# Information-Theoretic Pilot Design for Downlink Channel Estimation in FDD Massive MIMO Systems

Yujie Gu, *Senior Member, IEEE* and Yimin D. Zhang, *Fellow, IEEE*

***Abstract*—Massive multiple-input multiple-output (MIMO) is one of the most promising techniques for next generation wireless communications due to its superior capability to provide high spectrum and energy efficiency. Considering the very large number of antennas employed at the base station, however, the pilot overhead for downlink channel estimation becomes unaffordable in frequency division duplex (FDD) multi-user massive MIMO systems. In this paper, we propose an information-theoretic metric to design the pilot for downlink channel estimation in FDD multi-user massive MIMO systems. By exploiting the low-rank nature of the channel covariance matrix, we first derive the minimum number of pilot symbols required to ensure perfect channel recovery, which is much less than the number of antennas at the base station. Further, under a general channel model that the channel vector of each user follows a Gaussian mixture distribution, the pilot symbols are designed by maximizing the weighted sum of the Shannon mutual information between the measurements of the users and their corresponding channel vectors on the complex Grassmannian manifold. Simulation results demonstrate the effectiveness of the proposed information-theoretic pilot design for the downlink channel estimation in FDD massive MIMO systems.**

***Index Terms*—Channel estimation, frequency division duplex (FDD), Gaussian mixture distribution, Grassmannian manifold, information-theoretic metric, massive multiple-input multiple-output (MIMO), pilot design.**

## I. Introduction

MASSIVE multiple-input multiple-output (MIMO) becomes a key enabling technology for next generation wireless communications benefited from its promising system capacity, energy efficiency, security and robustness [2–6]. In massive MIMO systems, the very large number of antennas at the base station simultaneously serve a much smaller number of users in the same radio frequency (RF) channel, where each user is usually equipped with a single antenna due to the physical size limitation. With the very large number of antennas, the number of degrees-of-freedom is high enough to eliminate multiuser interference using transmit precoding or receive beamforming. In addition, massive MIMO also improves the detection and estimation performance in wireless sensor networks [7–14].

Similar to general wireless communication systems, there are two commonly used duplex modes in massive MIMO systems, i.e., time division duplex (TDD) and frequency division duplex (FDD). Among them, the TDD is popular for massive MIMO [2, 15–18], where the principle of reciprocity can be leveraged, that is, the downlink channel vector (or matrix) is simply the transpose of the uplink channel vector (or matrix). As such, the number of required pilot symbols for the downlink channel estimation (via the uplink channel estimation) is independent of the number of antennas at the base station, which is very large in massive MIMO systems. It is worth noting that the accurate channel state information is essential in wireless communications for reliable signal transmission and efficient resource allocation.

Despite the attractiveness of the TDD mode in channel estimation, many current wireless cellular systems are primarily dominated by the FDD mode. Unlike the TDD mode, the channel reciprocity property does not apply for the FDD mode where the downlink and uplink channels occupy different frequency bands. In the FDD mode, each user estimates its own downlink channel from the received pilot symbols transmitted from the base station, and feeds the estimated downlink channel information back to the base station for subsequent signal transmission and resource allocation. Considering that the number of required pilot symbols for downlink channel estimation in the FDD mode is proportional to the number of antennas at the base station, there is a huge pilot overhead for massive MIMO systems equipped with a very large number of antennas. Hence, there is a more urgent requirement in FDD massive MIMO systems to use much less pilot symbols for downlink channel estimation.

Reducing the downlink pilot overhead in FDD massive MIMO systems has been the subject of recent studies due to its importance. These techniques are mainly developed by exploiting the sparsity of the channel on the virtual angular domain [19, 20], the low-rank nature of the channel covariance matrix [21–26], or joint sparse and low-rank structures [27]. Both the channel sparsity and the low-rank structure are due to the narrow angular spread of the incoming/outgoing rays at the base station in typical cellular systems, which subsequently leads to a high correlation between different paths that link the base station and the user. The sparsity and/or the low rankness enable the estimation of the downlink channel with much less pilot symbols than the number of antennas employed at the base station in a sparse reconstruction manner. Specifically, the distributed compressive channel estimation in [19] and the compressive sensing based adaptive channel estimation and feedback scheme in [20] exploit the sparsity

Part of this work was presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing [1].

The authors are with the Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA, 19122 USA (e-mail: guyujie@hotmail.com; ydzhang@temple.edu).

of massive MIMO channels to reduce the pilot overhead significantly. By reducing the dimensionality of the effective channels via the *correlated* channel covariance matrix, the joint spatial division and multiplexing scheme in [21] and the beam division multiple access transmission scheme in [24] enable significant savings in both the downlink pilot and the uplink feedback. By assuming that each training signal has a much lower rank than the number of antennas, open-loop and closed-loop training frameworks in [22] reduce the overhead of the downlink training phase by exploiting prior channel information such as the long-term channel statistics. In [23], the rank-deficient channel covariance matrices were exploited to design efficient downlink pilot symbols with dimensionality reduction. By exploiting the low-rank property of the massive MIMO channel matrix caused by correlation among users, the joint channel estimation for all users is performed at the base station based on the matrix completion [25], based on which the overhead of downlink channel training as well as uplink channel feedback can be reduced. By exploiting the low-rank property of the channel covariance matrices, the required number of pilot symbols was proved to be significantly reduced [26].

In addition, other researchers tried to utilize additional knowledge beyond the sparsity in order to further reduce the pilot overhead. For example, by exploiting the reciprocity of the angular scattering function [28] and the uplink-downlink covariance interpolation [29], the downlink channel estimation in FDD massive MIMO systems performs well even when the pilot dimension is less than the inherent dimension of the channel vectors. By estimating the directions-of-arrival (DOAs) of the users, the pilot symbols can then be designed such that the transmit energy concentrates over the known DOAs to achieve a beamforming gain [30–32]. Specifically, a two-stage compressive sensing scheme was proposed for channel estimation in [30], where the first stage randomly generates the pilot to coarsely estimate candidates for DOAs whereas the second stage adaptively designs the pilot to refine the candidates exhaustively. A joint RF training and compressive channel estimation scheme was proposed in [31], which achieves a better tradeoff between pilot overhead reduction using random RF training vectors and beamforming gain using narrow-beam RF training vectors. In [32], the out-of-band spatial information extracted from a sub-6 GHz channel is exploited to design a structured random codebook (pilot) and the associated weighted sparse channel recovery algorithm.

Typical pilot design criteria include maximizing the system spectral efficiency [21], minimizing the mean squared error (MSE) [22, 26], maximizing the average received signal-to-noise ratio (SNR) [22], maximizing the summation of the conditional mutual information [23], maximizing the sum-rate upper bound [24], minimizing the sum of MSEs [26], and minimizing the weighted sum MSE [33]. However, they all assume that the channel vector can be modeled as a single smooth Gaussian variable, which limits their applications in more complex environments. In [34], the uplink channel component in the beam domain is modeled to a more general Gaussian mixture, i.e., a weighted sum of multiple Gaussian distributions with different variances. To the best of our knowledge, however, there is no pilot design for the downlink channel estimation in massive MIMO systems based on the Gaussian mixture distribution.

In this paper, in addition to exploiting the sparsity and the low-rank nature, we further model the downlink channel in FDD massive MIMO systems to follow a general Gaussian mixture distribution for the pilot design [34, 35]. First, we study the asymptotic behavior of the minimum mean-squared error (MMSE) estimator. It reveals that a perfect channel recovery can be asymptotically reached in the high-SNR regime, provided that the number of pilot symbols is no less than the maximum rank of the channel covariance matrices of all Gaussian components of all users. Then, we adopt the information-theoretic approach to optimize the downlink pilot symbols for the estimation of Gaussian mixture channels in FDD massive MIMO systems. More specifically, the allocation of pilot symbols is optimized by maximizing the weighted sum of the Shannon mutual information between the measurements of multiple users served in massive MIMO systems and their corresponding downlink channel vectors. It is noted that the weighted sum of the mutual information is invariant when the pilot matrix undergoes an arbitrary unitary rotation, i.e., the weighted sum of the mutual information is a function of the subspace spanned by the pilot symbol vectors. Consequently, it enables us to optimize the pilot matrix on the complex Grassmannian manifold. Unlike the general sparse reconstruction problem, there is a closed-form MMSE solution for the downlink channel estimation under the Gaussian mixture assumption. Simulation results demonstrate the effectiveness of the proposed pilot design approach for the downlink channel estimation for an arbitrary number of users served in the massive MIMO system by exploiting the available *a priori* knowledge of the desired channel vector.

By generalizing the downlink channels in FDD massive MIMO systems to follow Gaussian mixture distributions, the main contributions of this paper can be summarized as follows:

- We analyze the asymptotic behavior of the MMSE estimator under Gaussian mixture distribution, and present the minimum number of pilot symbols required for Gaussian mixture channel estimation.
- We perform the Taylor series expansion to approximate the weighted sum of the mutual information associated with all users, and optimize the pilot matrix on a complex Grassmannian manifold.

The rest of the paper is organized as follows. In Section II, we describe the multi-user model in massive MIMO systems, and present the MMSE channel estimator based on the Gaussian mixture assumption. In Section III, we analyze the asymptotic behavior of the MMSE estimator in the high-SNR regime. In Section IV, we propose an information-theoretic pilot optimization on the Grassmannian manifold. In Section V, we compare the proposed pilot symbols with random pilot symbols in terms of channel estimation performance. We make our conclusions in Section VI.

## II. System Model

Assume that $N \gg 1$ transmit antennas equipped at the base station serve $M \geq 1$ single-antenna users in the FDD mode,

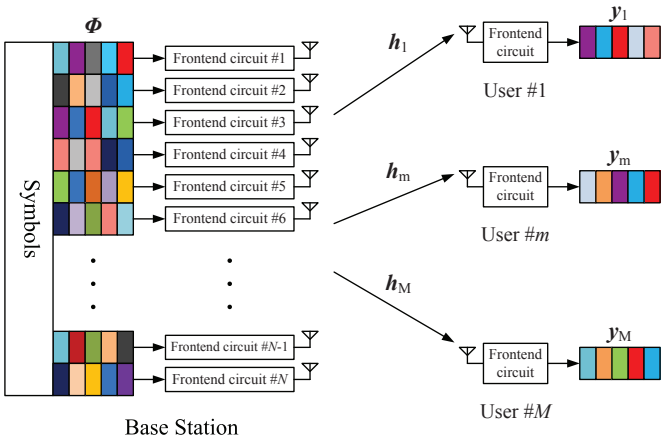


Fig. 1. Pilot demo in the FDD multi-user massive MIMO system.

as shown in Fig. 1. The downlink channel vector of the $m$-th user, $\boldsymbol{h}_m \in \mathbb{C}^N$, is estimated at the $m$-th user terminal from the noisy measurements of the known pilot symbols transmitted from the base station. The estimated downlink channel information is then sent back to the base station for subsequent scheduling, precoding and transmission.

Without loss of generality, the downlink channel of each user is assumed to be a frequency-flat fading channel under a narrowband assumption. This assumption can be easily extended to wideband frequency-selective channels in the context of multi-carrier transmission schemes such as orthogonal frequency-division multiplexing (OFDM).

Assume that the base station transmits a set of pilot symbols $\{\boldsymbol{\phi}(l), l = 1, 2, \cdots, L\}$, where $L$ is the length of the pilot (i.e., the number of pilot symbols in time) transmitted from each antenna. The $l$-th baseband signal received at the $m$-th user terminal is expressed as

$$y_m(l) = \boldsymbol{\phi}^{\mathrm{T}}(l)\boldsymbol{h}_m + n_m(l), \quad \forall\, m \in \mathcal{M}, \tag{1}$$

where $\mathcal{M} = \{1, 2, \cdots, M\}$ is the set denoting the indexes of users served in the massive MIMO system, $\boldsymbol{\phi}(l) \in \mathbb{C}^N$ is the $l$-th pilot symbol vector transmitted from $N$ base station antennas, and $n_m(l) \sim \mathcal{CN}(0, \sigma_{n_m}^2)$ denotes the zero-mean additive white Gaussian noise with variance $\sigma_{n_m}^2$. Here, $(\cdot)^{\mathrm{T}}$ denotes the transpose operator. Note that, in this signal model, the inter-cell interference leaked from neighboring frequency-reuse cells is ignored for the simplicity and clarity of the problem.

Stacking the received signals of the $m$-th user terminal over all $L$ pilot symbols as $\boldsymbol{y}_m = [y_m(1), y_m(2), \cdots, y_m(L)]^{\mathrm{T}} \in \mathbb{C}^L$ gives

$$\boldsymbol{y}_m = \boldsymbol{\Phi}\boldsymbol{h}_m + \boldsymbol{n}_m, \tag{2}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi}(1), \boldsymbol{\phi}(2), \cdots, \boldsymbol{\phi}(L)]^{\mathrm{T}} \in \mathbb{C}^{L\times N}$ is the pilot symbol matrix, and $\boldsymbol{n}_m = [n_m(1), n_m(2), \cdots, n_m(L)]^{\mathrm{T}} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_{n_m}^2 \boldsymbol{I})$ is the zero-mean additive Gaussian noise vector with $\boldsymbol{I}$ denoting the identity matrix with an appropriate dimension. The least squares (LS) estimate of $\boldsymbol{h}_m$ is given by

$$\hat{\boldsymbol{h}}_m^{\mathrm{LS}} = \left[\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Phi}\right]^{-1}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{y}_m, \tag{3}$$

where $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose operator. In order to perform the matrix inversion in the above expression, the number of pilot symbols must not be less than the number of antennas, i.e., $L \geq N$. Unfortunately, in massive MIMO systems, $N$, the number of antennas at the base station, is typically very large, which means that such a pilot overhead requirement becomes unaffordable in the FDD mode.

In order to guarantee the transmission efficiency, the number of pilot symbols in FDD massive MIMO systems should be kept much less than the number of antennas at the base station, i.e., $L \ll N$. In such a case, the above downlink channel estimation in (3) becomes ill-conditioned because the number of unknowns is much larger than the number of measurements. As a result, the least squares method is no longer applicable.

On the other hand, because of the compact configuration of the antennas in massive MIMO systems, the channel is a correlated random vector depending on the scattering geometry, which leads to a low-rank channel covariance matrix [23, 26]. In addition, the channel can be regarded as sparse or approximately sparse in some suitable representation bases, where the channel only contains a small number of dominant components while the others are negligible [34]. When the channel vector is sparse or approximately sparse, it can be estimated in the framework of compressive sensing with less measurements than the number of unknowns [20].

In addition to the sparsity and low-rank property, in this paper, we further model the channel vector in massive MIMO systems as a Gaussian mixture distribution. This mixture distribution is very popular and well verified in practice to describe the real environment signals with high flexibility and tractability (see, for example, [36, 37] and the references therein). Let the probability density function (pdf) of the channel vector $\boldsymbol{h}_m$ be modeled by a mixture of Gaussian distributions given by

$$\begin{aligned} f(\boldsymbol{h}_m) &= \sum_{k\in\mathcal{K}_m} p_k^{(m)} f^{(k)}(\boldsymbol{h}_m) \\ &= \sum_{k\in\mathcal{K}_m} p_k^{(m)} \mathcal{CN}\big(\boldsymbol{u}_{\boldsymbol{h}_m}^{(k)}, \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big), \end{aligned} \tag{4}$$

where $\mathcal{K}_m = \{1, 2, \cdots, K_m\}$ has a cardinality $K_m$, and $\sum_{k\in\mathcal{K}_m} p_k^{(m)} = 1$. The Gaussian mixture distribution of the channel vector $\boldsymbol{h}_m$ implies that it contains $K_m$ Gaussian components, and the $k$-th Gaussian component is activated with probability $p_k^{(m)} > 0$ and, when activated, that component generates a complex-valued Gaussian vector with distribution $f^{(k)}(\boldsymbol{h}_m) = \mathcal{CN}\big(\boldsymbol{u}_{\boldsymbol{h}_m}^{(k)}, \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)$. The parameter set $\left\{p_k^{(m)}, \boldsymbol{u}_{\boldsymbol{h}_m}^{(k)}, \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}; k \in \mathcal{K}_m\right\}$ defines the Gaussian mixture distribution of the channel vector $\boldsymbol{h}_m$. The parameters characterizing the Gaussian mixture distribution can be estimated by using, e.g., the expectation maximization (EM) algorithm [38, 39], the sparse Bayesian learning method [40, 41], or the piecewise-Gaussian approximation method [1, 36].

Under the channel distribution model (4), it can be shown that the measurement $\boldsymbol{y}_m$ in (2) also follows the Gaussian mixture distribution with $K_m$ components, and its pdf is

expressed as

$$f(\boldsymbol{y}_m) = \sum_{k\in\mathcal{K}_m} p_k^{(m)} f^{(k)}(\boldsymbol{y}_m), \tag{5}$$

where the $k$-th component $f^{(k)}(\boldsymbol{y}_m) = \mathcal{CN}\big(\boldsymbol{u}_{\boldsymbol{y}_m}^{(k)}, \boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\big)$ is complex Gaussian distributed with mean vector and covariance matrix given by

$$\begin{aligned} \boldsymbol{u}_{\boldsymbol{y}_m}^{(k)} &= \boldsymbol{\Phi}\boldsymbol{u}_{\boldsymbol{h}_m}^{(k)}, \\ \boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)} &= \boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{\Phi}^{\mathrm{H}} + \sigma_{n_m}^2\boldsymbol{I}. \end{aligned} \tag{6}$$

The MMSE estimate of the channel vector $\boldsymbol{h}_m$, defined by

$$\min_{\hat{\boldsymbol{h}}_m} \ \mathbb{E}\left\{\left\|\boldsymbol{h}_m - \hat{\boldsymbol{h}}_m\right\|_2^2\right\}, \tag{7}$$

is given by [42]

$$\hat{\boldsymbol{h}}_m^{\text{MMSE}} = \mathbb{E}\{\boldsymbol{h}_m|\boldsymbol{y}_m\} = \sum_{k\in\mathcal{K}_m} p_{k|\boldsymbol{y}_m}\boldsymbol{u}_{\boldsymbol{h}_m|\boldsymbol{y}_m}^{(k)}, \tag{8}$$

where $\mathbb{E}\{\,\cdot\,\}$ denotes the statistical expectation operator,

$$\boldsymbol{u}_{\boldsymbol{h}_m|\boldsymbol{y}_m}^{(k)} = \boldsymbol{u}_{\boldsymbol{h}_m}^{(k)} + \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{\Phi}^{\mathrm{H}}\big[\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\big]^{-1}\big(\boldsymbol{y}_m - \boldsymbol{u}_{\boldsymbol{y}_m}^{(k)}\big) \tag{9}$$

is the $k$-th component of the MMSE estimator of $\boldsymbol{h}_m$ given the measurement $\boldsymbol{y}_m$, and

$$p_{k|\boldsymbol{y}_m} = \frac{p_k^{(m)} f^{(k)}(\boldsymbol{y}_m)}{f(\boldsymbol{y}_m)} \tag{10}$$

is the corresponding posterior probability [43]. When the downlink channel estimate $\hat{\boldsymbol{h}}_m^{\text{MMSE}}$ is obtained, it will be sent back to the base station.

Note that, although the MMSE estimator for $\boldsymbol{h}_m$ has an analytical form, its performance measure, the MMSE itself, does not have such a closed form because of the Gaussian mixture distribution. According to [42], the MSE of $\boldsymbol{h}_m$ is upper and lower bounded as (11), shown at the bottom of the page, where $\mathrm{Tr}[\,\cdot\,]$ denotes the trace of a matrix, and

$$\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m} = \sum_{k\in\mathcal{K}_m} p_k^{(m)}\Big[\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)} + \boldsymbol{u}_{\boldsymbol{h}_m}^{(k)}\big(\boldsymbol{u}_{\boldsymbol{h}_m}^{(k)}\big)^{\mathrm{H}}\Big] - \boldsymbol{u}_{\boldsymbol{h}}^{(m)}\big(\boldsymbol{u}_{\boldsymbol{h}}^{(m)}\big)^{\mathrm{H}} \tag{12}$$

is the covariance matrix of the channel vector $\boldsymbol{h}_m$ with the mean vector given by

$$\boldsymbol{u}_{\boldsymbol{h}}^{(m)} = \sum_{k\in\mathcal{K}_m} p_k^{(m)}\boldsymbol{u}_{\boldsymbol{h}_m}^{(k)}. \tag{13}$$

It is indicated in [42] that the upper and lower bounds approach each other as the SNR increases, and they must coincide as the SNR tends to infinity.

When the Gaussian mixture distribution degrades into a single Gaussian distribution (i.e., $K_m = 1$), the upper and lower bounds of the MSE of the MMSE estimator for the channel vector $\boldsymbol{h}_m$ in (11) become identical, i.e., the MSE has a closed-form expression, thereby facilitating the pilot design via minimizing the MSE [26] or minimizing the weighted sum MSE [33]. However, the existing pilot design methods under the MMSE criterion for the Gaussian channel are not directly suitable for the Gaussian mixture channel which does not have an analytic MSE expression. Considering the relationship between mutual information and MMSE [44], we propose an information-theoretic pilot design approach for the Gaussian mixture channel estimation in FDD massive MIMO systems. In the sequel, we first prove the asymptotic behavior of the MMSE estimator for the Gaussian mixture channel in Section III, and then propose the information-theoretic pilot optimization on the Grassmannian manifold in Section IV.

## III. Asymptotic Behavior of the MMSE Estimator

In this section, we analyze the behavior of the MMSE estimator in the asymptotic high-SNR regime. The asymptotic analysis verifies the possibility of perfect channel recovery from a small number of pilot symbols, because of the low rank of the channel covariance matrix in massive MIMO systems.

*Theorem*: Let $r_m^{(k)}$ be the rank of the channel covariance matrix of the $k$-th Gaussian component in the Gaussian mixture distribution of the $m$-th user, $\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}$, i.e., $r_m^{(k)} = \mathrm{rank}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)$, where $\mathrm{rank}(\,\cdot\,)$ denotes the rank of a matrix. Let $\boldsymbol{V}_m^{(k)}\boldsymbol{\Gamma}_m^{(k)}\big(\boldsymbol{V}_m^{(k)}\big)^{\mathrm{H}}$ denote the eigen-decomposition of $\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Phi}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}$, where $\boldsymbol{V}_m^{(k)} = \big[\boldsymbol{v}_{m_1}^{(k)}, \cdots, \boldsymbol{v}_{m_N}^{(k)}\big]$ is a unitary matrix consisting of the eigenvectors of $\boldsymbol{V}_m^{(k)}\boldsymbol{\Gamma}_m^{(k)}\big(\boldsymbol{V}_m^{(k)}\big)^{\mathrm{H}}$, and the diagonal matrix $\boldsymbol{\Gamma}_m^{(k)} = \mathrm{diag}\big(\gamma_1, \gamma_2, \cdots, \gamma_{r_m^{(k)}}, 0, \cdots, 0\big)$ consists of the corresponding eigenvalues with $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{r_m^{(k)}} > 0$. Assume that the number of randomly generated pilot symbols, $L$, is no less than the maximum of $r_m^{(k)}$ of all Gaussian components for each user, i.e., $L \geq \max_{m\in\mathcal{M}}\max_{k\in\mathcal{K}_m} r_m^{(k)}$, then the *lower* bound of the MSE of the MMSE estimate of $\boldsymbol{h}_m$ is given by

$$\begin{aligned} \varepsilon_{\text{Lower}}^{(m)} &\leq \mathbb{E}\left\{\left\|\boldsymbol{h}_m - \hat{\boldsymbol{h}}_m^{\text{MMSE}}\right\|_2^2\right\} \\ &= \sum_{k\in\mathcal{K}_m} p_k^{(m)} \sum_{i=1}^{r_m^{(k)}} \left(1 + \frac{\gamma_i}{\sigma_{n_m}^2}\right)^{-1} \big(\boldsymbol{v}_{m_i}^{(k)}\big)^{\mathrm{H}} \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)} \boldsymbol{v}_{m_i}^{(k)}, \\ &\qquad\qquad \forall\ m \in \mathcal{M}, \end{aligned} \tag{14}$$

which approaches zero in the asymptotic low-noise regime, i.e., $\sigma_{n_m}^2 \to 0$. Considering that the upper and lower bounds

$$\begin{aligned} &\sum_{k\in\mathcal{K}_m} p_k^{(m)} \mathrm{Tr}\Big[\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)} - \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{\Phi}^{\mathrm{H}}\big(\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{\Phi}^{\mathrm{H}} + \sigma_{n_m}^2\boldsymbol{I}\big)^{-1}\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\Big] \\ &\leq \mathbb{E}\left\{\left\|\boldsymbol{h}_m - \hat{\boldsymbol{h}}_m^{\text{MMSE}}\right\|_2^2\right\} \\ &\leq \mathrm{Tr}\Big[\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m} - \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}\boldsymbol{\Phi}^{\mathrm{H}}\big(\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}\boldsymbol{\Phi}^{\mathrm{H}} + \sigma_{n_m}^2\boldsymbol{I}\big)^{-1}\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}\Big], \end{aligned} \tag{11}$$

approach each other with increasing SNR and they must coincide as the SNR tends to infinity [42], we have

$$\lim_{\mathrm{SNR}_m \to \infty} \mathbb{E}\left\{ \left\| \boldsymbol{h}_m - \hat{\boldsymbol{h}}_m^{\mathrm{MMSE}} \right\|_2^2 \right\} = 0, \quad \forall\, m \in \mathcal{M}, \quad (15)$$

where $\mathrm{SNR}_m = \|\boldsymbol{h}_m\|^2/\sigma_{n_m}^2$ denotes the SNR of the $m$-th user's channel. That is, perfect channel recovery is possible for each user.

*Proof*: See Appendix A. ■

The above asymptotic analysis proves that an accurate channel estimation from as few as $L = \max_{m\in\mathcal{M}} \max_{k\in\mathcal{K}_m} \mathrm{rank}\big(\boldsymbol{R}^{(k)}_{\boldsymbol{h}_m\boldsymbol{h}_m}\big)$ pilot symbols is possible in the asymptotic regime. Nevertheless, it does not tell us how to design pilot symbols in the non-asymptotic regime. In the following section, we propose a novel information-theoretic pilot design method for the channel estimation in FDD massive MIMO systems serving for an arbitrary number of users.

## IV. Information-Theoretic Pilot Optimization on the Grassmannian Manifold

In this section, we adopt the maximum mutual information criterion to optimize the pilot symbols on the Grassmannian manifold for the estimation of Gaussian mixture channels. Define $I(\boldsymbol{y}_m;\boldsymbol{h}_m)$ as the Shannon mutual information between the measurement vector $\boldsymbol{y}_m$ at the $m$-th user and the corresponding channel vector $\boldsymbol{h}_m$, i.e., [45]

$$I(\boldsymbol{y}_m;\boldsymbol{h}_m) = h(\boldsymbol{y}_m) - h(\boldsymbol{y}_m|\boldsymbol{h}_m), \quad (16)$$

where $h(\boldsymbol{y}_m) = -\mathbb{E}\{\log[f(\boldsymbol{y}_m)]\}$ denotes the differential entropy of the measurement vector $\boldsymbol{y}_m$, and $h(\boldsymbol{y}_m|\boldsymbol{h}_m) = -\mathbb{E}\{\log[f(\boldsymbol{y}_m|\boldsymbol{h}_m)]\}$ denotes the differential entropy of the measurement vector $\boldsymbol{y}_m$ conditioned on the channel vector $\boldsymbol{h}_m$. Note that, the mutual information $I(\boldsymbol{y}_m;\boldsymbol{h}_m)$ is an implicit function of the pilot matrix $\boldsymbol{\Phi}$ via the measurement equation (2). It is difficult, if not impossible, to analytically derive the differential entropy even for a simple estimation problem, let alone the parameter estimation problem with the high dimensionality and non-Gaussianity.

In massive MIMO systems, the pilot symbols transmitted from the base station are common to all served users and should be optimized to maximize the mutual information associated with all these users. It is assumed that the channels of different users are independent to each other because of the separation between the users and the rich multipath. Let

$$J(\boldsymbol{\Phi}) = \sum_{m\in\mathcal{M}} w_m I(\boldsymbol{y}_m;\boldsymbol{h}_m) \quad (17)$$

denote the weighted sum of the mutual information associated with all users served in the massive MIMO system, where $w_m$ is the weight assigned to the $m$-th user with $\sum_{m\in\mathcal{M}} w_m = 1$ according to different criteria, such as the commonly used equal weights and different fairness weights (e.g., max-min fairness, weighted fairness, and proportional fairness) [46]. Unlike in (16), here we express the weighted sum of the mutual information as an explicit function of $\boldsymbol{\Phi}$.

The information-theoretic design of pilot symbols can be formulated as an optimization problem to maximize the weighted sum of the mutual information associated with all users, i.e.,

$$\begin{aligned} \max_{\boldsymbol{\Phi}} \quad & J(\boldsymbol{\Phi}) \\ \text{subject to} \quad & \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}} = \boldsymbol{I}, \end{aligned} \quad (18)$$

where the orthonormal constraint $\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}} = \boldsymbol{I}$ is introduced to allocate the equal power to each pilot symbol so as to avoid increasing the mutual information by simply scaling $\boldsymbol{\Phi}$ to be larger because scaling $\boldsymbol{\Phi}$ only affects the channel rather than the noise. Another commonly adopted pilot constraint is the total transmit power constraint [23, 26, 33], i.e., $\mathrm{Tr}[\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}}] \le L$. It is noted that the orthonormal constraint always satisfies the power constraint, but not vice versa. That is to say, the feasible set with orthonormal constraint is a subset of the feasible set with power constraint, i.e, $\{\boldsymbol{\Phi}|\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}} = \boldsymbol{I}\} \subset \{\boldsymbol{\Phi}|\mathrm{Tr}\big[\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}}\big] \le L\}$.

Note that the objective function in (18) is non-convex with respect to the optimization variable $\boldsymbol{\Phi}$, which belongs to the $L$-dimensional subspace in $\mathbb{C}^N$. In general, the entropy calculation does not have a closed-form solution for most pdf's of interest, including the Gaussian mixture model used in this paper. In order to obtain a feasible solution, therefore, we perform a Taylor series expansion of the logarithm of the Gaussian mixture pdf required in the definition of the entropy, which enables a gradient-based search method.

By performing the first-order Taylor series expansion of the logarithm of the Gaussian mixture distribution in the definition of the differential entropy of the measurement vector $\boldsymbol{y}_m$, we obtain the approximated differential entropy as [1, 36]

$$h(\boldsymbol{y}_m) \approx -\log\left[ \sum_{k\in\mathcal{K}_m} p_k^{(m)} f^{(k)}\big(\boldsymbol{y}_0^{(m)}\big) \right], \quad (19)$$

where $\boldsymbol{y}_0^{(m)} = \mathbb{E}\{\boldsymbol{y}_m\}$ is the mean value of the measurement $\boldsymbol{y}_m$. The derivation of the above approximated differential entropy can be found in Appendix B. Following the zero mean assumption of the channel vector [21, 23, 26, 33, 34, 47], we have $\boldsymbol{u}^{(k)}_{\boldsymbol{y}_m} = \boldsymbol{\Phi}\boldsymbol{u}^{(k)}_{\boldsymbol{h}_m} = \boldsymbol{0}$ for all individual Gaussian components of $\boldsymbol{y}_m$. In this case, it is natural to set the Taylor series expansion point to $\boldsymbol{y}_0^{(m)} = \boldsymbol{0}$, resulting in

$$f^{(k)}\big(\boldsymbol{y}_0^{(m)}\big) = \frac{1}{\pi^L \left|\boldsymbol{R}^{(k)}_{\boldsymbol{y}_m\boldsymbol{y}_m}\right|}, \quad (20)$$

where $|\cdot|$ denotes the determinant of a matrix. We now have the approximated differential entropy of $\boldsymbol{y}_m$ expressed as

$$h(\boldsymbol{y}_m) \approx -\log\left[ \sum_{k\in\mathcal{K}_m} p_k^{(m)} \left|\boldsymbol{R}^{(k)}_{\boldsymbol{y}_m\boldsymbol{y}_m}\right|^{-1} \right] + L\log\pi. \quad (21)$$

Because the additive white Gaussian noise vector $\boldsymbol{n}_m \sim \mathcal{CN}(\boldsymbol{0}, \sigma_{n_m}^2\boldsymbol{I})$ is independent of the channel vector $\boldsymbol{h}_m$, the conditional differential entropy is given by

$$h(\boldsymbol{y}_m|\boldsymbol{h}_m) = h(\boldsymbol{n}_m) = L\log(\pi e\sigma_{n_m}^2). \quad (22)$$

Hence, the mutual information $I(\boldsymbol{y}_m;\boldsymbol{h}_m)$ is approximated as

$$I(\boldsymbol{y}_m;\boldsymbol{h}_m) \approx -\log\left[ \sum_{k\in\mathcal{K}_m} p_k^{(m)} \left|\boldsymbol{R}^{(k)}_{\boldsymbol{y}_m\boldsymbol{y}_m}\right|^{-1} \right] - L\log(e\sigma_{n_m}^2). \quad (23)$$

Accordingly, the weighted sum of the mutual information associated with all served users is approximated as

$$J(\boldsymbol{\Phi}) \approx -\sum_{m\in\mathcal{M}} w_m \log\left[\sum_{k\in\mathcal{K}_m} p_k^{(m)} \left|\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\right|^{-1}\right] - L\log(e\sigma_{n_m}^2), \tag{24}$$

where the second term is a constant independent of the pilot matrix $\boldsymbol{\Phi}$. It is noted that $J(\boldsymbol{\Phi})$, the objective function in (18), remains *unchanged* when the pilot matrix $\boldsymbol{\Phi}$ undergoes an arbitrary unitary rotation, i.e.,

$$J(\boldsymbol{Q}\boldsymbol{\Phi}) = J(\boldsymbol{\Phi}), \tag{25}$$

where $\boldsymbol{Q} \in \mathbb{C}^{L\times L}$ is a unitary matrix satisfying $\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{H}} = \boldsymbol{I}$. Hence, $J(\boldsymbol{\Phi})$ is a function of the subspace spanned by the rows of the pilot matrix, or equivalently, the row space of $\boldsymbol{\Phi}$, because the row spaces of $\boldsymbol{\Phi}$ and $\boldsymbol{Q}\boldsymbol{\Phi}$ are the same.

Based on (25), we can optimize the mutual information maximization problem (18) over the complex Grassmannian manifold $\mathcal{G}(N, L)$, which composes of all $L$-dimensional subspaces in $\mathbb{C}^N$. The Grassmannian is a compact Riemannian manifold, and its geodesics can be explicitly computed [48, 49]. Following the framework in [50], we derive the Grassmannian gradient ascent algorithm to search the pilot matrix $\boldsymbol{\Phi}$ on the Grassmannian manifold $\mathcal{G}(N, L)$ such that the weighted sum of the mutual information $J(\boldsymbol{\Phi})$ is maximized.

To compute the gradient of $J(\boldsymbol{\Phi})$ on the Grassmannian manifold, we first need to compute the derivative of $J(\boldsymbol{\Phi})$ with respect to $\boldsymbol{\Phi}$. Taking the derivative of the weighted sum of the approximated mutual information in (24) with respect to the pilot matrix $\boldsymbol{\Phi}$, we have $\frac{dJ(\boldsymbol{\Phi})}{d\boldsymbol{\Phi}}$ in (26), shown at the bottom of the page, where the first denominator is a positive real number that scales the assigned weight of the mutual information of a specific user, which, in turn, affects the derivative direction, and the second denominator affects the derivative direction by scaling the activated probabilities of the Gaussian components in the mixture.

For the special case that there is only a *single* user in the massive MIMO system, by ignoring the index $m$ in (4), the Gaussian mixture distribution of the channel vector $\boldsymbol{h}$ is given by

$$f(\boldsymbol{h}) = \sum_{k\in\mathcal{K}} p_k f^{(k)}(\boldsymbol{h}) = \sum_{k\in\mathcal{K}} p_k \mathcal{CN}\big(\boldsymbol{u}_{\boldsymbol{h}}^{(k)}, \boldsymbol{R}_{\boldsymbol{h}\boldsymbol{h}}^{(k)}\big), \tag{27}$$

where $K$ denotes the cardinality of the index set $\mathcal{K} = \{1, 2, \cdots, K\}$. Accordingly, the derivative of the weighted sum of the approximated mutual information degenerates into the approximated mutual information gradient as [1]

$$\frac{d}{d\boldsymbol{\Phi}} I(\boldsymbol{y};\boldsymbol{h}) \approx \frac{\sum\limits_{k\in\mathcal{K}} p_k \left|\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}^{(k)}\right|^{-1} \left[\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}^{(k)}\right]^{-1} \boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}\boldsymbol{h}}^{(k)}}{\sum\limits_{k\in\mathcal{K}} p_k \left|\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}^{(k)}\right|^{-1}}, \tag{28}$$

where the denominator does not affect the derivative direction but the convergence rate.

According to [48], for the function $J(\boldsymbol{\Phi})$ defined on the Grassmannian manifold, the gradient of $J(\boldsymbol{\Phi})$ at $\boldsymbol{\Phi}$ is defined to be the tangent vector $\nabla_{\boldsymbol{\Phi}} J(\boldsymbol{\Phi})$ such that

$$\mathrm{Tr}\left[\left(\frac{dJ(\boldsymbol{\Phi})}{d\boldsymbol{\Phi}}\right)^{\mathrm{H}} \Delta\right] \equiv \mathrm{Tr}\left[(\nabla_{\boldsymbol{\Phi}} J(\boldsymbol{\Phi}))^{\mathrm{H}} \Delta\right] \tag{29}$$

for all tangent vectors $\Delta$ at $\boldsymbol{\Phi}$. Solving (29) for $\nabla_{\boldsymbol{\Phi}} J(\boldsymbol{\Phi})$ such that $(\nabla_{\boldsymbol{\Phi}} J(\boldsymbol{\Phi}))\,\boldsymbol{\Phi}^{\mathrm{H}} = \boldsymbol{0}$ yields the Grassmannian gradient as

$$\nabla_{\boldsymbol{\Phi}} J(\boldsymbol{\Phi}) = \frac{dJ(\boldsymbol{\Phi})}{d\boldsymbol{\Phi}}\big(\boldsymbol{I} - \boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Phi}\big), \tag{30}$$

which is the steepest ascent direction on the manifold. Using the steepest ascent method on the Grassmannian manifold $\mathcal{G}(N, L)$, the pilot matrix is updated according to

$$\hat{\boldsymbol{\Phi}} = \boldsymbol{\Phi} + \gamma \nabla_{\boldsymbol{\Phi}} J(\boldsymbol{\Phi}), \tag{31}$$

where $\gamma > 0$ is a small step size. The updated pilot matrix is a linear combination of the current pilot matrix and the Grassmannian gradient of the weighted sum of the approximated mutual information with respect to the pilot matrix, which generally lies outside the manifold $\mathcal{G}(N, L)$.

To project the updated pilot $\hat{\boldsymbol{\Phi}}$ back onto $\mathcal{G}(N, L)$, the orthonormal constraint in (18) is enforced by seeking the closet row-orthonormal matrix to the updated pilot matrix $\hat{\boldsymbol{\Phi}}$, which is a orthogonal Procrustes problem [51]. The orthonormal pilot matrix closest to $\hat{\boldsymbol{\Phi}}$ is given by

$$\tilde{\boldsymbol{\Phi}} = \boldsymbol{D}\boldsymbol{I}_{L\times N}\boldsymbol{G}^{\mathrm{H}}, \tag{32}$$

where $\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{G}^{\mathrm{H}} = \hat{\boldsymbol{\Phi}}$ is the singular value decomposition (SVD) of $\hat{\boldsymbol{\Phi}}$ with $\boldsymbol{D} \in \mathbb{C}^{L\times L}$ and $\boldsymbol{G} \in \mathbb{C}^{N\times N}$ being the unitary matrices, $\boldsymbol{I}_{L\times N} = \big[\boldsymbol{I}_{L\times L}, \boldsymbol{0}_{L\times(N-L)}\big]$ is a rectangular diagonal matrix composed by an identity matrix and a zero matrix. Namely, the orthonormal pilot matrix $\tilde{\boldsymbol{\Phi}}$ has the same unitary matrices as the updated pilot matrix $\hat{\boldsymbol{\Phi}}$, but with all one singular values. Then, $\tilde{\boldsymbol{\Phi}}$ is used to substitute $\boldsymbol{\Phi}$ in the next iteration in calculating (26) and then updated using (31)

$$\begin{aligned}
\frac{dJ(\boldsymbol{\Phi})}{d\boldsymbol{\Phi}} &\approx -\sum_{m\in\mathcal{M}} w_m \frac{d}{d\boldsymbol{\Phi}}\left\{\log\left[\sum_{k\in\mathcal{K}_m} p_k^{(m)} \left|\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\right|^{-1}\right]\right\} \\
&= -\sum_{m\in\mathcal{M}} \frac{w_m}{\sum\limits_{k\in\mathcal{K}_m} p_k^{(m)} \left|\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\right|^{-1}} \sum_{k\in\mathcal{K}_m} p_k^{(m)} \frac{d}{d\boldsymbol{\Phi}}\left\{\left|\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\right|^{-1}\right\} \\
&= \sum_{m\in\mathcal{M}} \frac{w_m}{\sum\limits_{k\in\mathcal{K}_m} p_k^{(m)} \left|\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\right|^{-1}} \sum_{k\in\mathcal{K}_m} \frac{p_k^{(m)}}{\left|\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\right|} \left[\boldsymbol{R}_{\boldsymbol{y}_m\boldsymbol{y}_m}^{(k)}\right]^{-1} \boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)},
\end{aligned} \tag{26}$$

**Algorithm 1** : Information-theoretic pilot optimization on the Grassmannian

**Initialize:** Random pilot matrix $\boldsymbol{\Phi}$;

**Repeat**

Step 1: Compute the derivative of the weighted sum of the approximated mutual information $\frac{dJ(\boldsymbol{\Phi})}{d\boldsymbol{\Phi}}$ via (26);

Step 2: Compute the Grassmannian gradient $\nabla_{\boldsymbol{\Phi}} J(\boldsymbol{\Phi})$ via (30);

Step 3: Update the pilot matrix $\hat{\boldsymbol{\Phi}}$ in the steepest ascent direction on $\mathcal{G}(N, L)$ via (31);

Step 4: Enforce the orthonormal constraint to obtain $\tilde{\boldsymbol{\Phi}}$ via (32), and let $\boldsymbol{\Phi} \leftarrow \tilde{\boldsymbol{\Phi}}$. Go back to Step 1.

**Until** convergence

**Output:** Information-theoretic pilot matrix $\boldsymbol{\Phi}$.

to achieve an iterative search procedure. The convergence criterion of the iterative process requires that the gradient of the weighted sum mutual information with respect to the pilot matrix converges to zero. Algorithm 1 summarizes the proposed information-theoretic pilot design for the downlink channel estimation in FDD massive MIMO systems.

From (26), the computational complexity of calculating the derivative of the weighted sum of the approximated mutual information with respect to the pilot matrix is $\mathcal{O}\left(LN^2 \sum_{m\in\mathcal{M}} K_m\right)$ because $L \ll N$. Hence, the overall computational complexity of the proposed pilot optimization method is $\mathcal{O}\left(TLN^2 \sum_{m\in\mathcal{M}} K_m\right)$, where $T$ denotes the number of iteration.

For effective channel estimation at the user terminal, the base station is required to send the optimized pilot symbols to all users served in the massive MIMO system. The signaling overhead of the optimized pilot matrix is generally very high. To reduce the actual overhead required in practical system implementations, one can predesign a set of sequences which are shared by the base station and the users [22, 33]. As such, the base station only needs to transmit the indexes of the chosen sequences, thus greatly relaxing the required signaling overhead.

## V. Simulation Results

In this section, we carry out simulations to demonstrate the performance advantages of the proposed pilot design method for downlink channel estimation in FDD massive MIMO systems. Throughout the simulations, we assume that the base station is equipped with a uniform linear array (ULA) with $N = 100$ omnidirectional antennas spaced a half wavelength apart. It is worth noting that there is no limit on the array geometry for the proposed pilot design to apply. That is, it can be applied to an arbitrary array geometry, such as two-dimensional array geometry [41].

In order to remove the power differences among different user channels and the effects of user channel SNR scaling on absolute error levels, we utilize the normalized MSE (NMSE), defined as

$$\text{NMSE}(\boldsymbol{h}_m) = \frac{1}{N_{\text{MC}}} \sum_{q=1}^{N_{\text{MC}}} \frac{\left\| \boldsymbol{h}_m(q) - \hat{\boldsymbol{h}}_m^{\text{MMSE}}(q) \right\|^2}{\left\| \boldsymbol{h}_m(q) \right\|^2}, \ \forall\ m \in \mathcal{M}, \tag{33}$$

to evaluate the channel estimation performance, where $\hat{\boldsymbol{h}}_m^{\text{MMSE}}(q)$ is the MMSE estimate of $\boldsymbol{h}_m(q)$, i.e., the $m$-th user's channel obtained in the $q$-th Monte Carlo trial. Here, $N_{\text{MC}} = 5,000$ is the number of Monte-Carlo trials. The step size for the iterative search for the optimal pilot is set as $\gamma = 0.1$.

In our signal model, both the pilot matrix optimization and the channel vector estimation depend on the *a prior* knowledge of the downlink channel vector. In practical applications, this knowledge is unavailable and needs to be estimated before performing the pilot optimization and channel estimation. Considering the time-varying nature of wireless communications, it is more appropriate to model the downlink channel vector as a mixture distribution, e.g., Gaussian mixture distribution considered in this paper, rather than a single, smooth Gaussian distribution despite that the latter offers a closed-form solution [26].

### A. Gaussian mixture approximation

It is well known that the EM algorithm performs maximum likelihood estimation of Gaussian mixture distribution [38], and there are some approximations (see [39] and references therein). However, the EM algorithm may not be the best choice for the channel characterization in massive MIMO systems due to not only the high dimensionality of channel vectors but also the limited training samples especially at the beginning of cell handover. Here, we adopt a piecewise-Gaussian approximation to model the Gaussian mixture distribution of channel vectors in massive MIMO.

In typical massive MIMO cellular systems, the channel vector from the base station to the user terminal is highly correlated due to the narrow angular spread [21, 47], and can be modeled by

$$\boldsymbol{h} = \boldsymbol{R}_{\boldsymbol{hh}}^{\frac{1}{2}} \boldsymbol{v}, \tag{34}$$

where $\boldsymbol{R}_{\boldsymbol{hh}} \succeq \boldsymbol{0}$ denotes a positive semi-definite channel covariance matrix, and $\boldsymbol{v} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ denotes a zero-mean complex-valued Gaussian random vector. In this paper, as we model the channel vector as a Gaussian mixture vector with $K_m$ components, the channel associated with the $m$-th user is expressed as

$$\boldsymbol{h}_m = \left(\boldsymbol{R}_{\boldsymbol{h}_m \boldsymbol{h}_m}^{(k)}\right)^{\frac{1}{2}} \boldsymbol{v}, \ \forall\ m \in \mathcal{M}, \tag{35}$$

with an activation probability of $p_k^{(m)}$, where $k \in \mathcal{K}_m$. Then, the pdf of channel vector is equivalent to the expression of $f(\boldsymbol{h}_m)$ in (4).

A Gaussian mixture model for channel vectors of users is learned from the *a priori* power azimuth spread of the user

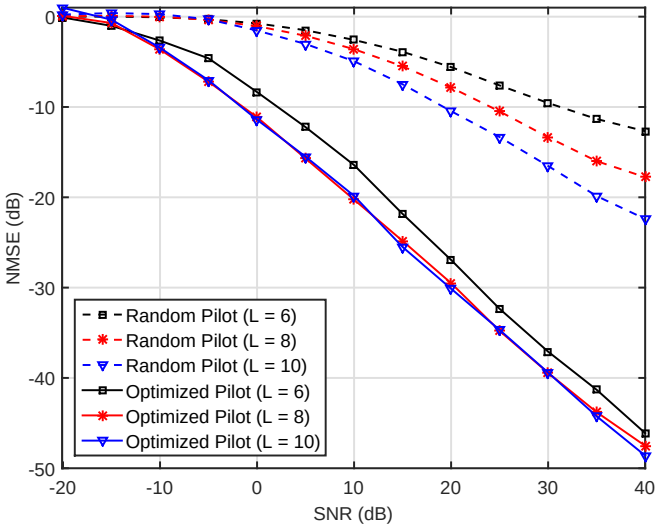


Fig. 2. NMSE performance comparison of the channel estimation versus the input SNR of the channel for different numbers of pilot symbols.

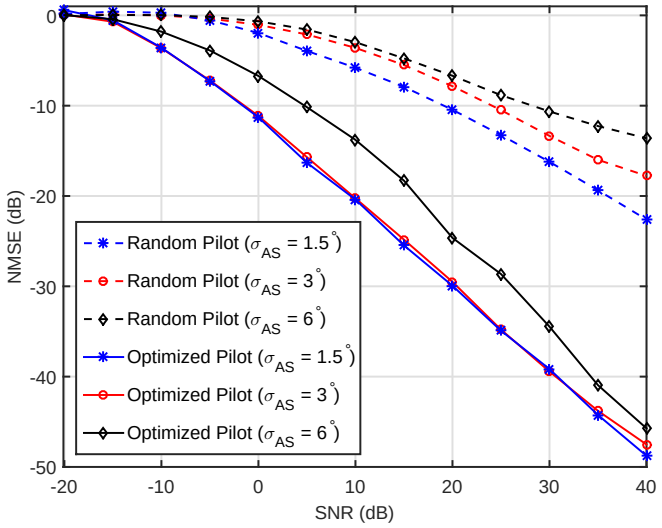


Fig. 3. NMSE performance comparison of the channel estimation versus the input SNR of the channel for different practical azimuth spread.

channel. Specifically, the channel covariance matrix in (35) is generated as

$$\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)} = \int_{\mathcal{A}^{(k)}} \sigma_{\boldsymbol{h}_m}^2 \boldsymbol{a}(\theta)\boldsymbol{a}^{\mathrm{H}}(\theta)d\theta \tag{36}$$

according to the piecewise-Gaussian approximation, where $\boldsymbol{a}(\theta) = \left[1, e^{-j\pi\sin\theta}, \cdots, e^{-j\pi(N-1)\sin\theta}\right]^{\mathrm{T}}$ is the steering vector of the ULA, $\sigma_{\boldsymbol{h}_m}^2$ is the channel power of the $m$-th user, and $\mathcal{A}^{(k)}$ denotes the $k$-th observation region at the base station (e.g., $\mathcal{A}^{(k)} \bigcap \mathcal{A}^{(k')} = \emptyset, \forall\ k, k' \in \mathcal{K}_m$ and $\bigcup_{k\in\mathcal{K}_m} \mathcal{A}^{(k)} = (-\pi/2, \pi/2]$ for the ULA). In the simulations, the observation regions are assumed to have the same value of $\mathcal{A}^{(k)} = 1°, \forall\ k \in \mathcal{K}_m$. The corresponding probability of the $k$-th Gaussian component, which reflects the power azimuth spread, can be modeled by a Laplacian distribution as [52–54]

$$p_k^{(m)} = \frac{1}{\sqrt{2}\sigma_{AS}^{(m)}} e^{-\frac{\sqrt{2}|\theta_k - \bar{\theta}_m|}{\sigma_{AS}^{(m)}}}, \tag{37}$$

where $\bar{\theta}_m$ and $\sigma_{AS}^{(m)}$ respectively denote the mean DOA and the azimuth spread of downlink channel associated with the $m$-th user. Although the Laplacian distribution is the most popular distribution in the typical outdoor propagations, there are other classes of distributions that are applicable in certain circumstances [54].

### *B. Single-user scenarios*

In the first example, a simple scenario with a single user is considered, where the mean DOA is randomly distributed as a uniform distribution over the interval $(-90°, 90°]$, i.e., $\bar{\theta} \sim \mathcal{U}(-90°, 90°]$, and the azimuth spread is set as $\sigma_{AS} = 3°$. Both the mean DOA and the azimuth spread are assumed to be known at the base station.

Numerical results show that the maximum rank of the channel covariance matrices over different Gaussian components is 8, i.e., $\max_{k\in\mathcal{K}} r^{(k)} = \max_{k\in\mathcal{K}} \mathrm{rank}\left(\boldsymbol{R}_{\boldsymbol{hh}}^{(k)}\right) = 8$. To understand the impact of the number of pilot symbols on the channel estimation performance, we consider three different choices of the number of pilot symbols compared with the maximum rank of the channel covariance matrices, i.e., $L = 10 > \max_{k\in\mathcal{K}} \mathrm{rank}\left(\boldsymbol{R}_{\boldsymbol{hh}}^{(k)}\right)$, $L = 8 = \max_{k\in\mathcal{K}} \mathrm{rank}\left(\boldsymbol{R}_{\boldsymbol{hh}}^{(k)}\right)$, and $L = 6 < \max_{k\in\mathcal{K}} \mathrm{rank}\left(\boldsymbol{R}_{\boldsymbol{hh}}^{(k)}\right)$, respectively. In Fig. 2, we depict the NMSEs of the channel estimation versus the input SNR of the channel with different lengths of pilot symbols, where the optimized pilot symbols and the random pilot symbols are compared. The NMSE performance is clearly a function of the input SNR for both the optimized and random pilot symbols, where the channels are scaled by $\sqrt{\mathrm{SNR}}$ to model varying the quality of channel. From Fig. 2, it is observed that the channel estimation performance can be greatly improved by using the proposed pilot symbols as compared to the random generated pilot symbols for the fixed number of pilot symbols. The performance advantage becomes more pronounced as the input SNR increases. It is also observed that, when the number of the optimized pilot symbols reaches the maximum rank of the channel covariance matrices over different Gaussian components, the channel estimation performance cannot be further improved by increasing the number of pilot symbols. On the contrary, when the number of pilot symbols is less than the maximum rank of the channel covariance matrices, the channel estimation performance with the proposed pilot is unstable, and is worse than that of the proposed pilot which number is no less than the maximum rank of the channel covariance matrices.

As the approximated mutual information gradient (28) shows, the proposed pilot optimization depends on the statistical information of the channel vectors to be estimated. Hence, it is necessary to study the robustness of the proposed pilot optimization for channel estimation against the perturbation of the channel knowledge. In Fig. 3, we consider a situation where the assumed azimuth spread in the *a priori* channel distribution remains 3°, whereas the actual azimuth spread varies between 1.5° (smaller than the assumed value) and 6° (larger than the assumed value). It is clear that there is no

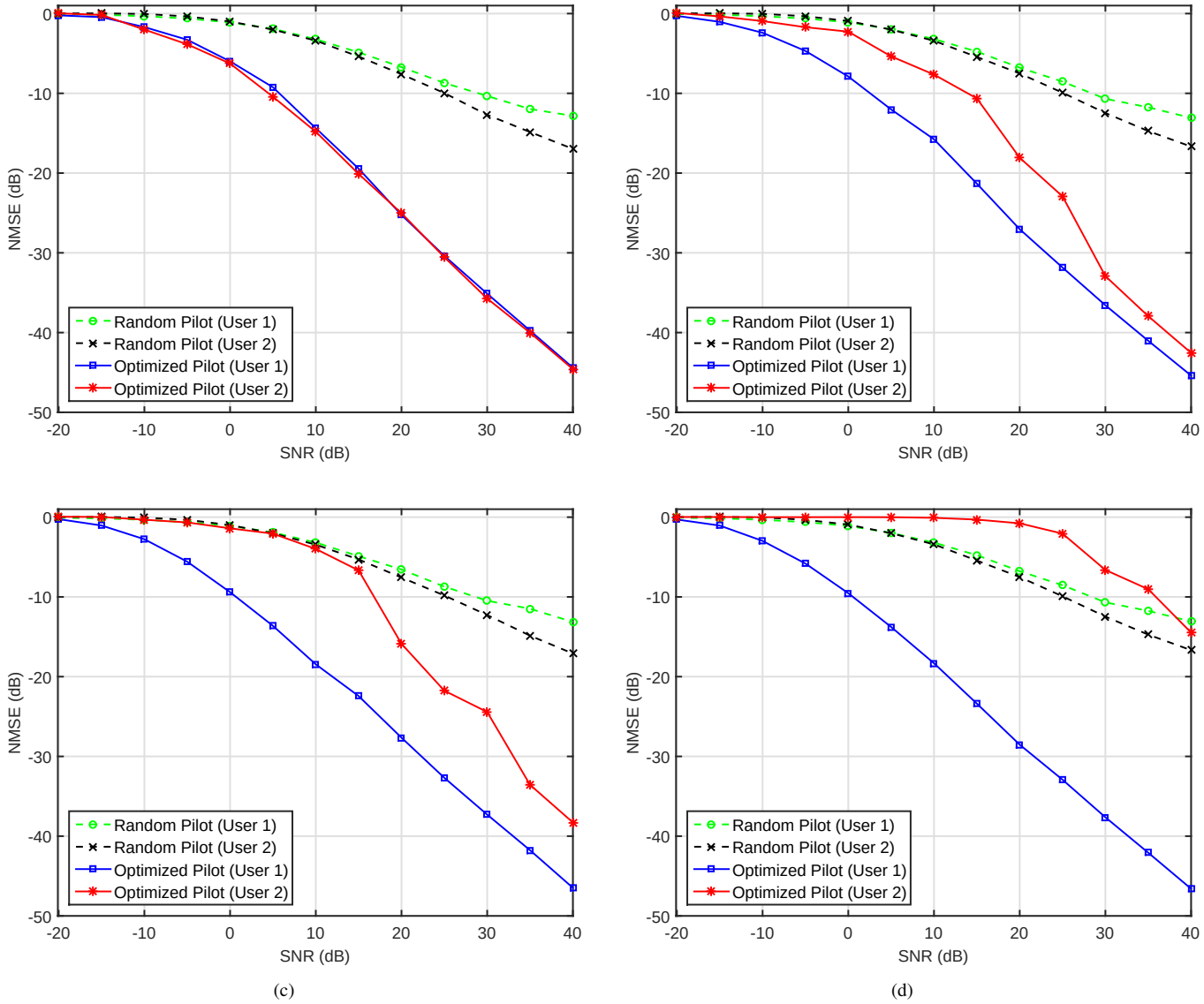


Fig. 4. NMSE performance comparison of the channel estimation versus the input SNR of the channels. (a) Equal user weight ($w_1 = w_2 = 0.5$); (b) Different user weight ($w_1 = 0.7, w_2 = 0.3$); (c) Different user weight ($w_1 = 0.9, w_2 = 0.1$); (d) Extreme user weight ($w_1 = 1, w_2 = 0$).

obvious performance loss in the estimated channel when the actual azimuth spread is smaller than the assumed one, while a significant performance loss is observed when the actual azimuth spread is higher than the assumed value. Hence, for the proposed pilot optimization to achieve effective channel estimation, the *a priori* distribution should at least cover the actual spreading of the channel to be estimated.

### C. Multi-user scenarios

In the second example, we examine the multi-user case. A system with two separated users is considered as an example without loss of generality. It is assumed that the two users respectively have the mean DOAs of $\bar{\theta}_1 = 0°$ and $\bar{\theta}_2 = 50°$, but with a same azimuth spread $\sigma_{AS}^{(m)} = 3°, m = 1, 2$. The mean DOAs and the azimuth spread are assumed to be known at the base station. The two user channels are further assumed to have the same power.

In Fig. 4, we compare the NMSE of the channel estimation versus the input SNR of the channel with different user weights ($w_m, m \in \mathcal{M}$), where the length of pilot symbols is fixed to be the maximum rank of the channel covariance matrices of all Gaussian components of all users, i.e., $L = 8 = \max_{m\in\mathcal{M}} \max_{k\in\mathcal{K}_m} \text{rank}\big(\boldsymbol{R}^{(k)}_{\boldsymbol{h}_m\boldsymbol{h}_m}\big)$. It is clear from Fig. 4(a) that the two users with equal weights have almost identical channel estimation performance. From Fig. 4(a) to Fig. 4(d), we observe that, as the user weight increases, the corresponding user achieves a better channel estimation performance while the performance of the other user degrades. In the extreme scenario where one user occupies the whole weight (i.e., $w_m$ = 1), as Fig. 3(d) shows, the corresponding user has the same channel estimation performance as that in the first example with a single-user scenario, while the channel estimation performance of the other user is even worse than that with random pilot symbols. Hence, in order to provide the desired quality of service (QoS) to all users served in the same massive MIMO system, we prefer to assign the same

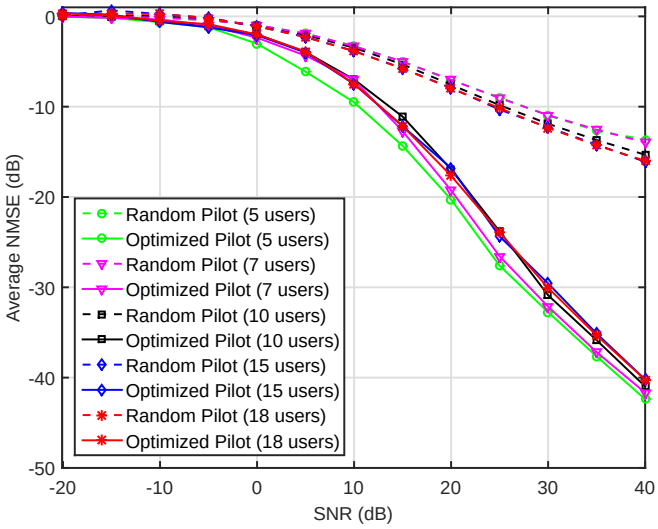


Fig. 5. Average NMSE performance comparison of the channel estimation versus the number of served users.

user weight in the proposed pilot design.

To further evaluate the channel estimation performance with a higher number of served users, we compare the average NMSEs of the channel estimation versus the input SNR of the channel with different numbers of the simultaneously served users. Here, we define the average NMSE over different users as

$$\text{Average NMSE} = \frac{1}{M}\sum_{m=1}^{M}\text{NMSE}(\boldsymbol{h}_m). \tag{38}$$

In the simulations, the number of simultaneously served users increases from 5 to 7, 10, 15 and 18, which mean DOAs are respectively uniform distributed over $[-30°, 30°]$, $[-36°, 36°]$, $[-60°, 75°]$, $[-84°, 84°]$, and $[-85°, 85°]$. All served users are assumed to have the same channel quality (i.e., SNR) with the same azimuth spread $\sigma_{AS}^{(m)} = 3°, \forall m \in \mathcal{M}$ and the same weights $w_m = \frac{1}{M}, \forall m \in \mathcal{M}$ (e.g., $w_m = \frac{1}{5}$ for 5 users). It is observed from Fig. 5 that the average channel estimation performance of the proposed pilot design method degrades as the number of served users increases. Nevertheless, the performance advantage of the proposed pilot design method over the random pilot design remains significant, especially for high-gain channels. An interesting observation is that, when we continue to increase the number of served users, the average channel estimation performance of the proposed pilot will no longer degrade.

## VI. Conclusion

Pilot overhead for FDD downlink channel estimation is a challenging problem in massive MIMO systems equipped with a very large number of antennas at the base station. By modeling the channel vector as a flexible and tractable Gaussian mixture distribution, we first proved that the channel can be perfectly recovered in the asymptotic high-SNR regime, when the number of pilot symbols is not less than the maximum rank of the channel covariance matrices of all Gaussian components of all users. Then, we proposed an information-theoretic pilot design by maximizing the weighted sum of the mutual information between the measurements of all served users and their corresponding channel vectors on the Grassmannian manifold. Hence, the proposed pilot can serve an arbitrary number of users in FDD massive MIMO systems by exploiting the *a priori* knowledge of the channel vectors to be estimated. With the available Gaussian mixture distribution, there is a closed-form solution to the underdetermined channel estimation problem under the MMSE criterion. Simulation results demonstrated that the proposed pilot outperforms the random pilot in terms of the NMSE of the channel estimation versus the input SNR of the channel.

## Appendix A

*Proof:* Similar to the proof in [26], the lower bound of the MSE of the MMSE estimator of the channel vector $\boldsymbol{h}_m$ can be written as (39), shown at the bottom of the page, where the first term vanishes as $\sigma_{n_m}^2 \to 0$.

Now, let us examine the second term. Because the rank of $\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}$ is $r_m^{(k)}$, the eigen-decomposition of $\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}$ can be denoted as

$$\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)} = \boldsymbol{U}_m^{(k)}\boldsymbol{\Lambda}_m^{(k)}\big(\boldsymbol{U}_m^{(k)}\big)^{\mathrm{H}}, \tag{40}$$

$$\begin{aligned}
\varepsilon_{\text{Lower}}^{(m)} &= \sum_{k\in\mathcal{K}_m} p_k^{(m)}\text{Tr}\Big[\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)} - \boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{\Phi}^{\mathrm{H}}\big(\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{\Phi}^{\mathrm{H}} + \sigma_{n_m}^2\boldsymbol{I}\big)^{-1}\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\Big] \\
&= \sum_{k\in\mathcal{K}_m} p_k^{(m)}\text{Tr}\Big[\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big(\boldsymbol{I} - \big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathrm{H}}\big(\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{\Phi}^{\mathrm{H}} + \sigma_{n_m}^2\boldsymbol{I}\big)^{-1}\boldsymbol{\Phi}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big)\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big] \\
&= \sum_{k\in\mathcal{K}_m} p_k^{(m)}\text{Tr}\Big[\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big(\boldsymbol{I} + \sigma_{n_m}^{-2}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Phi}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big)^{-1}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big] \\
&= \sum_{k\in\mathcal{K}_m} p_k^{(m)}\text{Tr}\Big[\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{V}_m^{(k)}\big(\boldsymbol{I} + \sigma_{n_m}^{-2}\boldsymbol{\Gamma}_m^{(k)}\big)^{-1}\big(\boldsymbol{V}^{(m_k)}\big)^{\mathrm{H}}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big] \\
&= \sum_{k\in\mathcal{K}_m} p_k^{(m)}\text{Tr}\Big[\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{V}_m^{(k)}\big(\boldsymbol{I} + \sigma_{n_m}^{-2}\boldsymbol{\Gamma}_m^{(k)}\big)^{-1}\big(\boldsymbol{V}^{(m_k)}\big)^{\mathrm{H}}\Big] \\
&= \sum_{k\in\mathcal{K}_m} p_k^{(m)}\sum_{i=1}^{r_m^{(k)}}\big(1+\gamma_i/\sigma_{n_m}^2\big)^{-1}\big(\boldsymbol{v}_{m_i}^{(k)}\big)^{\mathrm{H}}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{v}_{m_i}^{(k)} + \sum_{k\in\mathcal{K}_m} p_k^{(m)}\sum_{i=r_m^{(k)}+1}^{N}\big(\boldsymbol{v}_{m_i}^{(k)}\big)^{\mathrm{H}}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\boldsymbol{v}_{m_i}^{(k)},
\end{aligned} \tag{39}$$

where $\boldsymbol{U}_m^{(k)} \in \mathbb{C}^{N\times r_m^{(k)}}$ and $\boldsymbol{\Lambda}_m^{(k)} \in \mathbb{C}^{r_m^{(k)}\times r_m^{(k)}}$. We can write

$$\begin{aligned}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathrm{H}} &= \boldsymbol{U}_m^{(k)}\big(\boldsymbol{\Lambda}_m^{(k)}\big)^{\frac{1}{2}}\big(\boldsymbol{U}_m^{(k)}\big)^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}}\\ &= \boldsymbol{U}_m^{(k)}\boldsymbol{C}_m^{(k)},\end{aligned} \tag{41}$$

where $\boldsymbol{C}_m^{(k)} = \big(\boldsymbol{\Lambda}_m^{(k)}\big)^{\frac{1}{2}}\big(\boldsymbol{U}_m^{(k)}\big)^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}} \in \mathbb{C}^{r_m^{(k)}\times L}$. When $L \geq \max_{m\in\mathcal{M}}\max_{k\in\mathcal{K}_m} r_m^{(k)}$ and the pilot symbols of $\boldsymbol{\Phi}$ are randomly generated according to some distribution, the matrix $\boldsymbol{C}_m^{(k)}$ has a full row rank with probability one, i.e., $\mathrm{rank}\big(\boldsymbol{C}_m^{(k)}\big) = r_m^{(k)}$. Hence, let $\mathrm{Range}(\cdot)$ denote the range of a matrix, i.e., the column space spanned by its column vectors. Then,

$$\begin{aligned}&\mathrm{Range}\Big(\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathrm{H}}\Big)\\ =\ &\mathrm{Range}\Big(\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Phi}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\Big)\\ =\ &\mathrm{Range}\big(\boldsymbol{U}_m^{(k)}\big), \quad \forall\ k\in\mathcal{K}_m, m\in\mathcal{M}.\end{aligned} \tag{42}$$

Therefore, we have

$$\begin{aligned}\big(\boldsymbol{u}_{m_i}^{(k)}\big)^{\mathrm{H}}\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)} &= \big(\boldsymbol{v}_{m_i}^{(k)}\big)^{\mathrm{H}}\boldsymbol{U}_m^{(k)}\\ &= \boldsymbol{v}_{m_i}^{(k)}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Phi}\big(\boldsymbol{R}_{\boldsymbol{h}_m\boldsymbol{h}_m}^{(k)}\big)^{\frac{1}{2}}\\ &= \boldsymbol{0}, \quad \forall\ i = r_m^{(k)}+1,\cdots,N.\end{aligned} \tag{43}$$

Hence, the second term in (39) disappears when the number of pilot symbols is no less than the maximum rank of the channel covariance matrices of all Gaussian components of all users, i.e., $L \geq \max_{m\in\mathcal{M}}\max_{k\in\mathcal{K}_m} r_m^{(k)}$. That is to say, the lower bound of the MSE of the MMSE estimate of $\boldsymbol{h}_m$ approaches zero in the limit of vanishing noise,

$$\lim_{\mathrm{SNR}_m\to\infty} \varepsilon_{\mathrm{Lower}}^{(m)} = 0. \tag{44}$$

In the Gaussian mixture distribution, the upper and lower bounds of the MSE of the MMSE estimate approach each other with increasing SNR, and they must coincide as the SNR tends to infinity, i.e., [42],

$$\lim_{\mathrm{SNR}_m\to\infty} \varepsilon_{\mathrm{Upper}}^{(m)} = \lim_{\mathrm{SNR}_m\to\infty} \varepsilon_{\mathrm{Lower}}^{(m)} = 0, \tag{45}$$

which implies that

$$\lim_{\mathrm{SNR}_m\to\infty} \mathbb{E}\left\{\left\|\boldsymbol{h}_m - \hat{\boldsymbol{h}}_m^{\mathrm{MMSE}}\right\|_2^2\right\} = 0. \tag{46}$$

■

## APPENDIX B

For completeness, we briefly recall the result in [36] for the derivation of the approximated differential entropy (19). Performing the first-order Taylor series expansion of $\log\left[f(\boldsymbol{y}_m)\right]$ around $\boldsymbol{y}_0^{(m)} = \mathbb{E}[\boldsymbol{y}_m] = \boldsymbol{u}_{\boldsymbol{y}_m}$, i.e., the mean value of $\boldsymbol{y}_m$, we have

$$\log\left[f(\boldsymbol{y}_m)\right] \approx \log\left[f\big(\boldsymbol{y}_0^{(m)}\big)\right] + \boldsymbol{g}^{\mathrm{H}}\big(\boldsymbol{y}_0^{(m)}\big)\left[\boldsymbol{y}_m - \boldsymbol{y}_0^{(m)}\right], \tag{47}$$

where $\log\left[f\big(\boldsymbol{y}_0^{(m)}\big)\right] = \log\left[\sum_{k\in\mathcal{K}_m} p_k^{(m)} f^{(k)}\big(\boldsymbol{y}_0^{(m)}\big)\right]$, and $\boldsymbol{g}\big(\boldsymbol{y}_0^{(m)}\big) = \frac{d}{d\boldsymbol{y}_m}\log\left[f(\boldsymbol{y}_m)\right]\big|_{\boldsymbol{y}_m=\boldsymbol{y}_0^{(m)}}$ denotes the first derivative of $\log\left[f(\boldsymbol{y}_m)\right]$ evaluated at the point $\boldsymbol{y}_0^{(m)}$. Substituting (47) into the definition of differential entropy, we obtain the approximated differential entropy (48), shown at the bottom of the page.


## ACKNOWLEDGMENT

The authors would like to thank the Associate Editor Prof. Bruno Clerckx and the anonymous reviewers for their helpful comments and suggestions.


## REFERENCES


[1] Y. Gu and Y. D. Zhang, "Pilot design for Gaussian mixture channel estimation in massive MIMO," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, Canada, Apr. 2018, pp. 3266–3270.

[2] F. Rusek, D. Persson, Buon K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[3] R. C. de Lamare, "Massive MIMO systems: Signal processing challenges and future trends," *URSI Radio Sci. Bull.*, vol. 2013, no. 347, pp. 8–20, Dec. 2013.

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

[5] 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.

[6] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, "MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 13–21, Oct. 2015.


$$\begin{aligned}h(\boldsymbol{y}_m) &= -\int \sum_{k\in\mathcal{K}_m} p_k^{(m)} f^{(k)}\big(\boldsymbol{y}_m\big)\log\left[f(\boldsymbol{y}_m)\right] d\boldsymbol{y}_m\\ &\approx -\sum_{k\in\mathcal{K}_m} p_k^{(m)} \int f^{(k)}\big(\boldsymbol{y}_m\big)\left\{\log\left[f\big(\boldsymbol{y}_0^{(m)}\big)\right] + \boldsymbol{g}^{\mathrm{H}}\big(\boldsymbol{y}_0^{(m)}\big)\left[\boldsymbol{y}_m - \boldsymbol{y}_0^{(m)}\right]\right\} d\boldsymbol{y}_m\\ &= -\sum_{k\in\mathcal{K}_m} p_k^{(m)}\left\{\log\left[f\big(\boldsymbol{y}_0^{(m)}\big)\right] + \boldsymbol{g}^{\mathrm{H}}\big(\boldsymbol{y}_0^{(m)}\big)\left[\boldsymbol{u}_{\boldsymbol{y}_m}^{(k)} - \boldsymbol{y}_0^{(m)}\right]\right\}\\ &= -\log\left[f\big(\boldsymbol{y}_0^{(m)}\big)\right] - \boldsymbol{g}^{\mathrm{H}}\big(\boldsymbol{y}_0^{(m)}\big)\left[\sum_{k\in\mathcal{K}_m} p_k^{(m)}\boldsymbol{u}_{\boldsymbol{y}_m}^{(k)} - \boldsymbol{y}_0^{(m)}\right]\\ &= -\log\left[f\big(\boldsymbol{y}_0^{(m)}\big)\right].\end{aligned} \tag{48}$$

[7] D. Ciuonzo, P. S. Rossi, and S. Dey, "Massive MIMO channel-aware decision fusion," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 604–619, Feb. 2015.

[8] X.-C. Gao, J.-K. Zhang, Z.-Y. Wang and J. Jin, "Optimal precoder design maximizing the worst-case average received SNR for massive distributed MIMO systems," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 589–592, Apr. 2015.

[9] A. Shirazinia, S. Dey, D. Ciuonzo, and P. S. Rossi, "Massive MIMO for decentralized estimation of a correlated source," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2499–2512, May 2016.

[10] Y. Gu, Y. D. Zhang, and N. A. Goodman, "Optimized compressive sensing-based direction-of-arrival estimation in massive MIMO," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, Mar. 2017, pp. 3181–3185.

[11] F. Zhu, N. Wu, and Q. Liang, "Channel estimation for massive MIMO with 2-D nested array deployment," *Phys. Commun.*, vol. 25, pp. 432–437, Dec. 2017.

[12] Z. Shao, R. C. de Lamare, L. T. N. Landau, "Iterative detection and decoding for large-scale multiple-antenna systems with 1-bit adcs," *IEEE Wirelesss Commun. Lett.*, vol. 7, no. 3, pp. 476–479, June 2018.

[13] C. Zhou, Y. Gu, X. Fan, Z. Shi, G. Mao, and Y. D. Zhang, "Direction-of-arrival estimation for coprime array via virtual array interpolation," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5956–5971, Nov. 2018.

[14] C. Zhou, Y. Gu, Z. Shi, and Y. D. Zhang, "Off-grid direction-of-arrival estimation using coprime array interpolation," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1710–1714, Nov. 2018.

[15] 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.

[16] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 2, pp. 264–273, Feb. 2013.

[17] R. R. Muller, L. Cottatellucci, and M. Vehkapera, "Blind pilot decontamination," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 773–786, Oct. 2014.

[18] 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, Sept. 2017.

[19] X. Rao and V. K. N. 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.

[20] 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.

[21] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing – The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[22] 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.

[23] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2868–2882, May 2015.

[24] C. Sun, X. Q. Gao, S. Jin, M. Matthaiou, Z. Ding, and C. Xiao, "Beam division multiple access transmission for massive MIMO communications," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2170–2184, Jun. 2015.

[25] W. Shen, L. Dai, B. Shim, S. Mumtaz, and Z. Wang, "Joint CSIT acquisition based on low rank matrix completion for FDD massive MIMO systems," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2178–2181, Dec. 2015.

[26] J. Fang, X. Li, H. Li, and F. Gao, "Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1935–1947, Mar. 2017.

[27] 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.

[28] M. B. Khalilsarai, S. Haghighatshoar, X. Yi, and G. Caire, "FDD massive MIMO via UL/DL channel covariance extrapolation and active channel sparsification", arXiv:1803.05754.

[29] S. Haghighatshoar, M. B. Khalilsarai, and G. Caire, "Multi-band covariance interpolation with applications in massive MIMO," in *Proc. IEEE Int. Symp. Inf. Theory*, Vail, CO, June 2018, pp. 386–390.

[30] Y. Han and J. Lee, "Two-stage compressed sensing for millimeter wave channel estimation," in *Proc. IEEE Int. Symp. Inf. Theory*, Barcelona, Spain, July 2016, pp. 860–864.

[31] A. Liu, V. Lau, M. L. Honig, and L. Lian, "Compressive rf training and channel estimation in massive mimo with limited rf chains," in *IEEE Int. Conf. Commun.*, Paris, France, May 2017.

[32] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.

[33] S. Bazzi and W. Xu, "Downlink training sequence design for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4732–4744, Sept. 2017.

[34] 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.

[35] K. Lemke-Rust and C. Paar, "Gaussian mixture models for higher-order side channel analysis," in *Proc. Workshop Cryptograph. Hardw. Embedded Syst.*, Vienna, Austria, Sept. 2007, pp. 14–27.

[36] Y. Gu, N. A. Goodman, and A. Ashok, "Radar target profiling and recognition based on TSI-optimized compressive sensing kernel," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3194–3207, June 2014.

[37] Y. Gu and N. A. Goodman, "Information-theoretic compressive sensing kernel optimization and Bayesian Cramér-Rao bound for time delay estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4525–4537, Sept. 2017.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[39] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.

[40] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, June 2001.

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

[42] J. T. Flåm, S. Chatterjee, K. Kansanen, and T. Ekman, "On MMSE: A linear model under Gaussian mixture statistics," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3840–3844, Jul. 2012.

[43] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[44] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition. Hoboken, NJ, USA: John Wiley & Sons, 2006.

[46] J.-Y. Le Boudec, "Rate adaptation, congestion control and fairness: A tutorial," Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, 2003, Tech. Rep. Available at http://ica1www.epfl.ch/PS_files/LEB3132.pdf.

[47] D. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multi-element antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.

[48] A. Edelman, T. Arias, and S. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Analysis Applicat.*, vol. 20, no. 2, pp. 303–353, 1998.

[49] T. E. Abrudan, J. Eriksson, and V. Koivunen, "Steepest descent algorithms for optimization under unitary matrix constraint," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1134–1147, Mar. 2008.

[50] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 635–650, Mar. 2002.

[51] P. H. Schönemann, "A generalized solution of the orthogonal Procrustes problem", *Psychometrika*, vol. 31, no. 1, pp. 1–10, Mar. 1966.

[52] K. I. Pedersen, P. E. Mogensen, and B. H. Fleury, "Power azimuth spectrum in outdoor environments," *Electron. Lett.*, vol. 33, no. 18, pp. 1583–1584, Aug. 1997.

[53] L. M. Correia, *Wireless Flexible Personalized Communications*. Hoboken, NJ, USA: John Wiley & Sons, 2001.

[54] S. Saunders and A. Aragón-Zavala, *Antennas and Propagation for Wireless Communication Systems*, 2nd Edition. Hoboken, NJ, USA: John Wiley & Sons, 2007.

**Yujie Gu** (SM'16) received the Ph.D. degree in electronic engineering from Zhejiang University, Hangzhou, China, in 2008. After graduation, he held multiple research positions in China, Canada, Israel, and the USA. He is currently a Postdoctoral Research Associate with Temple University, Philadelphia, PA, USA. His research interests are in statistical and array signal processing including adaptive beamforming, compressive sensing, MIMO systems, radar imaging, target localization, waveform design, etc.

Dr. Gu is currently a Subject Editor for *Electronics Letters*, an Associate Editor for *Signal Processing*, *IET Signal Processing*, *Circuits, Systems and Signal Processing*, and *EURASIP Journal on Advances in Signal Processing*. He is also the Lead Guest Editor of special issue on Source Localization in Massive MIMO for the *Digital Signal Processing*. He has been a member of the Technical Program Committee of multiple international conferences including IEEE ICASSP. He is also a Special Sessions Co-Chair of the 2020 IEEE Sensor Array and Multichannel Signal Processing Workshop.

**Yimin D. Zhang** (F'19) received the Ph.D. degree from the University of Tsukuba, Tsukuba, Japan, in 1988. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, College of Engineering, Temple University, Philadelphia, PA, USA. From 1998 to 2015, he was a Research Faculty with the Center for Advanced Communications, Villanova University, Villanova, PA, USA. His research interests lie in the areas of statistical signal and array processing, including compressive sensing, machine learning, convex optimization, time-frequency analysis, MIMO systems, radar imaging, direction finding, target localization and tracking, wireless and cooperative networks, and jammer suppression, with applications to radar, wireless communications, and satellite navigation. He has authored/coauthored more than 340 journal and conference papers and 14 book chapters in these areas.

Dr. Zhang is an Associate Editor for the *IEEE Transactions on Signal Processing* and an Editor for the *Signal Processing* journal. He was an Associate Editor for the *IEEE Signal Processing Letters* during 2006–2010, and an Associate Editor for the *Journal of the Franklin Institute* during 2007–2013. He is a member of the Sensor Array and Multichannel Technical Committee and the Signal Processing Theory and Methods Technical Committee of the IEEE Signal Processing Society. He was the Technical Co-Chair of the 2018 IEEE Sensor Array and Multichannel Signal Processing Workshop. He was the recipient of the 2016 IET Radar, Sonar & Navigation Premium Award and the 2017 IEEE Aerospace and Electronic Systems Society Harry Rowe Mimno Award, and coauthored a paper that received the 2018 IEEE Signal Processing Society Young Author Best Paper Award.