat the limited channel coherence in activity detection is to apply pilot hopping, which has been considered in [16]. Specifically, the L samples in a channel block are partitioned into P sub-blocks of size τ, with $L = P\tau$ for simplicity. Within each sub-block $p \in [P]$, J sequences $\boldsymbol{\psi}_1^{(p)}, \cdots, .\boldsymbol{\psi}_J^{(p)}$ of length τ are generated, and we refer to $\boldsymbol{\psi}_j^{(p)}$ as the jth sub-pilot in the pth sub-block. Each user k is assigned a unique hopping pattern $\mathbf{z}_k \triangleq [z_k^{(1)}, \cdots, z_k^{(P)}]^\mathsf{T}$, where $z_k^{(p)} \in \{0\} \cup [J]$. If $z_k^{(p)} \in [J]$, user k transmits the $z_k^{(p)}$th sub-pilot in the pth sub-block when it is active; no pilot transmission if $z_k^{(p)} = 0$. By defining the $J \times K$ sequence selection matrices $\{\mathbf{U}^{(p)}\}$ with $[\mathbf{U}^{(p)}]_{j,k} = \mathbb{1}\{z_k^{(p)} = j\}$, the pilot matrix can be written as

$$\boldsymbol{\Phi} = \begin{bmatrix}\boldsymbol{\Phi}^{(1)}\\ \vdots\\ \boldsymbol{\Phi}^{(P)}\end{bmatrix}, \quad \boldsymbol{\Phi}^{(p)} \triangleq \boldsymbol{\Psi}^{(p)}\mathbf{U}^{(p)}, \tag{15}$$

where $\boldsymbol{\Psi}^{(p)} \triangleq [\boldsymbol{\psi}_1^{(p)}, \cdots, \boldsymbol{\psi}_J^{(p)}]$. From (15), we can observe that the pilot hopping scheme can be seen as introducing a special structure in the pilot matrix; see Fig. 1.

In [16], the channel is assumed to be quasi-static so that it stays constant within each sub-block and changes independently across sub-blocks. The sub-pilots are made mutually orthogonal (i.e., $J = \tau$) and the users are distinguished by their unique hopping patterns. A practical activity detection algorithm for this scheme is proposed in [17], relying on the channel hardening and the favorable propagation properties of massive MIMO. More specifically, as shown in [17], when the number of receive antennas $M \to \infty$, the received energies on $\boldsymbol{\psi}_j^{(p)}$ has the form

$$e_j^{(p)} \to \underbrace{[\mathbb{1}\{z_1^{(p)} = j\}\tau\beta_1, \cdots, \mathbb{1}\{z_K^{(p)} = j\}\tau\beta_K]}_{\triangleq(\boldsymbol{\omega}_j^{(p)})^\mathsf{T}}\mathbf{a}. \tag{16}$$

By stacking $\{e_j^{(p)}\}$ into a vector $\mathbf{e}$, the relation can be written as $\mathbf{e} \xrightarrow{M\to\infty} \boldsymbol{\Omega}\mathbf{a}$, where $\boldsymbol{\Omega}$ has rows $\{(\boldsymbol{\omega}_j^{(p)})^\mathsf{T}\}$. The $P\tau \times K$ matrix $\boldsymbol{\Omega}$ can be seen as a sparse measurement (sensing) matrix, and a corresponding CS problem can be formulated to recover the user activities $\mathbf{a}$:

$$\min_{\mathbf{a}\in[0,1]^K} \; \|\boldsymbol{\Omega}\mathbf{a} - \mathbf{e}\|_2^2 + \lambda\|\mathbf{a}\|_1, \quad \lambda \geq 0. \tag{17}$$

Notice that (17) is a least absolute shrinkage and selection operator (LASSO) problem that can be solved using standard optimization toolboxes (e.g., MOSEK [26]). Furthermore, as suggested in [17], the matrix $\boldsymbol{\Omega}$ is self-regularizing so that the regularization term $\lambda\|\mathbf{a}\|_1$ can be removed. The problem then reduces to a non-negative least squares (NNLS) problem, which is solved by the Lawson-Hanson algorithm [27] in [17].

A. Pilot Hopping in Covariance-Based Activity Detection

It is straightforward to combine the covariance-based approach with pilot hopping. To do this, we ignore the correlation between different channel sub-blocks and keep only the diagonal blocks in the channel covariance matrix, i.e.,

$$\mathbf{R} \approx \mathrm{bdiag}(\mathbf{R}^{(1)}, \cdots, \mathbf{R}^{(P)}), \tag{18}$$

where $\mathbf{R}^{(p)}$ is the channel covariance within the pth sub-block that is assumed to be known and to have the Gramian factorization $\mathbf{R}^{(p)} = \mathbf{G}^{(p)}(\mathbf{G}^{(p)})^\mathsf{H}$. A motivation for ignoring the inter-block correlation is to avoid overfitting the learned low-dimensional structure to the training data, which might be generated from an oversimplified channel model that does not generalize well to real-world radio environments. To see this, notice that one can always learn a perfect two-dimensional representation of the standard two-ray model in [11, Fig. 2.1] on an infinitely large time-frequency block, while the learned representation is most probably not useful in practice.

By utilizing the (assumed) block-diagonal structure of $\mathbf{R}$, the signal covariance can be expressed as

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \mathrm{bdiag}\big(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{(1)}, \cdots, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{(P)}\big), \tag{19}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{(p)} \triangleq \sum_{k\in[K]}\gamma_k\mathbf{S}_k^{(p)}(\mathbf{S}_k^{(p)})^\mathsf{H}$ with $\mathbf{S}_k^{(p)} = \boldsymbol{\phi}_k^{(p)}\mathbf{G}^{(p)}$. Here, $\boldsymbol{\phi}_k^{(p)}$ is the pth sub-block in the pilot sequence of user k, i.e., the kth column of $\boldsymbol{\Phi}^{(p)}$ in (15). The ML estimation of $\boldsymbol{\gamma}$ can then be formulated as

$$\min_{\boldsymbol{\gamma}\in\Gamma} \; \widetilde{f}(\boldsymbol{\gamma}) \triangleq \sum_{p=1}^{P} f^{(p)}(\boldsymbol{\gamma}), \tag{20}$$

$$\text{where} \quad f^{(p)}(\boldsymbol{\gamma}) \triangleq \log\big|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{(p)}\big| + \mathrm{tr}\Big(\big(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{(p)}\big)^{-1}\widehat{\boldsymbol{\Sigma}}^{(p)}\Big) \tag{21}$$

Algorithm 2 Activity Detection with Pilot Hopping

```
Input: sample covariances $\{\widehat{\mathbf{\Sigma}}^{(p)}\}$, effective pilots $\{\mathbf{S}_k^{(p)}\}$,
    and hopping patterns $\{\mathcal{P}_k\}$
Initialize: $\gamma_k \leftarrow 0, \forall k$ and $(\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{(p)})^{-1} \leftarrow \sigma^{-2}\mathbf{I}, \forall p$
 1: for $i = 1, 2, \cdots, I$ do
 2:     Generate $\mathcal{K}$ by randomly permuting $[K]$
 3:     for $k$ taken from $\mathcal{K}$ by order do
 4:         $\forall p \in \mathcal{P}_k$: $\mathbf{V}^{(p)}\mathbf{D}_{\boldsymbol{\lambda}}^{(p)}\mathbf{V}^{(p)\mathsf{H}} \xleftarrow{\text{EVD}} \mathbf{S}_k^{(p)\mathsf{H}}(\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{(p)})^{-1}\mathbf{S}_k^{(p)}$
 5:         $\forall p \in \mathcal{P}_k$: $\widetilde{\mathbf{V}}^{(p)} \leftarrow (\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{(p)})^{-1}\mathbf{S}_k^{(p)}\mathbf{V}^{(p)}$
 6:         $\forall p \in \mathcal{P}_k, \forall n$: $\xi_n^{(p)} \leftarrow [\widetilde{\mathbf{V}}^{(p)\mathsf{H}}\widehat{\mathbf{\Sigma}}^{(p)}\widetilde{\mathbf{V}}^{(p)}]_{n,n}$
 7:         $d^* = \operatorname{argmin} \widetilde{\mathcal{L}}_k(d)$ for $d \in [-\gamma_k, \gamma_{\max} - \gamma_k]$
 8:         $\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{(p)-1} \leftarrow \mathbf{\Sigma}_{\boldsymbol{\gamma}}^{(p)-1} - d^*\widetilde{\mathbf{V}}^{(p)}(\mathbf{I} + d^*\mathbf{D}_{\boldsymbol{\lambda}}^{(p)})^{-1}\widetilde{\mathbf{V}}^{(p)\mathsf{H}}$
 9:         $\gamma_k \leftarrow \gamma_k + d^*$
10:     end for
11: end for
Output: activity estimate $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_K]^{\mathsf{T}}$
```

with $\widehat{\mathbf{\Sigma}}^{(p)}$ being the diagonal block in the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ corresponding to the pth sub-block.

It is clear that (21) has the same form as (4), and the same techniques from Section III-C can be applied here to obtain the cost function for updating the kth coordinate as

$$\widetilde{\mathcal{L}}_k(d) \triangleq \sum_{p\in\mathcal{P}_k}\sum_{n\in[N]}\left(\log(1+d\lambda_{kn}^{(p)}) - \frac{d\xi_{kn}^{(p)}}{1+d\lambda_{kn}^{(p)}}\right), \quad (22)$$

where $\mathcal{P}_k$ denotes the set of sub-blocks in which user k is allocated a sub-pilot, i.e., $\mathcal{P}_k = \{p : z_k^{(p)} \neq 0\}$. $\{\lambda_{kn}^{(p)}\}$ and $\{\xi_{kn}^{(p)}\}$ are calculated similarly as $\{\lambda_{kn}\}$ and $\{\xi_{kn}\}$ in (7) for each channel sub-block. Since (22) has the same structure as (7), it can be optimized in the same way - either comparing the roots of its first-order derivative and the boundary points or performing a one-dimensional search.

The detection procedure is summarized in Algorithm 2.

Complexity: The matrix multiplications in steps 4, 5, 6, and 8 have complexity $\mathcal{O}(|\mathcal{P}_k|\tau^2 N)$. The EVD in step 4 has complexity $\mathcal{O}(|\mathcal{P}_k|N^3)$. However, as the derivative of $\widetilde{\mathcal{L}}_k(d)$ in (22) is a polynomial of degree $2|\mathcal{P}_k|N-1$, finding the root requires $\mathcal{O}(|\mathcal{P}_k|^3N^3)$ arithmetic operations. In case that $|\mathcal{P}_k|N$ is large, employing a one-dimensional search can be more efficient, as the complexity does not scale with $|\mathcal{P}_k|$ nor N. The overall complexity of Algorithm 2 is $\mathcal{O}(IK|\mathcal{P}_k|N(\tau^2 + N^2))$. Notice that Algorithm 2 can be parallelized when the users can be partitioned into groups that transmit in disjoint sub-blocks. For example, when each user transmits in only one sub-block, i.e., $|\mathcal{P}_k| = 1, \forall k$, the BS can perform activity detection in each sub-block separately.

B. Hopping Pattern Design

Another critical design aspect in pilot hopping-based activity detection is the selection of hopping patterns for each user. In [17], the hopping patterns are selected uniformly at random, where the patterns $\{z_k^{(p)}\}$ are generated in an i.i.d. manner with uniform distribution across $[J]$. We modify this method to restrict each user to selecting only D sub-blocks, i.e., $|\mathcal{P}_k| = D, \forall k$, to control the average overload factor $DK/(P\tau)$ in each sub-block. The method in [17] can be seen as a special case when $D = P$.

Algorithm 3 Hopping Pattern Generation

```
Input: user vertices $\mathcal{U}$, sub-pilot vertices $\{\mathcal{V}^{(p)}\}$, and $D$
Initialize: $\mathcal{E}^{(p)} \leftarrow \emptyset, \forall p \in [P]$
for $u \in \mathcal{U}$ do
    $\widetilde{\mathcal{P}} \leftarrow \emptyset, i \leftarrow 0$
    while $|\widetilde{\mathcal{P}}| < D$ do
        $\widetilde{\mathcal{P}} \leftarrow \{p : |\mathcal{E}^{(p)}| \leq \min_{p'\in[P]} |\mathcal{E}^{(p')}| + i\}$
        $i \leftarrow i + 1$
    end while
    while $\deg(u) < D$ do
        choose $p \in \widetilde{\mathcal{P}}$ randomly and remove $p$ from $\widetilde{\mathcal{P}}$
        $\widetilde{\mathcal{V}}^{(p)} \leftarrow \operatorname{argmin}_{v\in\mathcal{V}^{(p)}} \deg(v)$
        choose $v \in \widetilde{\mathcal{V}}^{(p)}$ randomly and add $(u, v)$ to $\mathcal{E}^{(p)}$
    end while
end for
Output: $\{\mathcal{E}^{(p)}\}$
```

In the random pilot hopping generation scheme, however, the number of users selecting different sub-blocks, as well as sub-pilots within each sub-block, can vary significantly. To address this imbalance in the hopping system, we propose a new hopping pattern generation scheme, inspired by the configuration model in random graph generation [18, Ch. 5.3]. Specifically, the system can be seen as a bipartite graph $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, as shown in Fig. 2, where the first vertex set $\mathcal{U}$ denotes the users, the second vertex set $\mathcal{V} \triangleq \cup_{p\in[P]}\mathcal{V}^{(p)}$ represents sub-pilots with $\mathcal{V}^{(p)}$ being the sub-pilots in the pth sub-block. We aim to generate the edge set $\mathcal{E} \triangleq \cup_{p\in[P]}\mathcal{E}^{(p)}$ with $\mathcal{E}^{(p)}$ representing the edges between $\mathcal{V}^{(p)}$ and $\mathcal{U}$ with all sub-pilot vertices in $\mathcal{V}$ having nearly the same degrees, differing at most by one, i.e., $\lfloor DK/(PJ)\rfloor \leq \deg(v) \leq \lfloor DK/(PJ)\rfloor + 1, \forall v \in \mathcal{V}$, all sub-blocks having nearly the same number of edges, i.e., $\lfloor DK/P\rfloor \leq |\mathcal{E}^{(p)}| \leq \lfloor DK/P\rfloor + 1, \forall p \in [P]$, and each user vertex in $\mathcal{U}$ having degree D with at most one edge to each $\mathcal{V}^{(p)}$. The procedure for generating $\{\mathcal{E}^{(p)}\}$ is summarized in Algorithm 3. The algorithm iterates over a randomly permuted list of users and randomly chooses for each user D sub-blocks with the minimal number of edges and one sub-pilot with the lowest degree in each selected sub-block to form a new edge. The conversion from the edge lists $\{\mathcal{E}^{(p)}\}$ to the hopping patterns $\{z_k^{(p)}\}$ is straightforward: if there is an edge connecting the vertex corresponding to user k and the vertex for the jth sub-pilot in the pth sub-block, we set $z_k^{(p)} = j$. If no vertex in $\mathcal{V}^{(p)}$ is connected to the vertex representing user k, we assign $z_k^{(p)} = 0$.

C. The Complete Framework

We can combine the covariance estimation using dedicated, all-one pilots with the pilot hopping to reduce the overhead. We assume the channel is nearly wide-sense stationary and uncorrelated scattering (WSSUS) so that the diagonal blocks in the channel covariance $\mathbf{R}$ are approximately identical, i.e., $\mathbf{R}^{(1)} \approx \cdots \approx \mathbf{R}^{(P)} \triangleq \mathbf{R}_0$. Then, it becomes sufficient to use

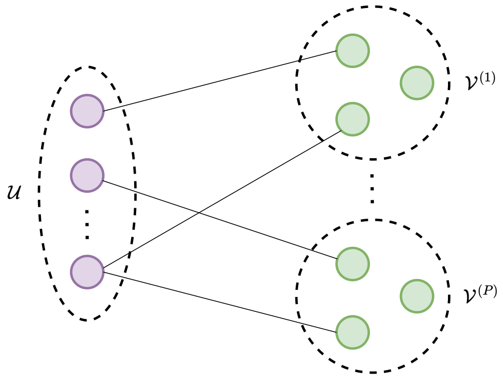


Fig. 2: Pilot hopping system as a bipartite graph.

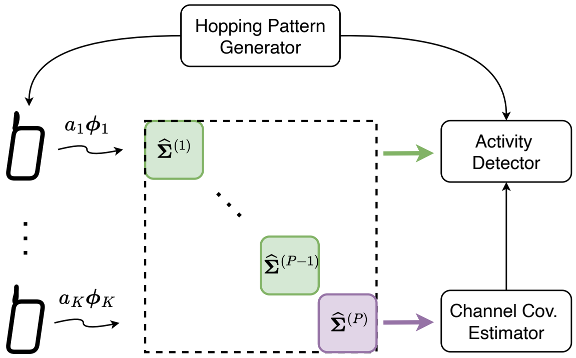

Hopping Pattern Generator
$a_1\boldsymbol{\phi}_1$
$\widehat{\boldsymbol{\Sigma}}^{(1)}$
$\widehat{\boldsymbol{\Sigma}}^{(P-1)}$
$\widehat{\boldsymbol{\Sigma}}^{(P)}$
$a_K\boldsymbol{\phi}_K$
Activity Detector
Channel Cov. Estimator

Fig. 3: The complete activity detection framework.

only one sub-block for covariance estimation, leaving the remaining ones for activity detection. Our detection framework, as shown in Fig. 3, is as follows: All active users send the all-one sequence in the Pth sub-block for the BS to estimate the channel covariance. In the remaining $P-1$ channel sub-blocks, these active users follow their unique hopping patterns for pilot transmission. The BS performs activity detection using the received pilot signals as well as the bases obtained from the covariance estimate.

Remark 4: The division of a channel block into multiple sub-blocks, and especially the size of these sub-blocks, need to be carefully designed. For strongly correlated fading, we should use larger sub-blocks to increase the gain from coherent processing of the samples. Conversely, in scenarios with less correlated fading, forming smaller sub-blocks helps exploit diversity. There is also a tradeoff between computational complexity and detection performance. From a data-driven perspective, we could first estimate the channel covariance matrix $\mathbf{R}$ and calculate the eigenvalues of its sub-blocks of different sizes to determine the sub-block size. We should select the sub-block size such that the number of dominant eigenvalues of the sub-blocks is relatively small, allowing for a good low-dimensional approximation. One numerical example is provided in Section VI-H, where we show that the selected sub-block size gives an accurate order-3 approximation of the channel covariance matrix. Alternatively, if the channel can be well approximated by an existing model, for example as in [28, Ch. 2], and the relevant parameters are known, we can choose the sub-block size based on the channel auto-correlation functions. In this paper we focus on the case where the channel is nearly WSSUS and the channel covariance needs to be frequently re-estimated. In case that the channel is not WSSUS but the channel covariance changes slowly so that re-estimation is infrequent, we can use more complicated strategies. One idea is to use the magnitudes of the elements in the channel covariance as edge weights to form a graph and apply graph partitioning methods (see, for example, [18, Ch. 9]) to determine the sub-block structure – in this way, the elements in the same sub-block are not necessarily adjacent to each other. The design of data-driven approaches to determine the optimal sub-block structure could be an interesting direction for future research.

VI. Numerical Results

A. Generation of Channels

We first generate the CIRs by using the improved sum-of-sinusoids method in [29]. Since the same procedure will be repeated for each user-antenna pair, we ignore the subscript "km". For a channel with N_{path} paths at different delays, we generate each path independently, and the time-varying (complex) amplitude of the ith path is given by

$$q_i(t)=\frac{1}{\sqrt{N_{\text{sin}}}}\sum_{n\in[N_{\text{sin}}]} e^{j(\omega_d t\cos\alpha_n+\psi_n)} \tag{23}$$

with $\alpha_n=\frac{2\pi n+\zeta_n}{N_{\text{sin}}}, n\in[N_{\text{sin}}]$, where N_{sin} is the number of sinusoids, ω_d is the maximum Doppler frequency in radians, ψ_n and ζ_n are i.i.d. distributed over $[-\pi,\pi)$.

We denote the sampling rate by B which is identical to the system bandwidth, and the impulse response of the pulse shaping filter as $p(\cdot)$. The discrete-time impulse response of the multipath-fading channel is given by

$$q_{t_l\ell}=\sum_{i\in[N_{\text{path}}]}\sqrt{c_i}q_i\left(\frac{t_l-1}{B}\right)p\left(\frac{\ell-\tau_i}{B}\right), \tag{24}$$

where $t_l\in[T]$, the fractional power c_i, and the delay τ_i of different paths are determined by a power delay profile defined in, for example, [30], [31]. Here, ℓ is the time-lag index, which is an integer ranging from $\ell_{\min}$ to $\ell_{\max}$. The smallest time-lag $\ell_{\min}$ is a negative integer, e.g., -6, and $\ell_{\max}=\lceil B\tau_{\text{exc}}\rceil-\ell_{\min}$ with $\tau_{\text{exc}}=\max_{i\in[N_{\text{path}}]}\{\tau_i\}$ being the maximum excess delay. For each t_l, we apply the DFT to $\mathbf{x}_{t_l}=[q_{t_l\ell_{\min}},\cdots,q_{t_l\ell_{\max}}]^{\mathsf{T}}$ to obtain the frequency response at $f_l\in[F]$ as

$$Q_{t_lf_l}=\sum_{i\in[\ell_{\max}-\ell_{\min}+1]}[\mathbf{x}_{t_l}]_i e^{-j\frac{2\pi(i-1)(f_l-1)}{N_{\text{sub}}}}, \tag{25}$$

where $N_{\text{sub}}\geq F$ is the total number of sub-carriers in this OFDM system (we use the first F sub-carriers for pilot transmission). These coefficients are arranged into the length-L channel vector $\mathbf{h}=[h_1,\cdots,h_L]^{\mathsf{T}}$ where h_l equals to $Q_{t_lf_l}$ with an invertible mapping from (t_l,f_l) to l. For simplicity, we assume that the total number of sub-blocks can be factorized as $P=P_{\text{time}}P_{\text{freq}}$, where P_{time} is an integer that divides the number of OFDM symbols T and P_{freq} divides the number of sub-carriers F. We use a mapping between (t_l,f_l) and l as illustrated in Fig. 4.

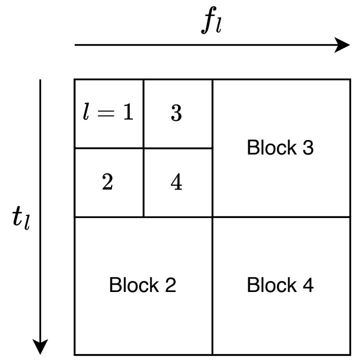


Fig. 4: The mapping from (t_l, f_l) to l for $T = F = 4$ and $P_{\text{time}} = P_{\text{freq}} = 2$.

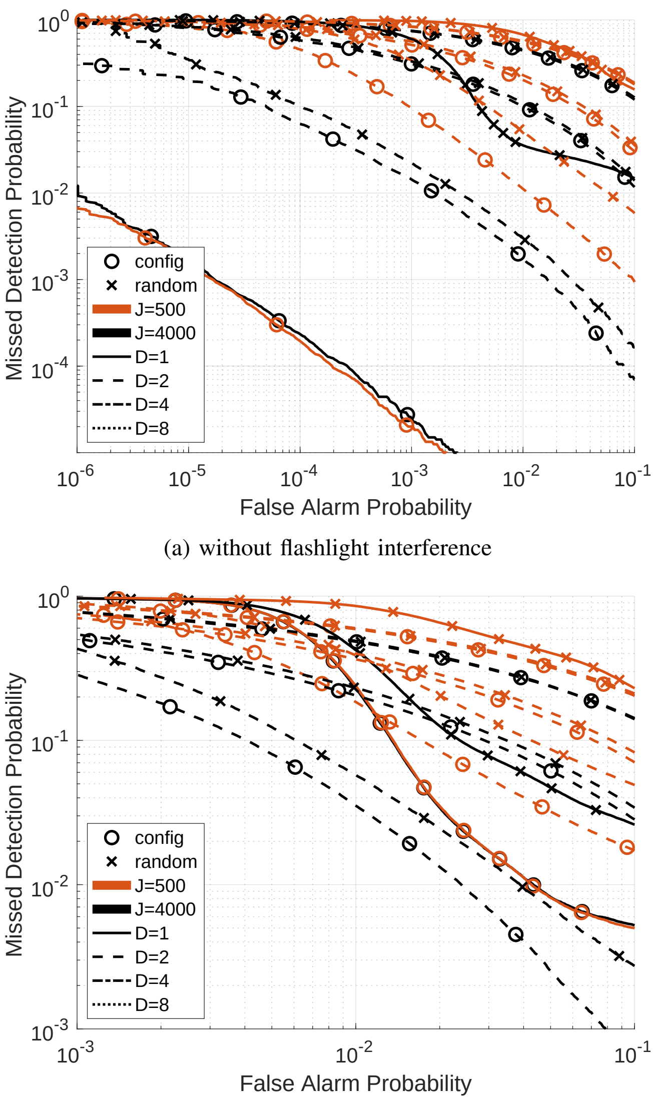
B. System Setup

We consider a single-cell network with $K = 4000$ users. Each user becomes active independently with a probability of $\epsilon = 0.1$. The BS has $M = 100$ receive antennas. We consider a peak power constraint and ideal channel inversion power control so that the signal-to-noise ratio (SNR) equals 0 dB for all users, i.e., $\beta_1/\sigma^2 = \cdots = \beta_K/\sigma^2 = 1$. The channels are generated using the procedure in Section VI-A with $T = 12$ and $F = 36$, resulting in a channel dimension of $L = 432$. The carrier frequency is set to 30 GHz which falls into the frequency range 2 in 5G New Radio. The sub-carrier spacing is set to 30 kHz [32]. The DFT size is set to 128. The number of sinusoids in the channel simulator is $N_{\text{sin}} = 20$. For pulse shaping, we use a root-raised-cosine (RRC) filter with a roll-off factor of 0.22. In each Monte-Carlo trial, the channel model is randomly selected from TDL-A, TDL-B, and TDL-C in [31, Sec. 7.7.2] and typical urban (TUx), rural area (RAx), and hilly terrain (HTx) in [30, Sec. 5], where the delays are randomly scaled to have a root-mean-square (RMS) delay spread between 0.5 and 1.5 microseconds. The mobile speed is randomly selected between 80 and 160 km/h. The time-frequency grid is partitioned into $P = 9$ channel sub-blocks with $P_{\text{time}} = P_{\text{freq}} = 3$. Unless otherwise stated, we use $D = 1$, $J = 4000$, and Algorithm 3 for hopping pattern generation. The sub-pilots are randomly generated from complex Gaussian and are normalized to have unit average energy per symbol. The performance of the detection algorithms is evaluated using receiver operating characteristic (ROC) curves, each one generated using 3000 independent Monte Carlo trials.

C. Is Pilot Hopping Useful? – A Case Study

The ROC curves of different pilot hopping schemes are presented in Fig. 5a. We observe that the hopping patterns generated by the configuration model, as described in Algorithm 3, significantly outperform the random patterns.

Choice of D: When using good hopping patterns, it does not appear advantageous to let users make multiple transmissions in different channel sub-blocks. Instead, the best detection performance is achieved when each active user transmits in only $D = 1$ sub-block, which suggests that one should simply partition the users into disjoint groups and allocate each group a dedicated channel sub-block – this scheme more closely resembles *scheduling* than hopping. One interpretation of this result is that when sufficient spatial diversity has been achieved by a large antenna array, it becomes unnecessary to explore extra time or frequency diversity at the cost of increased interference. The concurrent work [33] reveals a similar observation: one should not trade sparsity for diversity in scenarios with limited radio resources.



(a) without flashlight interference

(b) with flashlight interference

Fig. 5: Detection performance with different hopping schemes.

Choice of J: Since our approach combines pilot hopping with covariance-based activity detection, using orthogonal sub-pilots ($J = \tau$) with collision but no contamination, and unique sub-pilots ($J = K$) with no collision but severe contamination, respectively, it is of interest to explore the tradeoff between these two design principles. However, as suggested by our results, users should not reuse sub-pilots, and better performance can be achieved by assigning a unique sub-pilot to each user within every channel sub-block. (In our setup, users have unique sub-pilots when $J \geq DK/P = 500D$.)

While the findings so far might suggest that pilot hopping is not useful, this conclusion could vary based on the scenario.

For instance, in real-world situations, the receiver might experience “flashlight interference” from neighboring cells (due to imperfect coordination/scheduling in multi-cell networks; see [34]) or encounter accidental blocking effects, leading to signal contamination or weakening in certain channel sub-blocks. Under such circumstances, exploiting additional diversity through pilot hopping becomes important. Specifically, when a user transmits its pilot in only one of the sub-blocks, there is a possibility that its signal will be contaminated by this interference, leading to a higher probability of detection error. Conversely, if users utilize multiple sub-blocks to exploit diversity, even though their signals may be contaminated in one sub-block, the BS can still leverage the diversity in the other sub-blocks to effectively detect the users. In Fig. 5b, we recalculate the results from Fig. 5a, introducing 100 interfering users who transmit randomly generated complex Gaussian signals within a randomly selected channel sub-block. The results show that enabling users to transmit pilots over multiple sub-blocks enhances robustness against flashlight interference.

D. Comparison of Different Bases

In addition to the learned bases (referred to as the principal component analysis (PCA) bases henceforth), we consider three other schemes as baselines:[2] (We use block size $T \times F$ for illustration, which will be adjusted accordingly for pilot hopping. We use $N = 3$ unless otherwise stated.)

- Block-fading: the channel is assumed to be constant within each sub-block, i.e., $\mathbf{G} = \mathbf{1}_L$ and $N = 1$.
- BWL bases: the channel varies linearly over time and frequency. To be specific, consider two vectors
$$\mathbf{u}_T \propto [-1, -1 + 2/(T-1), \cdots, 1]^{\mathsf{T}} \in \mathbb{R}^T$$
$$\mathbf{u}_F \propto [-1, -1 + 2/(F-1), \cdots, 1]^{\mathsf{T}} \in \mathbb{R}^F$$
representing the linear variations, where $\propto$ means that the vectors are normalized, i.e., $\|\mathbf{u}_T\|^2 = T, \|\mathbf{u}_F\|^2 = F$. Let $\widetilde{\mathbf{G}} = [\mathbf{1}_L, \mathbf{1}_F \otimes \mathbf{u}_T, \mathbf{u}_F \otimes \mathbf{1}_T]$, where $\otimes$ represents the Kronecker product, and $\mathbf{1}_L$ accounts for the average channel gain. The basis matrix is given by $\mathbf{G} = \widetilde{\mathbf{G}}\mathbf{D}_{\mathbf{x}}^{\frac{1}{2}}$, where $\mathbf{x} \in \mathbb{R}^3$ denotes the variances after projecting the channel vectors onto each basis.
- DFT bases: We select $\mathbf{v}_T$ to be one column from the $T \times T$ DFT matrix, and $\mathbf{v}_F$ from the $F \times F$ DFT matrix. (We traverse all columns and choose the ones giving the best approximation.) Similar to the BWL case, we choose $\mathbf{G} = \widetilde{\mathbf{G}}\mathbf{D}_{\mathbf{x}}^{\frac{1}{2}}$ with $\widetilde{\mathbf{G}} = [\mathbf{1}_L, \mathbf{1}_F \otimes \mathbf{v}_T, \mathbf{v}_F \otimes \mathbf{1}_T]$.

Similar to the learning of the low-dimensional structure in Section VII, to determine the vector $\mathbf{x}$ in the BWL/DFT bases, we can form a target matrix $\boldsymbol{\Upsilon} \triangleq \widehat{\boldsymbol{\Sigma}}_0 - \sigma^2\mathbf{I}$, where $\widehat{\boldsymbol{\Sigma}}_0$ is the sample covariance after sending all-one pilots, and set $\mathbf{x} = \operatorname{diag}(\widetilde{\mathbf{G}}^{\mathsf{H}}\boldsymbol{\Upsilon}\widetilde{\mathbf{G}})$ which minimizes $\|\boldsymbol{\Upsilon} - \widetilde{\mathbf{G}}\mathbf{D}_{\mathbf{x}}\widetilde{\mathbf{G}}^{\mathsf{H}}\|_{\mathsf{F}}$.

The detection performance with different bases for the low-dimensional channel approximation is shown in Fig. 6. One can observe that the PCA-based bases outperform the competing models by a considerable margin. However, using a higher approximation order ($N = 5$) does not necessarily give significantly better performance than a lower approximation order ($N = 3$) even at the cost of increased computational complexity. This is because our knowledge of the low-dimensional channel subspace is imperfect. The higher-order bases may be subject to larger estimation noise.

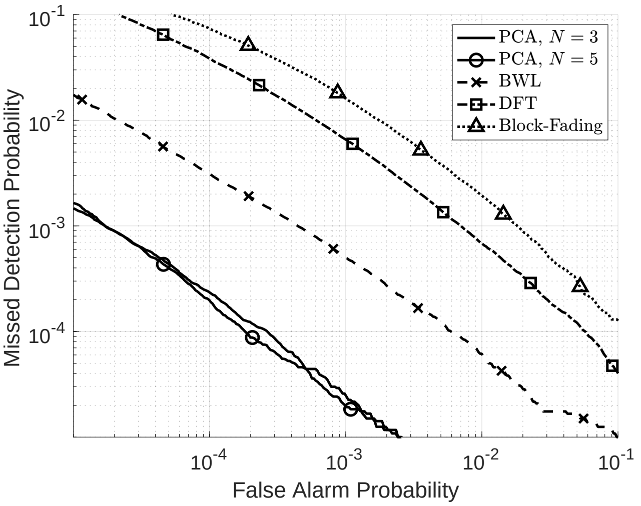


Fig. 6: Performance comparison of different bases.

E. Performance with Different Learning Overheads

It is of interest to know how much overhead one should spend on learning the low-dimensional channel structure. On the one hand, we want to learn the structure more accurately so that the activity detection algorithm can work better. On the other hand, due to the limited radio resources, we have fewer resources for activity detection as we spend more resources on structure learning. It is difficult to draw a general conclusion since there is no explicit objective function that can be feasibly optimized. Additionally, as discussed in Section IV, the pilot design can also play an important role in this problem, and our choice of all-one pilots is only a special case. In Fig. 7, we compare the detection performance when choosing different numbers of dedicated sub-blocks for low-dimensional channel structure learning. We achieve the best performance when using only one sub-block for structure learning. This suggests that by sending all-one pilots, we can estimate the low-dimensional structure with sufficient accuracy, and there is no need to use more sub-blocks.

F. Performance with Different Active Probabilities

In Fig. 8, we can observe that the detection performance degrades rapidly when users access the channel with higher probabilities. Furthermore, compared with a quasi-static block-fading channel, where the channel stays constant within each sub-block, the continuously varying channels make the detection problem more challenging, as fewer users can be simultaneously supported with the same amount of radio resources. This is unavoidable, as there is less coherence that can be exploited in the continuously varying channels, and also because additional radio resources need to be allocated to the learning of the low-dimensional channel structure.

[2] The BWL/DFT bases were used in [12] and [14], respectively, for frequency selectivity. We extend their methods by incorporating time variations.

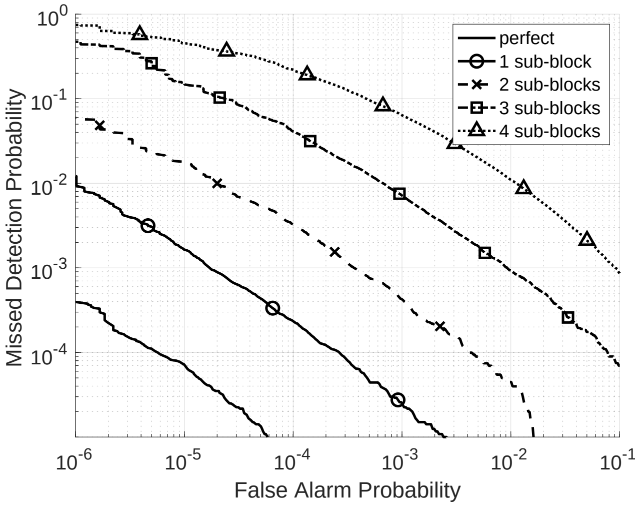


Fig. 7: Detection performance with different numbers of sub-blocks dedicated for low-dimensional structure learning.

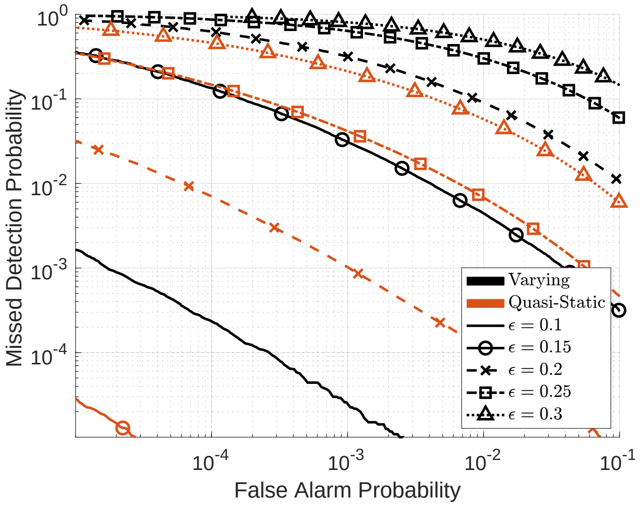


Fig. 8: Performance with different active probabilities ϵ.

G. Runtime

Activity detection using Algorithm 2 for 10 global iterations takes around 3 seconds for the system setup in Section VI-B. (We use a one-dimensional search to minimize (22) directly. The runtime was obtained on a standard PC using Matlab.)

We note that our implementation of the algorithms is by no means optimized for speed. There are potential improvements that could reduce runtime. For instance, as showcased in [35], the runtime of the coordinate decent algorithm can be significantly reduced by using properly designed computational architectures. Additionally, the active set selection algorithm in [36] could be used to avoid iterating over all users in each global iteration of Algorithm 2 to accelerate the coordinate descent algorithm. Designing more computationally efficient algorithms is an important direction for future research.

H. Visualization of Channels

We choose the TDL-B channel model with a mobile speed of 120 km/h and a RMS delay spread of 1 microsecond.

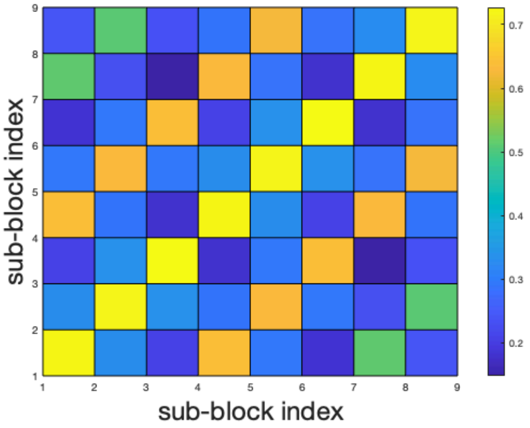
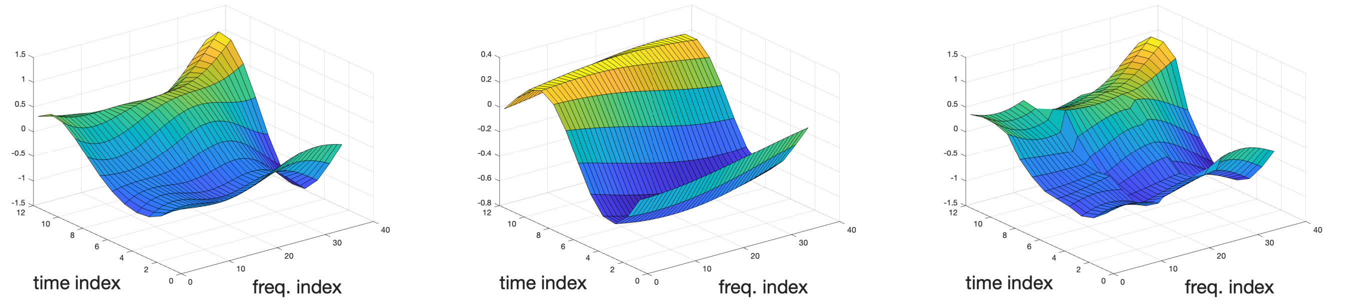


Fig. 9: The magnitude of elements in the channel covariance matrix averaged over each sub-block pair.

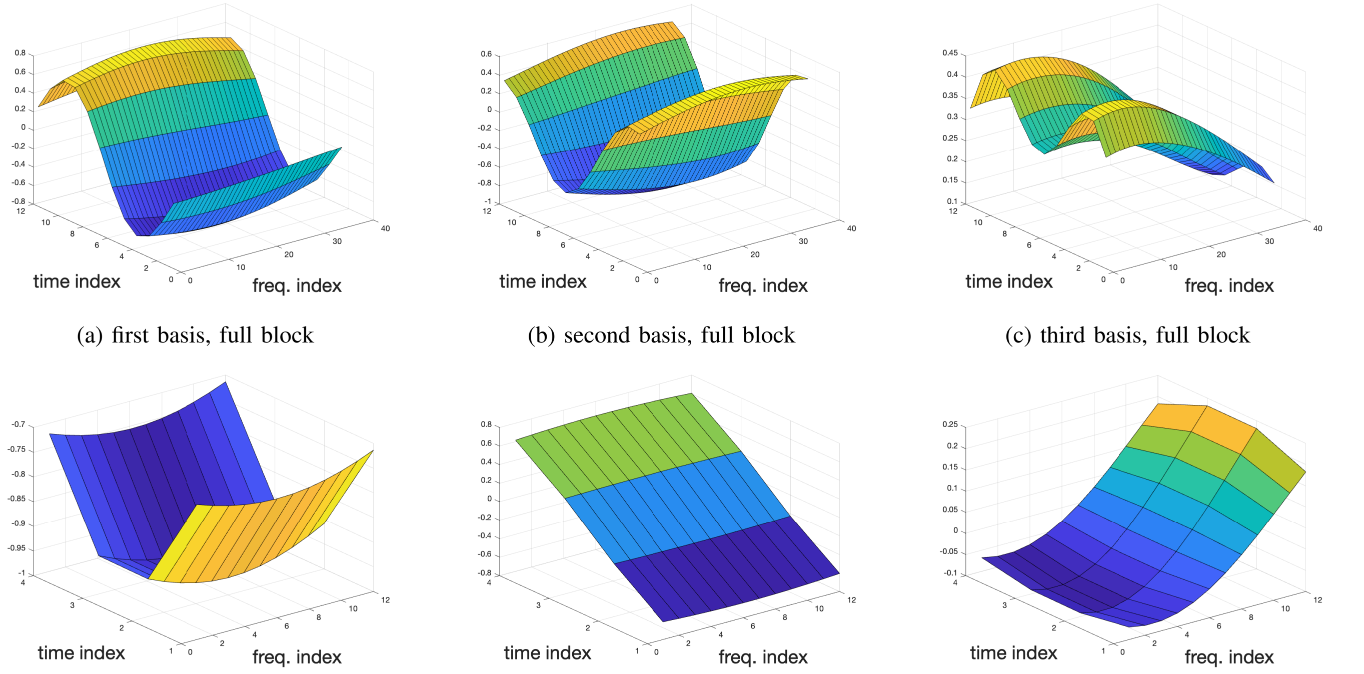
We generate a sufficiently large number ($KM = 10000$) of channel vectors using the procedure described in Section VI-A to form a sample covariance matrix that is an accurate estimate of the true channel covariance matrix. We first check the average magnitude of the corresponding elements in the sample covariance matrix for each sub-block pair. As shown in Fig. 9, the diagonal elements have dominant magnitudes, suggesting strong intra-block correlations and weak inter-block correlations. We consider two approaches to low-dimensional approximations. In the first approach, we directly perform the eigenvalue decomposition on the $L \times L$ sample covariance matrix to obtain the $L \times N$ basis matrix $\mathbf{G}$ in (2) with its columns as the dominant eigenvectors (scaled by the square roots of their corresponding eigenvalues). In the second approach, we take the average of the diagonal blocks in the sample covariance matrix corresponding to the channel sub-blocks. This new matrix has dimension $(L/P) \times (L/P)$, and we perform the eigenvalue decomposition on it to obtain a $(L/P) \times N$ basis matrix for approximating the channel variations in the sub-blocks.

One instance of the true channel (more specifically, its real part), its brute-forth full-block approximation, and the per-sub-block approximation are visualized in Fig. 10. The bases are visualized in Fig. 11. One can observe that by using the per-sub-block approximation, we obtain much simpler bases and a more accurate approximation. To quantify the approximation accuracy, we define $\mathbf{H} = [\cdots, \mathbf{h}_{km}, \cdots]$ as an $L \times KM$ matrix consisting of the fading coefficients of all user-antenna pairs, $\widehat{\mathbf{H}} = [\cdots, \widehat{\mathbf{h}}_{km}, \cdots]$ as the channels obtained by the low-dimensional approximation, and $\overline{\mathbf{H}} = [\cdots, \overline{\mathbf{h}}_{km}, \cdots]$ with $\overline{\mathbf{h}}_{km} = (\frac{1}{L} \sum_{l \in [L]} h_{lkm})\mathbf{1}$ as the block-fading approximation. We define the metric $\kappa = \|\widehat{\mathbf{H}} - \mathbf{H}\|_{\mathsf{F}} / \|\overline{\mathbf{H}} - \mathbf{H}\|_{\mathsf{F}}$ as the relative approximation error compared with the block-fading model. The value of κ for different approximation order N is depicted in Fig. 12. The per-sub-block approximation achieves a much more accurate approximation.



(a) true channel (b) approx. (c) approx. per sub-block

Fig. 10: Channel visualization and its order-3 approximation.



(a) first basis, full block (b) second basis, full block (c) third basis, full block

(d) first basis, per sub-block (e) second basis, per sub-block (f) third basis, per sub-block

Fig. 11: Visualization of the basis vectors.

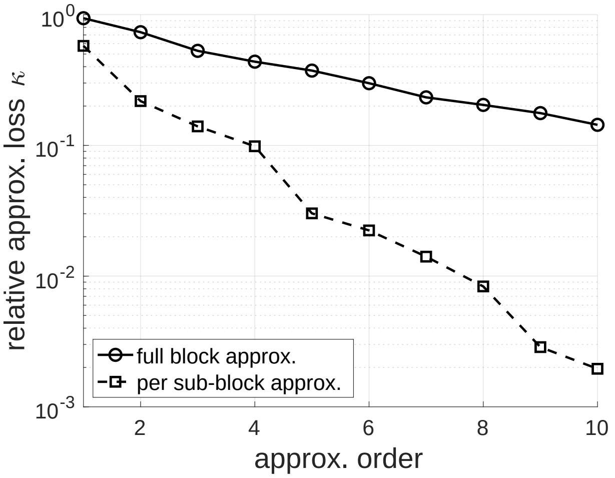


Fig. 12: The value of κ with different approximation order N.

VII. Conclusion

In this paper, we introduce a unified framework for robust user activity detection for massive access. Instead of assuming a block-fading channel as in most existing work, our framework allows for symbol-by-symbol variations of the channel by exploring a low-dimensional representation of the variations. This low-dimensional structure can be learned directly from the received pilot signals, and it provides considerable performance improvement compared with several existing baselines. Another important component in our framework is pilot hopping which allows users to explore extra time- and/or frequency diversity. Through case studies, we show that pilot hopping can improve the robustness of activity detection in the presence of flashlight interference or accidental blocking effects. However, we also observe that when those effects are not present, partitioning users into disjoint groups with dedicated radio resources achieves significantly better performance – one should not trade sparsity for diversity in this case. The choice of hopping patterns is also critical – carefully designed patterns work considerably better than randomly generated ones.

Appendix: WLRMA for Covariance Estimation

We consider the estimation of $\mathbf{R} = \mathbf{G}\mathbf{G}^{\mathsf{H}}$ when an activity estimate $\widehat{\boldsymbol{\gamma}}$ is given, without restriction to the use of all-one pilots. We note that this approach is not used in the proposed activity detection framework, as explained earlier. We include this section mainly for completeness and to describe potential directions for future research.

For a weight matrix $\mathbf{C}$, we consider the WLRMA problem

$$\begin{aligned}\min_{\mathbf{X}\in\mathbb{C}^{L\times N}} J(\mathbf{X};\mathbf{C}) &\triangleq \|\mathbf{C}\odot(\boldsymbol{\Upsilon}-\mathbf{X}\mathbf{X}^{\mathsf{H}})\|_{\mathsf{F}}^2\\ &= \sum_{i,j\in[L]} |[\mathbf{C}]_{i,j}|^2 \left|[\boldsymbol{\Upsilon}-\mathbf{X}\mathbf{X}^{\mathsf{H}}]_{i,j}\right|^2,\end{aligned} \tag{26}$$

with

$$\boldsymbol{\Upsilon} \triangleq (\widehat{\boldsymbol{\Sigma}} - \sigma^2\mathbf{I}) \oslash \mathbf{C}(\widehat{\boldsymbol{\gamma}}), \tag{27}$$

where $\oslash$ denotes the Hadamard (element-wise) division. One can observe from (26) that the problem depends only on the magnitude of the entries in $\mathbf{C}$. Therefore, it is sufficient to consider only real-valued matrices $\{\mathbf{C}\}$ with non-negative entries. When the weight matrix $\mathbf{C}$ is constructed by assigning each entry the absolute value of the corresponding entry of $\mathbf{C}(\widehat{\boldsymbol{\gamma}})$ in (12), denoted $\mathbf{C} = \mathbf{C}(\widehat{\boldsymbol{\gamma}})_{\text{abs}}$, the solution of (26) gives the desired estimate of the basis matrix $\mathbf{G}$ in (2), i.e., $\widehat{\mathbf{G}} = \operatorname{argmin} J(\mathbf{X};\mathbf{C}(\widehat{\boldsymbol{\gamma}}))$. Notice also that due to the Hadamard division in (27), when $\mathbf{C}(\widehat{\boldsymbol{\gamma}})$ has entries with small magnitudes, the noise in $\widehat{\boldsymbol{\Sigma}}$ will be magnified, and the objective becomes sensitive to the estimation error in $\widehat{\boldsymbol{\gamma}}$. (Strictly speaking, the Hadamard division is undefined when $\mathbf{C}(\widehat{\boldsymbol{\gamma}})$ has zero entries. But this case happens with zero probability when the pilots are randomly drawn from a continuous distribution.) This partially explains the ill-conditioning of the joint estimation problem.

1) EM Algorithm: An expectation maximization (EM) procedure for the WLRMA problem was developed in [37]. The algorithm performs the following update in each iteration:

$$\mathbf{X}_{i+1} \leftarrow \mathrm{LRA}_N(\widetilde{\mathbf{C}}\odot\boldsymbol{\Upsilon} + (1-\widetilde{\mathbf{C}})\odot(\mathbf{X}_i\mathbf{X}_i^{\mathsf{H}})), \tag{28}$$

where

$$\mathrm{LRA}_N(\mathbf{A}) \triangleq \operatorname*{argmin}_{\mathbf{X}\in\mathbb{C}^{L\times N}} \|\mathbf{A}-\mathbf{X}\mathbf{X}^{\mathsf{H}}\|_{\mathsf{F}}, \tag{29}$$

and $\widetilde{\mathbf{C}}$ is obtained by scaling $\mathbf{C}$ so that the maximum element equals one, i.e., $\widetilde{\mathbf{C}} \triangleq \mathbf{C}/\max\{[\mathbf{C}]_{i,j}\}$. Notice that the EM procedure in (28) does not necessarily find the optimal solution due to the non-convexity of (26).

2) Sequential Approximation: As it is difficult to directly solve (26) for $\mathbf{C} = \mathbf{C}(\widehat{\boldsymbol{\gamma}})_{\text{abs}}$, the authors of [38] proposed to solve (26) sequentially for a series of weight matrices $\mathbf{C}_0,\cdots,\mathbf{C}_I$. The first weight matrix is chosen to be $\mathbf{C}_0 = \mathbf{1}\mathbf{1}^{\mathsf{T}}$ and the corresponding optimal solution $\mathbf{X}_0$ can be easily obtained as $\mathrm{LRA}_N(\boldsymbol{\Upsilon})$. (When a good estimate $\widehat{\mathbf{G}}_{\text{old}}$ has been obtained from previous transmissions and if the channel covariance does not change abruptly, one may also use the corresponding $\mathbf{C}(\widehat{\boldsymbol{\gamma}}_{\text{old}})$ as $\mathbf{C}_0$ and use the corresponding estimate as the initial solution.) The final weight matrix is chosen as $\mathbf{C}_I = \mathbf{C}(\widehat{\boldsymbol{\gamma}})_{\text{abs}}$ so that its optimal solution $\mathbf{X}_I$ is the desired basis matrix. The remaining weight matrices $\{\mathbf{C}_i\}_{i=1}^{I-1}$ are selected as convex combinations of $\mathbf{C}_0$ and $\mathbf{C}_I$, i.e.,

Algorithm 4 Sequential Approximation for WLRMA

```
Input: sample covariance $\widehat{\boldsymbol{\Sigma}}$, and $\{\mathbf{C}_i\}_{i=1}^I$ in (30)
Initialize: $\boldsymbol{\Upsilon} \leftarrow (\widehat{\boldsymbol{\Sigma}} - \sigma^2\mathbf{I}) \oslash \mathbf{C}(\widehat{\boldsymbol{\gamma}})$
    $\mathbf{X}_0 \leftarrow \mathrm{LRA}_N(\boldsymbol{\Upsilon})$ with $\mathrm{LRA}_N(\cdot)$ defined in (29)
for $i = 0, 1, \cdots I-1$ do
    Obtain $\Delta_i$ by solving the QCQP in (32)
    $\mathbf{X}_{i+1} \leftarrow \mathbf{X}_i + \Delta_i$
end for
Output: basis matrix $\mathbf{G} = \mathbf{X}_I$
```

$$\mathbf{C}_i = \frac{I-i}{I}\mathbf{C}_1 + \frac{i}{I}\mathbf{C}_I, \quad 1 \le i \le I-1. \tag{30}$$

The key idea in [38] is that the minimizer of $J(\mathbf{X};\mathbf{C})$ is a continuous function of $\mathbf{C}$. And the difference $\|\mathbf{X}_{i+1}-\mathbf{X}_i\|$ can be made arbitrarily small if $\|\mathbf{C}_{i+1}-\mathbf{C}_i\|$ is sufficiently small. This can be guaranteed when I is sufficiently large.

By substituting $\mathbf{X}_{i+1} = \mathbf{X}_i + \Delta_i$, the WLRMA problem for the weight matrix $\mathbf{C}_{i+1}$ can be reformulated as

$$\min_{\Delta_i\in\mathbb{C}^{L\times N}} \|\mathbf{C}_{i+1}\odot(\widetilde{\boldsymbol{\Upsilon}}_i - \mathbf{X}_i\Delta_i^{\mathsf{H}} - \Delta_i\mathbf{X}_i^{\mathsf{H}} - \Delta_i\Delta_i^{\mathsf{H}})\|_{\mathsf{F}}^2 \tag{31}$$

with $\widetilde{\boldsymbol{\Upsilon}}_i \triangleq \boldsymbol{\Upsilon} - \mathbf{X}_i\mathbf{X}_i^{\mathsf{H}}$. As discussed, when I is sufficiently large, Δ_i represents a small perturbation, we can omit the second-order term $\Delta_i\Delta_i^{\mathsf{H}}$ and add the constraint[3] $\|\Delta_i\|_{\mathsf{F}} \le \varepsilon$, where $\varepsilon > 0$ is a pre-determined maximum perturbation, to obtain the approximated problem

$$\begin{aligned}\min_{\Delta_i\in\mathbb{C}^{L\times N}} \quad & \|\mathbf{C}_{i+1}\odot(\widetilde{\boldsymbol{\Upsilon}}_i - \mathbf{X}_i\Delta_i^{\mathsf{H}} - \Delta_i\mathbf{X}_i^{\mathsf{H}})\|_{\mathsf{F}}^2\\ \text{s.t.} \quad & \|\Delta_i\|_{\mathsf{F}} \le \varepsilon.\end{aligned} \tag{32}$$

Problem (32) is a convex quadratically constrained quadratic program (QCQP) w.r.t. $\mathrm{vec}(\Delta_i)$ that can be solved using standard optimization toolboxes [26], [39].

This approach is summarized in Algorithm 4.

A. Numerical Results

We consider that 50 active users are transmitting their randomly generated Gaussian pilot sequences in a time-frequency block of size 5 by 10. The identities of those users are assumed to be known, so that $\mathbf{C}(\boldsymbol{\gamma})$ is perfectly known. (This is the ideal case for applying the EM and the sequential approximation (SA) algorithms. When the user activities are not perfectly known, the estimation error of $\boldsymbol{\gamma}$ can further degrade the performance of channel covariance estimation. In contrast, using the all-one pilots does not require knowing the user activities.) The performance is evaluated in terms of the relative approximation error, defined as $\|\mathbf{R}-\widehat{\mathbf{R}}\|_{\mathsf{F}}/\|\mathbf{R}\|_{\mathsf{F}}$. We first check the convergence of the covariance estimation algorithms. As shown in Fig. 14, the SA approach achieves slightly better

[3] In [38], the perturbation is measured in terms of the maximum magnitude of the entries in Δ_i. However, we found that the problem can be solved more efficiently using the Frobenius norm constraint.

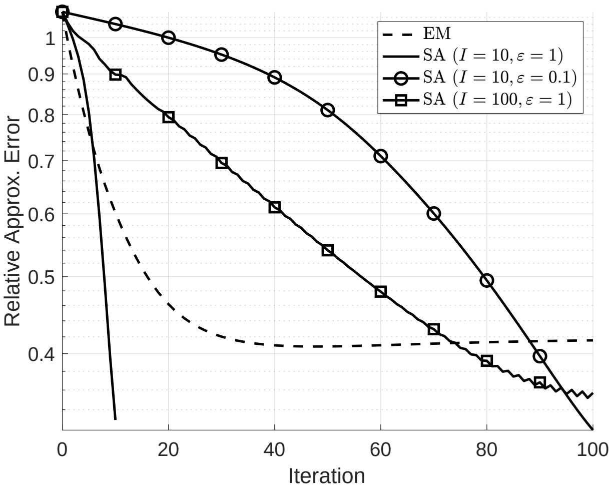


Fig. 13: Convergence of the covariance estimation algorithms.

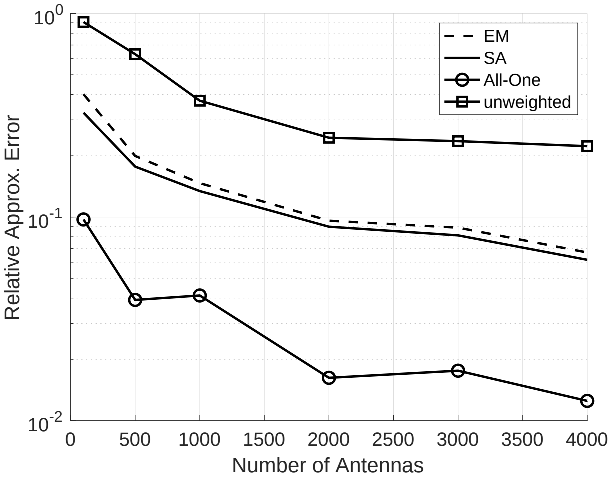


Fig. 14: Comparison of the covariance estimation methods.

performance than the EM algorithm. Additionally, the SA approach requires properly choosing the parameters I and ε, which represent the number of sub-problems to solve and the maximum perturbation in each sub-problem, respectively.

We then compare the performance of the EM algorithm and the SA approach with directly solving the unweighted problem (brute-forcely using $\mathbf{C} = \mathbf{1}\mathbf{1}^\mathsf{T}$ in (26)), and with the case of using all-one pilots. Suggested by the results in Fig. 13, we select $I = 10$ and $\varepsilon = 1$ for the SA approach, and run the EM algorithm for 40 iterations. As shown in Fig. 14, using the dedicated all-one pilots achieves much more accurate channel covariance estimation. When $\mathbf{C}(\boldsymbol{\gamma})$ is diagonally dominant, the WLRMA problem becomes very sensitive to the noise in the sample covariance matrix. In this case, the EM algorithm and the SA approach require an unrealistically large number of antennas to achieve a satisfactory performance.

Regarding the runtimes of different methods, the EM algorithm takes around 0.025 seconds, the SA approach takes around 1.5 seconds, and using the all-one pilots takes only around 0.002 seconds. (The QCQP in (32) is solved using the YAMIP toolbox [39] with the MOSEK solver [26].)

The idea of joint activity detection and covariance estimation outlined here bears some promise, but underperforms our main algorithm (Section V-C). It is a possible topic for future work to consider other approaches to the joint problem.

References

[1] J. Bai and E. G. Larsson, "Robust covariance-based activity detection for massive access," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Oct. 2023, pp. 304–308.

[2] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Sep. 2020.

[3] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. De Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[4] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.

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

[6] H. Djelouat, M. Leinonen, and M. Juntti, "Spatial correlation aware compressed sensing for user activity detection and channel estimation in massive MTC," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6402–6416, Feb. 2022.

[7] M. Leinonen, M. Codreanu, and G. B. Giannakis, "Compressed sensing with applications in wireless networks," *Foundations and Trends® in Signal Processing*, vol. 13, no. 1-2, pp. 1–282, 2019.

[8] Q. Wang, L. Liu, S. Zhang, and F. C. Lau, "Exploiting temporal side information in massive IoT connectivity," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1432–1447, Feb. 2023.

[9] Y. Jin and B. D. Rao, "Support recovery of sparse signals in the presence of multiple measurement vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3139–3157, May 2013.

[10] D. Jiang and Y. Cui, "ML and MAP device activity detections for grant-free massive access in multi-cell networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3893–3908, Jun. 2022.

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

[12] W. Jiang, M. Yue, X. Yuan, and Y. Zuo, "Massive connectivity over MIMO-OFDM: Joint activity detection and channel estimation with frequency selectivity compensation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6920–6934, Sep. 2022.

[13] A. L. Scharf, B. F. Uchôa-Filho, and D. Le Ruyet, "User activity detection and channel estimation in frequency-selective faded grant-free access," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2451–2455, Sep. 2023.

[14] W. Jiang, Y. Jia, and Y. Cui, "Statistical device activity detection for OFDM-based massive grant-free access," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3805–3820, Jun. 2022.

[15] Y. Zhu, G. Sun, W. Wang, L. You, F. Wei, L. Wang, and Y. Chen, "OFDM-based massive grant-free transmission over frequency-selective fading channels," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4543–4558, Jul. 2022.

[16] E. De Carvalho, E. Björnson, J. H. Sørensen, E. G. Larsson, and P. Popovski, "Random pilot and data access in massive MIMO for machine-type communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7703–7717, Dec. 2017.

[17] E. Becirovic, E. Björnson, and E. G. Larsson, "Detection of pilot-hopping sequences for grant-free random access in massive MIMO systems," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 8380–8384.

[18] V. Latora, V. Nicosia, and G. Russo, *Complex Networks: Principles, Methods and Applications*. Cambridge University Press, 2017.

[19] E. Björnson and Ö. T. Demir, *Introduction to Multiple Antenna Communications and Reconfigurable Surfaces*. Now Publishers, 2024.

[20] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, Jun. 1901. [Online]. Available: https://api.semanticscholar.org/CorpusID:125037489

[21] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 303–318, Jan. 2016.

[22] A. Kobel, F. Rouillier, and M. Sagraloff, "Computing real roots of real polynomials... and now for real!" in *Proc. ACM Int. Symp. Symbolic and Algebraic Comput. (ISSAC)*, Jul. 2016, pp. 303–310.

[23] J. M. McNamee, *Numerical Methods for Roots of Polynomials - Part I*. Elsevier, Jun. 2007.

[24] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection in multi-cell massive MIMO exploiting channel large-scale fading," *IEEE Trans. Signal Process.*, vol. 69, pp. 3768–3781, Jun. 2021.

[25] R. P. Brent, *Algorithms for Minimization without Derivatives*. Courier Corporation, 2013.

[26] M. ApS, *MOSEK optimization toolbox for MATLAB 10.2.0*, 2024. [Online]. Available: http://docs.mosek.com/latest/toolbox/index.html

[27] C. L. Lawson and R. J. Hanson, *Solving Least-Squares Problems*. Society for Industrial and Applied Mathematics, 1995. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611971217

[28] G. L. Stüber, *Principles of Mobile Communication, Fourth Edition*. Springer, 2017.

[29] C. Xiao, Y. R. Zheng, and N. C. Beaulieu, "Novel sum-of-sinusoids simulation models for Rayleigh and Rician fading channels," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3667–3679, Dec. 2006.

[30] 3GPP, "Universal mobile telecommunication systems (UMTS) deployment aspects," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 25.943, Apr. 2022, version 17.0.0.

[31] ——, "Study on channel model for frequency spectrum above 6 GHz," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.900, Jun. 2018, version 15.0.0.

[32] X. Lin, J. Li, R. Baldemair, J.-F. T. Cheng, S. Parkvall, D. C. Larsson, H. Koorapaty, M. Frenne, S. Falahati, A. Grovlen, and K. Werner, "5G New Radio: Unveiling the essentials of the next generation wireless access technology," *IEEE Commun. Standards Mag.*, vol. 3, no. 3, pp. 30–37, Sep. 2019.

[33] G. M. de Jesus, O. L. A. Lopez, R. D. Souza, N. H. Mahmood, M. Juntti, and M. Latva-Aho, "Assessment of the sparsity-diversity trade-offs in active users detection for mMTC," *arXiv preprint arXiv:2402.05687*, Feb. 2024.

[34] S. R. Khosravirad, O. Tirkkonen, Ü. Parts, L. Zhou, D. Korpi, P. Baracca, and M. A. Uusitalo, "Communications survival strategies for industrial wireless control," *IEEE Network*, vol. 36, no. 2, pp. 66–72, Mar. 2022.

[35] M. Henriksson, O. Gustafsson, U. K. Ganesan, and E. G. Larsson, "An architecture for grant-free random access massive machine type communication using coordinate descent," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Oct. 2020, pp. 1112–1116.

[36] Z. Wang, Z. Chen, Y.-F. Liu, F. Sohrab, and W. Yu, "An efficient active set algorithm for covariance based joint data and activity detection for massive random access with massive mimo," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 4840–4844.

[37] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. Int. Conf. Machine Learning (ICML)*, 2003, pp. 720–727.

[38] W.-S. Lu and A. Antoniou, "New method for weighted low-rank approximation of complex-valued matrices and its application for the design of 2-D digital filters," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, May 2003, pp. 694–697.

[39] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. IEEE Int. Conf. Robotics and Automation*, Taipei, Taiwan, Sep. 2004.