Structured channel covariance estimation from limited samples for large antenna arrays

Tianyu Yang1, Mahdi Barzegar Khalilsarai1, Saeid Haghighatshoar2,3 and Giuseppe Caire1*

Abstract
In massive multiuser multiple antenna systems, the knowledge of the users’ channel covariance matrix is crucial for minimum mean square error channel estimation in the uplink and it plays an important role in several multiuser beamforming schemes in the downlink. Due to the large number of base station antennas, accurate covariance estimation is challenging especially in the case where the number of samples is limited and thus comparable to the channel vector dimension. As a result, the standard sample covariance estimator may yield a too large estimation error which in turn may yield significant system performance degradation with respect to the ideal channel covariance knowledge case. To address such problem, we propose a method based on a parametric representation of the channel angular scattering function. The proposed parametric representation includes a discrete specular component which is addressed using the well-known MUltiple SIgnal Classification (MUSIC) method, and a diffuse scattering component, which is modeled as the superposition of suitable dictionary functions. To obtain the representation parameters, we propose two methods, where the first solves a nonnegative least-squares problem and the second maximizes the likelihood function using expectation–maximization. Our simulation results show that the proposed methods outperform the state of the art with respect to various estimation quality metrics and different sample sizes.

Keywords: Massive MIMO, Covariance estimation, MUSIC, Maximum likelihood, Nonnegative least squares, Angular scattering function

1 Introduction
Massive multiple-input multiple-output (MIMO) communication system, where the number of base station (BS) antennas $M$ is much larger than the number of single antenna users, has been shown to achieve high spectral efficiency in wireless cellular networks and to enjoy various system level benefit, such as energy efficiency, inter-cell interference reduction, and dramatic simplification of user scheduling (e.g., see [2, 3]). In a large number of papers on the subject, the knowledge of the uplink (UL) and downlink (DL) channel covariance matrix, i.e., of the correlation structure of the channel antenna coefficients at the BS array, is assumed and used for a variety of purposes, such as minimum mean square error (MMSE) UL channel estimation and pilot decontamination [4–6], efficient DL multiuser precoding/beamforming design, especially in the frequency...
division duplexing (FDD) case [6–10], and multiuser DL precoding design based on statistical channel state information (CSI) [11–13].

Under the usual assumption of wide-sense stationary (WSS) uncorrelated scattering (US) [6, 7, 9, 14, 15], the channel vector evolves over time as a vector-valued WSS process and its spatial correlation is frequency-invariant over a frequency interval significantly larger than the signal bandwidth, although much smaller than the carrier frequency. In particular, in an orthogonal frequency division multiplexing (OFDM) system, the channel spatial covariance is independent of time (OFDM symbol index) and frequency (subcarrier index). In order to capture the actual WSS statistics, samples sufficiently spaced over time and frequency must be collected (e.g., one sample for every resource block in the UL slots over which a given user is active). On the other hand, the WSS model holds only locally, over time intervals where the propagation geometry (angle and distances of multipath components) does not change significantly. Such time interval, referred to as “geometry coherence time,” is several orders of magnitude larger than the coherence time of the small-scale channel coefficients. For a typical mobile urban environment, the channel geometry coherence time is of the order of seconds, while the small-scale fading coherence time is of the order of milliseconds (see [19] and references therein). Hence, the BS can collect a window of tens-to-hundreds of noisy channel snapshots from UL pilot symbols sent by any given user and use them to produce a “local” estimate of the corresponding user covariance matrix which remains valid during a channel geometry coherence time. Such estimation must be repeated, or updated, at a rate that depends on the propagation scenario and user-BS relative motion. This discussion points out that the number of samples $N$ available for covariance estimation is limited and often comparable or even less than the number of BS antennas $M$. In general, the accurate estimation of a high-dimensional $M \times M$ spatial channel covariance matrix (both in the UL and in the DL) from a limited number of noisy channel realizations (samples) is generally a difficult task.

The simplest way to estimate the channel covariance matrices is the sample covariance estimator. Such estimation is asymptotically unbiased and consistent and works well when $N \gg M$. Unfortunately, as already noticed above, this is typically not the case in massive MIMO. Hence, the goal of this paper is to devise new parametric estimators that outperform the sample covariance estimator, as well as the other state-of-the-art methods proposed in the literature. In addition, an attractive feature of the proposed parametric estimation is that it lends itself to the extrapolation of the estimated channel covariance matrix from the UL to the DL frequency band. As the channel samples are collected by the BS through pilot symbols sent by the users on the UL, the UL channel covariance can be directly estimated via our method. However, as pointed out above, several schemes for DL multiuser precoding/beamforming in FDD systems make use of the user channel covariance matrix in the DL, which differs from the UL covariance since the frequency separation between the UL and the DL bands is large. The estimation of the DL covariance from UL channel samples has been considered in several works and

---

1 The channel models in 3GPP standard TR 38.901 Subclause 7.5 equation 7.5-22 [16] and TR 25.996 equation 5.4-1 [17] specify the US property. Section 3 of the QuaDRiGa documentation [18] also implicitly includes this assumption.

2 See also discussion in [15, Section I] and [20] for a recent study based on data gathered from a channel sounder in an urban environment with a receiver moving at vehicular speed.
it is generally another challenging task [7, 21–25]. We shall see that our scheme is able to accurately estimate the DL channel covariance by extrapolating (over frequency) the estimated parametric model in the UL.

1.1 Related work
Covariance estimation from limited samples is a very well-known classical problem in statistics. In the “large system regime,” when both the vector dimension $M$ and the number of samples $N$ grow large with fixed sampling ratio $N/M$ of samples per dimension, a vast literature focused on the asymptotic eigenvalue distribution of sample covariance matrices estimator under specific statistical assumptions (see, e.g., [26–28] and the references therein). Covariance estimation has reemerged recently in many problems in machine learning, compressed sensing, biology, etc. (see, e.g., [29–32] from some recent results). What makes these recent works different from the classical ones is the highly structured nature of the covariance matrices in these applications. A key challenge in these new applications is to design efficient estimation algorithms able to take advantage of the underlying structure to recover the covariance matrix with a small sample size.

As explained in the sequel, MIMO covariance matrices are highly structured due to the particular spatial configuration of the BS antennas and wireless propagation scenario. So, it is very important to design efficient algorithms that are able to take advantage of this specific structure. Previous works have considered either diffuse scattering (the received power from any given angle direction is infinitesimal [21]), or separable discrete components [33–35]. Interestingly, both classes of state-of-the-art methods incur significant degradation when the actual channel model does not conform to the assumptions. No current work has addressed the simultaneous presence of both discrete components and diffuse scattering. In this paper, we propose a method to address such mixed scattering model. We shall compare our proposed method with alternative approaches which can be regarded as the state of the art for the specific case of wireless massive MIMO channels. The competitor methods shall be briefly discussed when presenting such comparison results.

1.2 Contributions
The main contributions of this work are listed below:

(1) In contrast to the work in [21] that only considers diffuse scattering, and the works in [33–35] that only consider separable discrete components, we propose a method that handles the more realistic mixed scattering model where discrete and diffuse scattering is simultaneously present in the angle domain. It should be noticed that our model is not just “one or many possible alternatives.” In fact, our model encompasses any possible received power distribution over the angle domain, referred to in this work as the angular scattering function (ASF).

(2) The new idea of the proposed general model consists of approximating the ASF in terms of atoms of a dictionary. This dictionary is formed by Dirac delta functions placed at the angle-of-arrival (AoA) of the discrete specular components and by a family of functions obtained by shifts of a template density function over the angle domain, in order to approximate the diffuse component of the ASF. Notice that
the proposed dictionary is partially “on the grid” (the shifts of the template density function at integer multiples of some chosen discretization interval) and partially “off the grid” (the Diracs placed at the unquantized AoAs). In this sense, the method differs substantially from standard compressed sensing schemes and does not assume sparsity of the ASF in the angle domain. In fact, the diffuse scattering component is not sparse at all, but is dense on given angular spread intervals.

(3) In order to estimate the AoA of the discrete components and their number (model order), we propose to use the well-known MUltiple SIgnal Classification (MUSIC) method [36]. In fact, this approach is particularly suited to the problem at hand thanks to the provable asymptotic consistency for the corresponding spiked model. Notice that this property has not been proven for other super-resolution methods based on convexity and atomic norm, which have also the disadvantage of being significantly more computationally complex (see, e.g., [37, Table III]) and numerically unstable.

(4) After obtaining the estimated AoAs of the spikes via MUSIC, we propose two estimators, namely a constrained least-squares estimator and maximum-likelihood (ML) estimator, to estimate the parametric model coefficients. In particular, the maximization of the likelihood function in the ML estimator is obtained via the expectation–maximization (EM) initialized by the constrained least-squares solution.

We demonstrate the advantage of the proposed approach through extensive numerical results based on the channel emulator QUAsi Deterministic Radio channel GenerAtor (QuaDriGa) [18], used routinely in 3GPP standardization to provide a common reference for comparison of different physical layer algorithms. We compare our method with other state-of-the-art schemes, showing that the proposed algorithms outperform the competitors over all benchmarks in a wide range of sampling ratio \(N/M\). It should be noticed that the QuaDriGa channel generator is agnostic and unaware of the proposed ASF representation and generates channel snapshots on the basis of a mixed physical/statistical model. Therefore, our results are not “biased” by generating channels that are “matched” to the assumed model. On the contrary, our results show that the proposed approach is robust to the underlying channel model and provides improvements on the state of the art even if the underlying channel statistics do not reflect exactly the assumed one.

1.3 Notations

An identity matrix with \(K\) columns is denoted as \(I_K\). An all-zero matrix with size \(m \times n\) is denoted as \(0_{m \times n}\). \(\text{diag}(\cdot)\) returns a diagonal matrix. \(\| \cdot \|_F\) returns Frobenius norm of a matrix. \(\delta(\cdot)\) denotes the Dirac delta function. We use \([n]\) to denote the ordered integer set \(\{1, 2, \ldots, n\}\).

2 Methods

This study is theoretical. The algorithms are developed through mathematical equations and implemented using standard scientific computation software (MATLAB). The numerical results of Sect. 5 are obtained by standard computer simulation using
MATLAB, with channel vector snapshots extracted from the publicly available channel simulator (QuaDriGa) [18], which is used in a large number of comparative studies in the framework of 3GPP standardization. All details of the numerical results, including comparison with competing state-of-the-art methods, are fully described in Sect. 5. No proprietary data have been used in this study.

### 3 System model

We consider a typical single-cell massive MIMO communication system, where a BS equipped with a uniform linear array (ULA) of \( M \) antennas communicates with multiple users through a multipath channel. Following the common assumption that the pilot sequences of different users are orthogonal to each other in the time–frequency domain, and since the observations used for channel covariance estimation use only the UL pilots, without loss of generality we focus on a generic user. Figure 1 visualizes the propagation model based on multipath clusters, which is physically motivated and widely adopted in standard channel simulation tools such as QuaDriGa [18]. During UL transmission, on each time–frequency resource block (RB) \( s \), the BS receives an UL user pilot carrying a measurement for the channel vector \( h[s] \). We assume that the window of \( N \) samples collected for covariance estimation is designed such that the samples are enough spaced in the time–frequency domain and resulting in statistically independent channel snapshots \( \{h[s]: s = [N]\} \). Meanwhile, the whole window spans a time significantly shorter than the geometry coherence time so that the WSS assumption holds (see discussion in Sect. 1) and the channel snapshots are identically distributed. The channel vectors are given by [38]

\[
h[s] = \int_{-1}^{1} \rho(\xi; s)a(\xi)d\xi, \ s \in [N],
\] (1)

Notice that in the case of a multi-cell system with pilot contamination the same techniques of this paper can estimate the covariance of the sum of multiple channels, and a pilot decontamination scheme as described in [6] can be applied. However, this goes beyond the scope of this paper.
where $\rho(\xi; s)$ is the channel complex coefficient at the normalized AoA $\xi = \frac{\sin(\theta)}{\sin(\theta_{\max})} \in [-1, 1]$, where $\theta_{\max} \in [0, \pi/2]$ is the maximum array angular aperture; $a(\xi) \in \mathbb{C}^M$ denotes the array response vector as a function of $\xi$, with $m$-th element given by $[a(\xi)]_m = e^{j2\pi d_{0m}^{'}/\lambda_{0}\sin(\theta_{\max})}$, where $d$ denotes the antenna spacing and $\lambda_{0}$ denotes the carrier wavelength. For convenience, we assume the antenna spacing to be $d = \frac{\lambda_{0}}{2\sin(\theta_{\max})}$.

Thus, the array response vector is given as

$$
a(\xi) = \left[1, e^{j\pi \xi}, \ldots, e^{j\pi (M-1)\xi}\right]^T. \quad (2)
$$

The channel coefficient $\rho(\xi; s)$ represents the small-scale multipath fading component at a given AoA, and it is modeled as a complex circularly symmetric Gaussian process with respect to $\xi$. Due to the WSS property, the channel second-order statistics are invariant with respect to the index $s \in [N]$. In particular, $\rho(\xi; s)$ has mean zero and variance $\mathbb{E}[\rho(\xi; s)\rho^*(\xi; s)] = \gamma(\xi)$. The function $\gamma : [-1, 1] \rightarrow \mathbb{R}_+$ is a real nonnegative measure that describes how the channel energy is distributed across the angle domain, and it is referred to as the channel ASF. From (1) and the ASF definition, it follows that the channel spatial covariance matrix, describing the correlation of the channel coefficients at the different antenna elements, is given by

$$
\Sigma_{h} = \mathbb{E}\left[h[s]h[s]^H\right] = \int_{-1}^{1} \gamma(\xi)a(\xi)a(\xi)^Hd\xi. \quad (3)
$$

Notice that $\Sigma_{h}$ is Toeplitz. This fact is verified when all the scattering clusters (see Fig. 1) are in the far field of the BS array. At RB $s$, the received pilot signal at the BS is given as

$$
y[s] = h[s]x[s] + z[s], \quad s \in [N], \quad (4)
$$

where $x[s]$ is the pilot symbol and $z[n] \sim \mathcal{CN}(0, N_0I_M)$ is the additive white Gaussian noise (AWGN). Without loss of generality, we assume that the pilot symbols are normalized as $x[n] = 1, \forall s \in [N]$. The goal of this work is to estimate the channel covariance matrix $\Sigma_{h}$ with the given set of $N$ noisy channel observations $\{y[s] : s \in [N]\}$.

### 3.1 Sample covariance matrix

We start by reviewing the sample covariance estimator. For known noise power $N_0$ at the BS, the sample covariance matrix is given by

$$
\hat{\Sigma}_{h} = \frac{1}{N}\sum_{s=1}^{N} y[s]y[s]^H - N_0I_M. \quad (5)
$$

---

1. Note that we use the standard spherical coordinate system (see, e.g., [16, Section 7.1]), where $\theta$ is the azimuth angle and limited by $\theta \in [-\theta_{\max}, \theta_{\max}]$.

2. As a note of caution, we hasten to say that the model must be reconsidered in the case of "extra-large aperture" arrays (e.g., see [39]), where the Toeplitz form does not hold any longer.

3. For finite $N$, the matrix $\hat{\Sigma}_{h}$ in (5) may not be positive semi-definite (PSD). In this case, an estimator would project $\hat{\Sigma}_{h}$ on the cone of PSD Hermitian matrices by setting to zero the negative eigenvalues. We neglect this detail here since it is irrelevant for the motivating argument made in this section. It is important to remark that all estimators considered in this work do produce PSD matrices.
This is a consistent estimator, in the sense that it converges to the true covariance matrix as \( N \rightarrow \infty \) [40, Section 1.2.2]. The mean square (Frobenius norm) error incurred by the sample covariance estimator is given as [40] 

\[
\mathbb{E} \left[ \left\| \hat{\Sigma}_h - \Sigma_h \right\|_F^2 \right] = \frac{\text{tr}(\hat{\Sigma}_h)\Sigma_1}{N}.
\]

By applying the Cauchy–Schwarz inequality to the singular values of \( \Sigma_h \), it is seen that 

\[
\text{tr}((\Sigma_h) \leq \| \Sigma_h \|_F \sqrt{\text{rank}(\Sigma_h)},
\]

which together with the estimation error expression yields the upper bound to the normalized mean squared error 

\[
\mathbb{E} \left[ \frac{\left\| \hat{\Sigma}_h - \Sigma_h \right\|_F^2}{\| \Sigma_h \|_F^2} \right] \leq \frac{\text{rank}(\Sigma_h)}{N}.
\]

As already discussed, a relevant and interesting regime for massive MIMO is when \( N \) and \( M \) are of the same order. From the above analysis, it is clear that the sample covariance estimator yields a small error if \( \text{rank}(\Sigma_h) \ll N \). For example, if the scattering contains only a finite number of discrete components (e.g., as assumed in [33–35]), \( \text{rank}(\Sigma_h) \) is small even if \( M \) is very large. In contrast, if \( \gamma(\xi) \) contains a diffuse scattering component, i.e., if its cumulative distribution function \( \Gamma(\xi) = \int_1^\xi \gamma(\nu) d\nu \) is piecewise continuous with strictly monotonically increasing segments, then \( \text{rank}(\Sigma_h) \) increases linearly with \( M \) (see [10]) and the error incurred by the sample covariance estimator may be large. On the other hand, the presence of discrete scattering components implies that \( \gamma(\xi) \) contains Dirac delta functions (spikes) and therefore it is not squared-integrable. This poses significant problems for estimation methods that assume \( \gamma(\xi) \) to be an element in a Hilbert space of functions (e.g., the method proposed in [21]). The challenge tackled in this work is to devise an estimator which is able to handle both the small sample regime \( N/M \leq 1 \) and the presence of discrete and diffuse scattering.

### 3.2 Structure of the channel covariance matrix

As said, the ASF describes how the received signal power is distributed over the AoA domain. The signal from the UE to the BS array propagates through a given scattering environment. The line-of-sight (LoS) path (if present), specular reflections, and wedge diffraction occupy extremely narrow angular intervals. This is usually modeled in a large number of papers as the superposition of discrete separable angular components coming at normalized AoAs \( \{\phi_i\} \). In particular, it is assumed that the general form (1) reduces to the discrete sum of \( r \) paths \( h[s] = \sum_{i=1}^r \rho_i[s]a(\phi_i) \), with corresponding ASF \( \gamma(\xi) = \sum_{i=1}^r c_i \delta(\xi - \phi_i) \) and covariance matrix \( \Sigma_h = \sum_{i=1}^r c_i a(\phi_i)a(\phi_i)^H \), where \( c_i = \mathbb{E}[|\rho_i[s]|^2] \). However, it is well known from channel sounding observations (e.g., see [41]) and widely treated theoretically (e.g., see [38]) that diffuse scattering is typically also present and may carry a very significant part of the received signal power especially at frequencies below 6GHz. In this case, scattering clusters span continuous intervals over the AoA domain. In order to encompass full generality, we model the ASF \( \gamma(\xi) \) as a mixed-type distribution [42, Section 5.3] including discrete and diffuse scattering components:

\[
\gamma(\xi) = \gamma_d(\xi) + \gamma_c(\xi) = \sum_{i=1}^r c_i \delta(\xi - \phi_i) + \gamma_c(\xi),
\]

(6)

where \( \gamma_d(\xi) \) models the power received from \( r \ll M \) discrete paths and \( \gamma_c(\xi) \) models the power coming from diffuse scattering clusters. Since the ASF can be seen as a (generalized) density function, we borrow the language of discrete and continuous random
variables and refer to $\gamma_d(\xi)$ and to $\gamma_c(\xi)$ as the discrete and the continuous parts of the ASF, respectively.\textsuperscript{7} This corresponds to a so-called spiked model in the language of asymptotic random matrix theory (e.g., see [28]), where the spikes are the discrete scattering components.

Plugging (6) into (3), we obtain a corresponding decomposition of the channel covariance matrix as

$$
\Sigma_h = \Sigma_h^d + \Sigma_h^c = \sum_{i=1}^{r} c_i \mathbf{a}(\phi_i)\mathbf{a}(\phi_i)^H + \int_{-1}^{1} \gamma_c(\xi)\mathbf{a}(\xi)\mathbf{a}(\xi)^H d\xi.
$$

(7)

Notice that, in a typical massive MIMO scenario, $\text{rank}(\Sigma_h^d) = r$ is much smaller than $M$.

4 Dictionary-based parametric representation and covariance estimation

An outline of the steps taken by the proposed method is given in the following:

i) **Spike Location Estimation for $\gamma_d(\xi)$**: We apply the MUSIC algorithm [36] to estimate the AoAs of the spike components, i.e., the angles $\{\phi_i\}_{i=1}^r$ in (6), from the $N$ noisy samples $\{y[s]: s \in [N]\}$. We let $\{\hat{\phi}_i\}_{i=1}^\hat{r}$ denote the estimated AoAs, where also the number of spikes $\hat{r}$ is estimated (not assumed known). This is detailed in Sect. 4.1. For complete estimation of the discrete part $\Sigma_h^d$, the weights $\{c_i\}_{i=1}^{r}$ need to be further recovered. This is done jointly with the coefficients of the continuous part, as elaborated in the next step.

ii) **Dictionary-based Method for Joint Estimation of $\gamma_d(\xi)$ and $\gamma_c(\xi)$**: We assume that the ASF continuous part can be written as

$$
\gamma_c(\xi) = \sum_{i=1}^{G} b_i \psi_i(\xi),
$$

(8)

where $\mathcal{G}_c = \{\psi_i(\xi): i \in [G]\}$ is a suitable dictionary of nonnegative density functions (not containing spikes). Figure 2 shows an example of Dirac delta and Gaussian dictionaries. Now, the goal is to estimate the model parameters $\{c_i\}_{i=1}^{\hat{r}}$ and $\{b_i\}_{i=1}^{G}$ assuming the form of ASF $\gamma(\xi) = \sum_{i=1}^{\hat{r}} c_i \delta(\xi - \hat{\phi}_i) + \sum_{i=1}^{G} b_i \psi_i(\xi)$. The parameters estimation can be formulated as a constrained least-squares problem, as detailed in Sect. 4.2. In particular, if the functions in $\mathcal{G}_c$ have disjoint supports (i.e., the nonzero parts of different functions are not overlapping, e.g., the Dirac delta functions in Fig. 2a), we obtain a nonnegative least-squares (NNLS) problem, while if the functions in $\mathcal{G}_c$ have overlapping

\textsuperscript{7}As for probability density functions, $\gamma_c(\xi)$ needs not be continuous, but its cumulative distribution function $\Gamma_c(\xi) = \int_{-1}^{\xi} \gamma_c(\nu) d\nu$ is a continuous function.
supports (e.g., the Gaussian functions in Fig. 2b), we obtain a quadratic programming (QP) problem. As an alternative, we can use ML estimation. The log-likelihood function is a difference of concave functions in the model parameters and can be maximized using majorization–minimization (MM) methods. However, general MM approaches are prohibitively computationally complex for typical values of $M$ arising in massive MIMO. It turns out that when $G_c$ is formed by Dirac deltas on a discrete grid, the likelihood function maximization can be obtained through EM with much lower complexity. Since in general EM is guaranteed to converge to local optima, the initialization plays an important role. We propose to use the result of the (low-complexity) NNLS estimator as initial point for the EM iteration. The resulting method is detailed in Sect. 4.3.

iii) From ASF to Covariance Estimation: Finally, having estimated $\gamma_d(\xi)$ and $\gamma_c(\xi)$, we estimate the covariance $\Sigma_h$ via (7). In particular, since $\gamma(\xi)$ depends only on the scattering geometry and it is invariant with frequency, the mapping $\gamma(\xi) \rightarrow \Sigma_h$ defined by (7) can be applied for different carrier frequencies by changing the wavelength parameter $\lambda_0$ in the expression of the array response vector $a(\xi)$. Specifically, replacing $a(\xi)$ in (7) with $a(\nu \xi)$ where $\nu$ is a wavelength expansion/contraction coefficient, we obtain an estimator at the carrier frequency $f_v = \nu f_0$, where $f_0 = c_0/\lambda_0$ is the UL carrier frequency and $c_0$ denotes the speed of light. The resulting covariance estimator is given by

$$
\hat{\Sigma}_h^{(v)} = \sum_{i=1}^{G+\hat{r}} u_i^* S_i^{(v)}, \quad (9)
$$

where $\{u_i^*: i \in [G+\hat{r}]\}$ are the estimated model parameters, and where we define $S_i^{(v)} = \int_{-\frac{1}{2}}^{1} \psi_i(\xi) a(\nu \xi) a^H(\nu \xi) d\xi$ for $i \in [G]$ and $S_i^{(v)} = a(\nu \phi_i) a^{H}(\nu \phi_i)$ for $i \in [\hat{r}]$. In particular, for the DL carrier we have $\nu > 1$ since in typical cellular systems the DL carrier frequency is higher than the UL carrier frequency.

4.1 Discrete ASF support estimation

In this part, we first estimate the model order $r$ by applying the well-known minimum description length (MDL) principle [44]. Then, we use the MUSIC method (e.g., see [45, 46] and references therein), to estimate the locations $\{\phi_i\}_{i=1}^{r}$ of the spikes in discrete part of ASF $\gamma_d(\xi)$. Given the noisy samples $Y = \{y[s]\}_{s=1}^{N}$, the sample covariance of $Y$ is given as

$$
\hat{\Sigma}_y = \frac{1}{N} \sum_{s=1}^{N} y[s] y[s]^H. \quad (10)
$$

Let $\hat{\Sigma}_y = \hat{U} \hat{\Lambda} \hat{U}^H$ be the eigendecomposition of $\Sigma_y$, where $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{M,M})$ denotes the diagonal matrix consisting of the eigenvalues of $\Sigma_y$. Without loss of generality, we assume that the eigenvalues are ordered as $\hat{\lambda}_{1,M} \geq \cdots \geq \hat{\lambda}_{M,M}$.  

---

8 This statement holds over not too large frequency ranges, where the scattering properties of materials are virtually frequency independent. For example, this property holds for the UL and DL carrier frequencies of the same FDD system; however, it does not generally hold over much larger frequency ranges. For example, the ASFs of the same environment at (say) 3.5 GHz and at 28 GHz are definitely different, although quite related [43].
4.1.1 Number of spikes estimation using MDL

First, we wish to estimate the model order \( r \). We adopt the classical MDL method provided in [47], which was designed for estimating the number of sources impinging on a passive array of sensors under white Gaussian noise with unknown noise power. Specifically, assume that the covariance matrix of the observation has the form

\[
\begin{align*}
\Sigma_y &= \Sigma_h + \sigma^2 I_M, \\
\end{align*}
\]

where \( \Sigma_h \) is the rank-\( r \) covariance matrix in (7) and \( \sigma^2 \) is an unknown parameter.\(^9\) We consider the following family of covariance matrices

\[
\begin{align*}
\tilde{\Sigma}_y^{(k)} &= \Sigma^{(k)} + \sigma^2 I_M, \\
\end{align*}
\]

where \( \Sigma^{(k)} \) denotes a semi-positive matrix of rank \( k \) and \( k \in \{0, 1, \ldots, M-1\} \) is the model order. \( \tilde{\Sigma}_y^{(k)} \) can be expressed in its spectral form as

\[
\tilde{\Sigma}_y^{(k)} = \sum_{i=1}^{k} (\tilde{\lambda}_i - \sigma^2) \tilde{u}_i \tilde{u}_i^H + \sigma^2 I_M,
\]

where \( \{\tilde{\lambda}_i\} \) and \( \{\tilde{u}_i\} \) are the eigenvalues and eigenvectors of \( \tilde{\Sigma}_y^{(k)} \), respectively. Then, we denote the parameter vector of the model by \( \Theta^{(k)} = [\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k, \sigma^2, \tilde{u}_1^T, \ldots, \tilde{u}_k^T]^T \). Using the parameter vector \( \Theta^{(k)} \), we can calculate the joint probability density of the noisy samples \( Y \), denoted as \( f(Y|\Theta^{(k)}) \). Denoting \( L(\Theta^{(k)}) = -\log f(Y|\Theta^{(k)}) \) as the minus log-likelihood function of \( Y \), we can calculate the maximum-likelihood estimate

\[
\hat{\Theta}^{(k)} = \arg\min_{\Theta^{(k)}} L(\Theta^{(k)}).
\]

Following similar results in [47] and [48], we have

\[
L(\hat{\Theta}^{(k)}) = N(M-k) \log \left( \frac{a(k)}{b(k)} \right),
\]

where\(^{10}\)

\[
a(k) = \frac{1}{M-k} \sum_{i=k+1}^{M} \tilde{\lambda}_{i,M}, \quad b(k) = \begin{cases} 
1, & k=0 \\
\left( \prod_{i=1}^{k} \tilde{\lambda}_{i,M} \right)^{-\frac{1}{M-k}}, & k > 0.
\end{cases}
\]

The MDL approach selects a model order \( k \in \{0, 1, \ldots, M-1\} \) from a parameterized family of probability densities \( f(Y|\Theta^{(k)}) \) that minimize the so-called total description length of the samples, which is tightly approximated by Rissanen bound [49] (after neglecting \( o(1) \) terms) as

\(^9\) Note that \( \Sigma_y \neq E[y|y(k)]^{(k)} \) is not the true covariance of samples, since the contribution of diffuse clusters is not considered. Nevertheless, in Remark 1 we argue that the MDL method in [47] can be applied to our problem without significant degradation by simply ignoring the presence of the diffuse component.

\(^{10}\) Note that our expression of \( b(k) \) is different from that in [47]. However, it can be easily shown that the resulting MDL expressions differ only by a constant \( M \), which has no influence in the minimization in (19). Here, \( b(k) \) has been modified to avoid the product of small (nearly zero when sample size is quite small) eigenvalues of \( \Sigma_y \).
where the first term is the result of the maximum likelihood estimator obtained in (15) and the second term is a penalty function including $|\Theta_1^{(k)}| = k(2M - k)$ being the number of free parameters in $\Theta^{(k)}$ [47, 48]. The number of spikes is estimated by minimizing the MDL metric

\[
\text{MDL}(k) = L\left(\hat{\Theta}^{(k)}\right) + \frac{1}{2}k(2M - k) \log(N),
\]

(17)

Remark 1 It is noticed that the proposed MDL approach does not explicitly consider the diffuse part. It considers only a spike signal subspace with dimension $r$ corresponding to the $r$ largest eigenvalues of $\hat{\Sigma}_y$ and a white noise with power $\sigma^2$ resulting in $M - r$ smallest eigenvalues of $\hat{\Sigma}_y$ equaling to $\sigma^2$. The diffuse component in our model can be considered as an additional spatially colored noise as far as the spike estimation is concerned. It is shown in [50] that in the presence of colored Gaussian noise, MDL tends to overestimate the model order with increasing number of samples $N$. However, we shall see later that overcounting the spikes is not catastrophic in the overall covariance estimation scheme, since the coefficients of fictitious spikes are typically estimated as near zero in the model coefficient estimation step. In other words, it is always better to overestimate the number of spikes than to underestimate them so that the true spikes are not missed. In [51], it is shown that undermodeling may happen when the signal and the noise eigenvalues are not well separated and the noise eigenvalues are clustered sufficiently closely. We show that the undermodeling is not likely to happen in our case by showing that the gap between $r$ large (containing the contribution of the spikes) and $M - r$ small eigenvalues (containing only the contribution of the diffuse scattering and the noise) of the sample covariance $\hat{\Sigma}_y$ becomes larger and larger as the number of antennas $M$ increases. This can be explained by noticing that as the array angular resolution increases, the amount of received signal energy in each angular bin decreases with the bin width in the bins that contain no spikes, while it remains roughly constant with $M$ if the angular bin contains a spike. More precisely, an angular bin of width $2/M$ and centered at $\xi$ contains a received signal power proportional to $2(\gamma_c(\xi) + N_0)/M$ if no spike falls in the interval, and to $c_i + 2(\gamma_c(\xi) + N_0)/M$ if the spike located at $\phi_i$ falls in the interval. By Szegö’s theorem (e.g., see [10] and references therein), the eigenvalues of the covariance matrix $\Sigma_y = \Sigma_h + N_0 I_M$ converge asymptotically as $M \rightarrow \infty$ to the energy received on equally spaced angular bins of width $2/M$ over the interval $[-1, 1]$ in the $\xi$ domain. Figure 3 corroborates this showing the separation between the eigenvalues of the sample covariance matrix $\hat{\Sigma}_y$ for different number of antennas $M = 25, 50, 100$ with fixed sample size to channel dimension ratio $N/M = 2$ and the same channel geometry defined by the ASF

Remark 1
where \( \text{rect}_A \) is 1 over the interval \( A \) and zero elsewhere. This ASF contains \( r = 2 \) spikes and two rectangular-shaped diffuse components. The SNR is set to 20 dB. For a large enough number of antennas (and even for a moderate number such as \( M = 25 \)), the eigenvalue distribution shows a significant jump, such that the two largest eigenvalues “escape” from the rest. Note that, by increasing the number of antennas, this separation becomes more and more significant. Hence, the undermodeling of MDL in the relevant case of massive MIMO is unlikely to occur.

### 4.1.2 Location of spikes estimation using MUSIC

Once the number of spikes is estimated using MDL, MUSIC proceeds to identify the locations of those spikes. Let \( \hat{\mathbf{u}}_{\tilde{r}+1:M}, \ldots, \hat{\mathbf{u}}_{MM} \) be the eigenvectors in \( \hat{\mathbf{U}} \) corresponding to the smallest \( M - \tilde{r} \) eigenvalues, and let us define \( \mathbf{U}_{\text{noi}} = [\hat{\mathbf{u}}_{\tilde{r}+1:M}, \ldots, \hat{\mathbf{u}}_{MM}] \) as the \( M \times (M - \tilde{r}) \) matrix corresponding to the noise subspace. The MUSIC objective function is defined as the pseudo-spectrum:

\[
\hat{\gamma}(\xi) = \left\| \mathbf{U}_{\text{noi}}^H \mathbf{a}(\xi) \right\|^2 = \sum_{k=\tilde{r}+1}^{M} \left| \mathbf{a}(\xi)^H \hat{\mathbf{u}}_{k:M} \right|^2.
\]

(21)

MUSIC estimates the support \( \{\hat{\phi}_1, \ldots, \hat{\phi}_r\} \) of the spikes by identifying \( \tilde{r} \) dominant minimizers of \( \hat{\gamma}_M(\xi) \). It can be shown that for a finite number of spikes \( r \), as the number of antennas \( M \) and the number of samples \( N \) grow to infinity with fixed ratio, MUSIC yields an asymptotically consistent estimate of the spike AoAs. The details are given in Appendix A for the sake of completeness. Figure 4 illustrates the normalized values of the pseudo-spectrum (21) for the example ASF in (20) and its corresponding sample covariance \( \hat{\mathbf{S}}_y \) for \( M = 25 \) and \( N/M = 2 \). As we can see, the \( r = 2 \) smallest minima of the pseudo-spectrum occur very close to the points \( \phi_1 = -0.2 \) and \( \phi_2 = 0.4 \), which are the locations of the spikes in the true ASF.

\[
\gamma(\xi) = \text{rect}_{[-0.7, -0.4]} + \text{rect}_{[0, 0.6]} + (\delta(\xi + 0.2) + \delta(\xi - 0.4))/2,
\]

(20)
4.2 Coefficients estimation by constrained least-squares

After obtaining an estimate of the number of spikes and their locations as described before, we need to find the spike coefficients $c = [c_1, \ldots, c_{\hat{r}}]^T \in \mathbb{R}^{\hat{r}}$ and the coefficients $b = [b_1, \ldots, b_G]^T \in \mathbb{R}^G$ of the continuous ASF component $\gamma_c(\xi)$ in the form (8). The dictionary functions $\{\psi_i(\xi)\}$ are selected according to the available prior knowledge about the propagation environment. Some typical choices include localized functions $\psi_i(\xi)$, such as Gaussian, Laplacian, or rectangular functions, with a suitably chosen support. The above representation of the ASF results in the parametric form of the channel covariance given by

$$\Sigma_h(u) = \sum_{i=1}^{G+\hat{r}} u_i S_i,$$  \quad (22)

where we define the model parameter vector $u = [u_1, \ldots, u_{G+\hat{r}}]^T = [b^T, c^T]^T$ and the positive semi-definite Hermitian symmetric matrices $S_i = \int_{-1}^{1} \psi_i(\xi) a(\xi) a(\xi)^H d\xi$, $\forall i \in [G]$, and $S_i = a(\hat{\phi}_i) a(\hat{\phi}_i)^H$, $\forall i \in [\hat{r}]$.

In order to estimate the coefficients vector $u$ from the noisy samples $\{y[s] : s \in [N]\}$, we propose three algorithms, namely NNLS estimator, QP estimator, and ML-EM estimator.

4.2.1 NNLS estimator

If the dictionary functions have disjoint support, since the overall $\gamma_c(\xi)$ must be non-negative, it follows that the coefficients $\{u_i : i \in [G]\}$ must take values in $\mathbb{R}_+$, i.e., the whole vector $u$ is nonnegative. We know that if the number of samples $N$ is large enough, the sample covariance matrix $\hat{\Sigma}_y$ would converge to $\Sigma_y = \Sigma_h + N_0 I_M$. Then, our goal is to find a good fitting to the sample covariance matrix $\hat{\Sigma}_y$ from the set of all covariance matrices of the form

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig4}
\caption{The pseudo-spectrum plotted for the example ASF in (20) with $M = 25$ and $N/M = 2$.}
\end{figure}
\[ \Sigma_{Y} = \Sigma_h(u) + N_0 I_M = \sum_{i=1}^{G+\hat{\tau}} u_i S_i + N_0 I_M. \]  

(23)

For this purpose, we use the Frobenius norm as a fitting metric and obtain an estimate of the model coefficients as

\[ u^* = \arg \min_{u \in \mathbb{R}^{G+\hat{\tau}}} \| \Sigma_{Y} - \sum_{i=1}^{G+\hat{\tau}} u_i S_i - N_0 I_M \|_F^2 = \arg \min_{u \in \mathbb{R}^{G+\hat{\tau}}} \| \hat{\Sigma}_h - \sum_{i=1}^{G+\hat{\tau}} u_i S_i \|_F^2. \]  

(24)

Applying vectorization and defining \( A = [\text{vec}(S_1), \ldots, \text{vec}(S_{G+\hat{\tau}})] \) and \( f = \text{vec}(\hat{\Sigma}_h) \), we can write this as a NNLS problem

\[ u^* = \arg \min_{u \in \mathbb{R}^{G+\hat{\tau}}} \| Au - f \|_2^2, \]  

(25)

which can be efficiently solved using a variety of convex optimization techniques (see, e.g., [52, 53]). In our simulation, we use the built-in MATLAB function \textit{lsqlin}.

Note that a ULA covariance is Hermitian Toeplitz. By leveraging this, we can reformulate the optimization problem so that the problem dimension is reduced. Concretely, let \( \tilde{\Sigma}_h \) denote the orthogonal projection of \( \hat{\Sigma}_h \) onto the space of Hermitian Toeplitz matrices. This is obtained by averaging the diagonals of \( \hat{\Sigma}_h \) and replacing the diagonal elements by the corresponding average value (see Appendix B for completeness). We define the first column of \( \tilde{\Sigma}_h \) as \( \tilde{\sigma} \). Then, (25) can be reformulated as

\[ u^* = \arg \min_{u \in \mathbb{R}^{G+\hat{\tau}}} \| W (\tilde{A} u - \tilde{\sigma}) \|_2^2, \]  

(26)

where \( \tilde{A} = [(S_1)_{-1}, \ldots, (S_{G+\hat{\tau}})_{-1}] \) is the matrix collecting the first columns \((S_i)_{-1}\) of the \( S_i \) and \( W = \text{diag}\left( \left[ \sqrt{M}, \sqrt{2(M-1)}, \sqrt{2(M-2)}, \ldots, \sqrt{2} \right]^T \right) \) is the weighting matrix to compensate for the number of times an element is repeated in a Hermitian Toeplitz matrix.

**Lemma 2**  
The optimization problems in (25) and (26) are equivalent.

**Proof**  
Note that the difference between (25) and (26) is only the replacement of \( \tilde{\Sigma}_h \) from \( \hat{\Sigma}_h \). Thus, it is sufficient to show the equivalence of the following two optimization problems:

\[ P1: \min_{X \in \mathcal{H}T} \| X - \tilde{\Sigma}_h \|_F^2, \quad P2: \min_{X \in \mathcal{H}T} \| X - \hat{\Sigma}_h \|_F^2, \]  

(27)

where \( \mathcal{H}T \) is the set of Hermitian Toeplitz matrices. Let \( \Delta = \tilde{\Sigma}_h - \hat{\Sigma}_h \). Then, P1 is rewritten as \( \min_{X \in \mathcal{H}T} \| X - \tilde{\Sigma}_h - \Delta \|_F^2 \). Since \( \tilde{\Sigma}_h \) is the orthogonal projection of \( \hat{\Sigma}_h \) on the set \( \mathcal{H}T \), by the orthogonality principle the difference \( \Delta = \hat{\Sigma}_h - \tilde{\Sigma}_h \) is orthogonal to the whole set, i.e., \( \langle \Delta, T \rangle := \text{tr}(\Delta^H T) = 0, \forall T \in \mathcal{H}T \). It follows that for any \( X \in \mathcal{H}T \) the difference \( X - \tilde{\Sigma}_h \) is also in \( \mathcal{H}T \). Thus, \( \| X - \tilde{\Sigma}_h - \Delta \|_F^2 = \| X - \tilde{\Sigma}_h \|_F^2 + \| \Delta \|_F^2 \) where \( \| \Delta \|_F^2 \)
is constant with respect to $X$ and therefore plays no role in minimization. This proves the equivalence of P1 and P2.

\[ \sum_{i=1}^{G} b_i \psi_i(\xi_j) \geq 0, \quad \forall j \in [G], \quad \iff \quad \Psi b \geq 0, \quad (28) \]

where the elements of the matrix $\Psi \in \mathbb{R}^{G \times G}$ are obtained as
\[ (\Psi)_{ij} = \psi_i(\xi_j), \quad \forall i \in [G], j \in [G]. \]

Then, the estimation problem under dictionary functions with overlapping support is given by
\[
\begin{align*}
\text{minimize} & \quad \|W(\tilde{A}u - \tilde{\sigma})\|, \quad \text{s.t.} \quad \Psi b \geq 0,
\end{align*}
\[ (29) \]

where $u = [b^T, c^T]^T$ is consistently with the definition in (22). Notice that (29) is a QP problem and can be solved using standard QP solvers, such as quadprog in MATLAB.

### 4.3 Coefficients estimation by maximum likelihood

Instead of directly fitting the Frobenius norm between the sample covariance and the parametric covariance, the model parameters can be estimate by the ML method. Give the matrix of the observed noisy channel samples $Y$, the likelihood function of $Y$ assuming $\Sigma_h = \Sigma_h(u)$ in the form of (22) is given by

\[ p(Y|u) = \prod_{s=1}^{N} p(y[s]|u) = \prod_{s=1}^{N} \frac{\exp\left(-y[s]^H(\Sigma_h(u) + N_0I_M)^{-1}y[s]\right)}{\pi^M\det(\Sigma_h(u) + N_0I_M)} \]
\[ = \frac{\exp\left(-\text{tr}\left((\Sigma_h(u) + N_0I_M)^{-1}YY^H\right)\right)}{\pi^{MN}\det(\Sigma_h(u) + N_0I_M)^N}. \quad (30) \]

Using (30), we can form the minus log-likelihood function $f_{ML}(u) := -\frac{1}{N} \log p(Y|u)$. Then, the ML-based covariance estimator is obtained by minimizing $f_{ML}(u)$ with respect to the real and nonnegative coefficients vector $u$, which is formulated as the optimization problem:

\[
\begin{align*}
\text{minimize}_{u \in \mathbb{R}^{G+r}^+} & \quad f_{ML}(u) = \log \det \left( \sum_{i=1}^{G+r} u_i S_i + N_0I_M \right) + \text{tr} \left( \left( \sum_{i=1}^{G+r} u_i S_i + N_0I_M \right)^{-1} \hat{\Sigma}_y \right),
\end{align*}
\[ (31) \]
where $\hat{\Sigma}_y$ is the sample covariance matrix of the observations defined in (10). Note that the objective function $f_{\text{ML}}(u) = f_{\text{cav}}(u) + f_{\text{ex}}(u)$ in (31) is the sum of a concave and a convex function and thus (31) is not a convex problem.

It is generally difficult to find the global optimum of a non-convex function such as $f_{\text{ML}}(u)$. A standard approach in such cases is to adopt a MM algorithm [54, 55], alternating through two steps with an updating surrogate function that has favorable optimization properties (e.g., convexity) and approximates the upper bound of the original objective function. Typical examples of MM algorithms are EM method [56], cyclic minimization [57], and the concave–convex procedure [58]. We choose the EM algorithm to iteratively find a good stationary point of $f_{\text{ML}}(u)$ as we will see that this algorithm yields a computationally efficient update rule and excellent empirical results for the task of estimating the parametric ASF coefficients. Note that although the likelihood function in (31) is in a general form for any family of dictionary functions, the EM method can be applied only in the case where all the matrices $S_i$ have rank 1, which is the case when the dictionary functions $\psi_i(\xi)$ are Dirac delta functions. In contrast, the more general concave–convex procedure (e.g., see [1] for the application in this case) can deal with any type of dictionary, but yields significantly higher computational complexity, so that it is not suited for large $M$ (massive MIMO case). In this work, we deal with general dictionary functions using constrained LS, while restrict the use of EM to the case of Dirac delta dictionary functions.

**Application of the EM algorithm.** When $\psi_i(\xi) = \delta(\xi - \xi_i)$ where $\{\xi_i : i \in [G]\}$ are uniformly spaced points in $[-1, 1]$, we have $S_i = a(\xi_i) a_H(\xi_i)$ for $i \in [G]$ and $S_{G+i} = a(\hat{\xi}_i) a_H(\hat{\xi}_i)$ for $i \in [\tilde{F}]$. Then, defining the extended grid $\{\hat{\xi}_i : i \in [G + \tilde{F}]\}$ with $\xi_{G+i} = \hat{\xi}_i$ for $i \in [\tilde{F}]$, the parametric form of channel covariance $\Sigma_h$ can be written as

$$
\Sigma_h(u) = \sum_{i=1}^{G+\tilde{F}} u_i S_i = DUD_H,
$$

where $D = [a(\xi_1), \ldots, a(\xi_{G+\tilde{F}})]$ and $U = \text{diag}(u)$. Then, $\Sigma_y = \Sigma_h + N_0 I_M$ can be formally considered as the covariance matrix of the approximated received samples

$$
Y[s] = \sum_{i=1}^{G+\tilde{F}} \rho_i [s] a(\xi_i) + z[s] = Dx[s] + z[s],
$$

where $x[s] = [\rho_1[s], \ldots, \rho_{G+\tilde{F}}[s]]^T$ is the latent variable vector containing the instantaneous random path gain, and we have $x[s] \sim \mathcal{CN}(\mathbf{0}, U), \forall s \in [N]$ where $u \in \mathbb{R}_{+}^{G+\tilde{F}}$ is the vector of component-wise variances of Gaussian vectors $\{x[s] : s = 1, \ldots, N\}$.

Under the formulation in (33), the $N$ noisy samples collected as columns of the matrix $Y$ can be written as $Y = DX + Z$, where $X = [x[1], \ldots, x[N]]$ and $Z = [z[1], \ldots, z[N]]$. The EM algorithm treats $(Y, X)$ as the *incomplete data*, where $X$ is referred to as *missing data*. Using the fact that $Y$ given $X$ is Gaussian with mean $DX$ and independent components with variance $N_0$, we have that $p(Y, X|u) = p(Y|X)p(X|u)$, where $p(Y|X)$ is a conditional Gaussian distribution that does not depend on $u$. By marginalizing with respect to $X$ and taking the logarithm, the log-likelihood function takes on the form of a conditional expectation $\mathcal{L}(u) := \log p(Y|u) = \log \mathbb{E}_{X|u}[p(Y|X)]$. The EM algorithm maximizes iteratively a lower bound on $\mathcal{L}(u)$ by alternating the expectation step (E-step) and the maximization
step (M-step) \[55\]. Let \( \tilde{u}^{(i)} \) be the estimate of \( u \) in the \( \ell \)-th iteration, by introducing a posterior density of \( X \) as \( p(X|Y, \tilde{u}^{(i)}) \) we have

\[
\mathcal{L}(u) = \log \mathbb{E}_{X|u} \left[ \frac{p(Y|X)p(X|Y, \tilde{u}^{(i)})}{p(X|Y, \tilde{u}^{(i)})} \right] = \log \mathbb{E}_{X,Y|\tilde{u}^{(i)}} \left[ \frac{p(Y|X)p(X|u)}{p(X|Y, \tilde{u}^{(i)})} \right]
\]

(a) \[34\]

\[
\geq \mathbb{E}_{X,Y|\tilde{u}^{(i)}} [\log p(Y|X) + \log p(X|u) - \log p(X|Y, \tilde{u}^{(i)})] := \tilde{\mathcal{L}}(u|\tilde{u}^{(i)}),
\]

where (a) follows Jensen’s inequality and the concavity of \( \log(\cdot) \).

The E-step consists of computing \( \tilde{\mathcal{L}}(u|\tilde{u}^{(i)}) \). Using the joint conditional Gaussianity of \( Y \) and \( X \) given \( u = \tilde{u}^{(i)} \), \( \tilde{\mathcal{L}}(u|\tilde{u}^{(i)}) \) can be evaluated in closed form by computing the conditional mean and covariance of \( x|s \) given \( y|s \) and \( \tilde{u}^{(i)} \), respectively, given by \[59\]

\[
\text{E-step: } \mu_{x|s}^{(i)} = \frac{1}{N_0} \Sigma_{x}^{(i)} \mathbf{D}^t y|s, \quad \Sigma_{x}^{(i)} = \left( \frac{1}{N_0} \mathbf{D}^t \mathbf{D} + \left( \tilde{U}^{(i)} \right)^{-1} \right)^{-1},
\]

where we define \( \tilde{U}^{(i)} = \text{diag}(\tilde{u}^{(i)}) \).

The M-step consists of the maximization:

\[
\tilde{u}^{(i+1)} = \arg \max_{u \in \mathbb{R}_+^G} \tilde{\mathcal{L}}(u|\tilde{u}^{(i)}).
\]

Note that \( p(Y|X) \) and \( p(X|Y, \tilde{u}^{(i)}) \) in \[35\] do not depend on \( u \) and thus can be neglected in the M-step. Hence, the function to be maximized in the M-step can be equivalently written as (details are omitted for brevity)

\[
\mathbb{E}_{X,Y|\tilde{u}^{(i)}} [\log p(X|u)] = \sum_{i=1}^{G+F} \left( -N \log(\pi u_i) - \frac{\sum_{s=1}^N \left[ \mu_{x|s}^{(i)} \right]^2 + N \left[ \Sigma_{x}^{(i)} \right]_{i,i}}{u_i} \right).
\]

It is observed from \[38\] that the maximization is decoupled with respect to each component \( u_i \) of \( u \). Then, the optimality in the \( \ell \)-th iteration is also easily obtained in closed form. Setting each partial derivative \( \frac{\partial}{\partial u_i} \) of \[38\] to zero, we find

\[
\text{M-step: } \tilde{u}_i^{(i+1)} = \frac{1}{N} \sum_{s=1}^N \left[ \mu_{x|s}^{(i)} \right]_{i}^2 + \left[ \Sigma_{x}^{(i)} \right]_{i,i}, \quad \forall i \in [G+F].
\]

With a initial point \( \tilde{u}^{(0)} \), the ML-EM algorithm iteratively runs the E-step and M-step until the stop condition \( f_{\text{ML}}(\tilde{u}^{(\ell)}) - f_{\text{ML}}(\tilde{u}^{(\ell+1)}) \leq \epsilon_{\text{EM}} \) is met, where \( \epsilon_{\text{EM}} \) is the predefined stop threshold. The initial point is usually set as an all ones vector. In our case, it can be set as the result of NNLS solution. An extensive comparison of these two initializations obtained by simulating several different channel scattering geometries is depicted in Fig. 5, which shows that both initializations converge within 100 iterations. It reveals that the NNLS initialization converges much faster and results in lower values of the objective function. Therefore, in our results we used the NNLS solution to initialize ML-EM algorithm. Furthermore, we provide computational complexity analysis of the proposed NNLS and ML-EM algorithms in Appendix C.
Remark 3 The use of Dirac delta functions as dictionary functions to approximate the continuous part of the ASF $\gamma_c$ may sound as a contradiction, since by definition this component does not contain spikes. However, there is a fundamental difference between the Dirac components corresponding to spikes and the equally spaced “picket-fence” used to approximate $\gamma_c$. This can be noticed by observing that the power associated with the $i$-th spike component is $c_i$, which is a constant independent on the number of grid points $G$, while the power associated with the $i$-th dictionary function $\psi_i(\xi) = \delta(\xi - \xi_i)$ is $b_i = 2\gamma_c(\xi_i)/G$ where $2/G$ is the spacing of the uniform grid on $[-1, 1]$. For sufficiently large $G$, the smooth function $\gamma_c$ can be approximated by the picket-fence of scaled Dirac deltas in the sense that, for any sufficiently smooth test function $f(\xi)$ ($L^1$-integrable on $[-1, 1]$, piecewise continuous, with at most a countably infinite number of discontinuities) we have $\lim_{G \to \infty} \int_{-1}^{1} (\gamma_c(\xi) - \frac{2}{G} \sum_{i=1}^{G} \gamma_c(\xi_i) \delta(\xi - \xi_i)) f(\xi) d\xi = 0$.

5 Results and discussion

In this section, the proposed constrained LS- and ML-EM-based estimators are numerically evaluated and compared with existing state-of-the-art methods. We use the realistic channel emulator QuaDriGa [18] to generate the channel samples. We adopt two communication scenarios in QuaDriGa: 3GPP 3D Urban Macro-Cell Line Of Sight (3GPP-3D-UMa-LOS) and 3GPP 3D Urban Macro-Cell Non-Line Of Sight (3GPP-3D-UMa-NLOS). The BS array adopts horizontal ULA with $M = 128$ antennas. The carrier frequency of the UL and DL is 1.9 GHz and 2.1 GHz, respectively. The SNR is set to 10 dB. The results are averaged over 20 random ASFs and 100 times of random channel realizations for each ASF.

5.1 Compared benchmarks

We compare the proposed algorithms to the following three benchmarks.

(1) Toeplitz–PSD Projection: The first benchmark is an intuitively simple approach. We know that the ULA channel covariance is a Toeplitz–PSD matrix. Thus, we can project
the sample covariance onto the space of Toeplitz–PSD by solving the convex optimization problem:

\[
\Sigma_h^{PSD} = \arg\min_{\Sigma \in HT} \| \Sigma - \bar{\Sigma}_h \|_F^2, \quad \text{s.t. } \Sigma \succeq 0. \tag{40}
\]

The projected matrix \( \Sigma_h^{PSD} \) is the covariance estimate. We use the projections onto convex sets (POCS) algorithm \([60]\) to solve (40). Specifically, the sample covariance matrix is alternately projected onto the convex set of Toeplitz and PSD cone. Projecting on Toeplitz follows the way in Appendix B, and projection on the PSD cone is achieved by eigendecomposition and setting all negative eigenvalues to zero. These two projections are repeated till convergence. Note that the complexity of this semi-definite programming problem can be high when the number of antennas \( M \) is large. Moreover, Toeplitz–PSD cannot provide UL-DL covariance transformation.

(2) The SPICE Method: The second method we use for comparison is known as sparse iterative covariance-based estimation (SPICE) \([33]\). This method also exploits the ASF domain but can only be applied with Dirac delta dictionaries. Similar to the parametric covariance model with only Dirac delta dictionaries introduced in (32), assuming Dirac delta dictionaries \( D = [a(\xi_1), \ldots, a(\xi_G)] \) of \( G \) array response vectors corresponding to \( G \) AoAs, and defining \( \Sigma = D\text{diag}(u)D^H \), the ASF coefficients \( u \) are estimated by solving the following convex optimization problem for two cases:

\[
u^* = \begin{cases} 
\arg\min_{u \in \mathbb{R}_{G+}^G} \| \Sigma^{-1/2} (\bar{\Sigma}_y - \Sigma) \|_F^2, & N < M, \\
\arg\min_{u \in \mathbb{R}_{G+}^G} \| \Sigma^{-1/2} (\bar{\Sigma}_y - \Sigma) \Sigma_y^{-1/2} \|_F^2, & N \geq M.
\end{cases} \tag{41}
\]

The channel covariance estimate is then obtained as \( \Sigma_h^{SPICE} = D\text{diag}(u^*)D^H \).

(3) Convex Projection Method: This method is proposed in \([21]\) for the ASF estimation by solving a convex feasibility problem \( \tilde{\gamma} = \text{find } \gamma, \text{subject to } \gamma \in \mathcal{S} \), where

\[
\mathcal{S} = \left\{ \gamma : \int_{-1}^{1} \gamma(\xi) e^{i\pi m\xi} d\xi = [\bar{\Sigma}_h]_{m,1}, \ m = [M], \ \gamma(\xi) \geq 0, \forall \xi \in [-1, 1] \right\}. \tag{42}
\]

This can be solved by applying an iterative projection algorithm, which produces a sequence of functions in \( L_2 \) that converges to a function satisfying the constraint \( \gamma \in \mathcal{S} \). Given the estimated ASF \( \tilde{\gamma} \), the channel covariance estimation is obtained following (3).

5.2 Considered metrics

Denoting a generic covariance estimate as \( \bar{\Sigma} \), we use three error metrics to evaluate the estimation quality:

(1) Normalized Frobenius norm Error: This error is defined as

\[
E_{NF} = \frac{\| \Sigma_h - \bar{\Sigma} \|_F}{\| \Sigma_h \|_F}. \tag{43}
\]
(2) **Normalized MSE of Channel Estimation:** Given a noisy channel observation \( y = h + z \), the optimal estimation of channel vector \( h \) is obtained via linear MMSE filter \( \hat{h} = \Sigma_1(N_0 \mathbf{I}_M + \hat{\Sigma})^{-1}y \). Then, this metric considers the normalized mean squared error (NMSE) of instantaneous channel estimation:

\[
E_{\text{NMSE}} = \frac{\mathbb{E}[\|h - \hat{h}\|^2]}{\mathbb{E}[\|h\|^2]}.
\]

By the optimality of linear MMSE estimation for Gaussian random vectors, the lower bound of this error is obtained using the true channel covariance.

(3) **Power Efficiency:** This metric evaluates the similarity of dominant subspaces between the estimated and true matrices, which is an important factor in various applications of massive MIMO such as user grouping and group-based beamforming [7, 10, 61]. Specifically, let \( p \in [M] \) denote a subspace dimension parameter and let \( U_p \in \mathbb{C}^{M \times p} \) and \( \hat{U}_p \in \mathbb{C}^{M \times p} \) be the \( p \) dominant eigenvectors of \( \Sigma_h \) and \( \hat{\Sigma} \) corresponding to their largest \( p \) eigenvalues, respectively. The power efficiency (PE) based on \( p \) is defined as

\[
E_{\text{PE}}(p) = 1 - \frac{\langle \Sigma_h, \hat{U}_p \hat{U}_p^H \rangle}{\langle \Sigma_h, U_p U_p^H \rangle} = \frac{\text{tr}(U_p^H \Sigma_h U_p) - \text{tr}(\hat{U}_p^H \Sigma_h \hat{U}_p)}{\text{tr}(U_p^H \Sigma_h U_p)}.
\]

It is noticed that \( E_{\text{PE}}(p) \in [0, 1] \) and the closer it is to 0, the more power is captured by the estimated \( p \)-dominant subspace.

### 5.3 Performance comparison

We first provide a performance comparison under only Dirac delta dictionaries with a relatively large number \( G \) of grid points. Then, we provide a comparison under different dictionaries to show the benefit of properly choosing the dictionary adapted to the case at hand.

(1) **Comparison under Dirac delta dictionaries:** We set the number of dictionary functions for the ASF continuous part as \( G = 2M \). SPICE is also applied to the same picket-fence dictionary without knowledge of the spike locations, since this is a feature of our own method and not intrinsic in the SPICE algorithm. The projection method is not a dictionary based method. To implement it, we discretize the ASF domain with 5000 grid points to approximate the continuous ASF.

In Fig. 6, the UL and DL channel covariance estimation error for scenario 3GPP-3D-UMa-LOS in terms of the normalized Frobenius norm error, normalized MSE of channel estimation, and power efficiency under different sample ratios \( N/M \) (from 0.0625 to 1) are depicted. It is observed that the results of the proposed NNLS and ML methods significantly outperform the other benchmarks in both UL and DL for all metrics under all range of sample size. Although the Toeplitz–PSD method has very similar Frobenius norm errors compared to our methods, it performs significantly worse than our methods in terms of channel estimation error especially under extremely small sample size. Moreover, it is worth to emphasize again that the Toeplitz–PSD method does not give directly an easy way to perform the UL-DL covariance transformation. In contrast (see
Fig. 6c, the proposed methods yield very good UL-DL transformation (according to the considered metrics) even under very small sample size. Furthermore, we also present the results of NNLS and ML-EM without MUSIC (i.e., without explicit spike location estimation) denoted as “NNLS, Delta, nM” and “ML-EM, Delta, nM” in Fig. 6a and b. It is observed that in this case the performances degrade dramatically. This indicates that the proposed MUSIC step is necessary and non-trivially improves the overall performance. It is also noticed that our results without MUSIC are still much better than the results of SPICE and projection method, which shows the advantage of the proposed NNLS and ML-EM algorithms.

Results for the NLOS scenario 3GPP-3D-UMa-NLOS are depicted in Fig. 7. First, we observe again that our methods outperform the other benchmarks. Interestingly, the results with MUSIC perform almost the same as the results without MUSIC. Notice that in this case the scheme is unaware of the fact that there are no spikes, and the MDL/MUSIC step may give some spurious spikes, which are then eliminated by the subsequent NNLS and ML/EM step (estimated coefficients near zero). This demonstrates the fact that the proposed method is robust to both LOS and NLOS cases.

(2) Comparison with Dirac delta and overlapping Gaussian dictionaries: In this part, we show the results based on Dirac delta and overlapping Gaussian dictionaries to indicate the importance of finding the proper dictionary. We set a fixed $N = 16$, i.e., $N/M = 0.125$. We test the NNLS algorithm with Dirac delta and Gaussian dictionaries as well as the QP estimator with Gaussian dictionaries with $\tilde{G} = 10000$ under different number of dictionaries ($G/M$ is from 0.0625 to 2) for the NLOS scenario 3GPP-3D-UMa-NLOS. The overlapping Gaussian dictionary functions are defined as follows. Given $G$, let $\tilde{\psi}(\xi)$ be a Gaussian density whose support is limited to $[-\frac{4}{G+3}, \frac{4}{G+3}]$. The dictionary $\{\psi_i(\xi) : i \in [G]\}$ consists of skewed shifted versions of $\tilde{\psi}(\xi)$, i.e., $\psi_i(\xi) = J(\xi)\tilde{\psi}(\xi + 1 - \frac{2(i+1)}{G+3})$, $i \in [G]$, where $J(\xi) = \frac{1}{\sqrt{1-\xi^2}}$ is a factor due to coordinate transformation from $\theta$ to $\xi$. Specifically, $\tilde{\psi}(\xi)$ is given by
where \( \mu = 0 \) and \( \sigma = \frac{4}{G+3} \) to ensure that the truncated interval \([-\frac{4}{G+3}, \frac{4}{G+3}]\) accounts for 6\( \sigma \) of the Gaussian function, and \( a_0 \) is a normalization scalar such that \( \int_{-\frac{4}{G+3}}^{\frac{4}{G+3}} \psi(\xi) d\xi = 1 \).

The results are shown in Fig. 8. From Fig. 8a and b, it is observed that the results under Gaussian dictionaries are much better than the results under Dirac delta dictionaries when the number of dictionaries is small, e.g., \( G/M \leq 1 \). It is also observed that both Frobenius norm error and channel estimate MSE of NNLS and SPICE under Dirac delta dictionaries decrease dramatically as \( G \) increases. In contrast, the results under Gaussian dictionaries become slightly worse when \( G/M > 0.5 \). As \( G \) becomes large, the results with Gaussian dictionary functions converge to the Dirac delta case. This is of course explained by the fact that the thin Gaussian density with normalized integral is more and more similar to Dirac delta functions. A similar behavior is observed in Fig. 8c and d. These results indicate that by choosing some template dictionary function may achieve advantages of the quantization of the AoA domain compared to Dirac delta function for small size of the dictionary, which reduces significantly the dimension and computational complexity.

6 Conclusion

In this work, we addressed the problem of estimating the covariance matrix of the channel vector from a set of noisy UL pilot observations in massive MIMO systems. By modeling the ASF of the channel as a parametric representation in the angle domain, we proposed the NNLS-, QP-, and ML-EM-based estimators to obtain the model parameters. In order to find the discrete scattering components (number and
location of the spikes in the ASF), we adopted MDL and MUSIC methods. Theoretical results guarantee that the separation of the spikes with respect to the clutter of eigenvalues due to the diffuse scattering components is large, for a large number of antennas. This yields that the spikes support estimation in the massive MIMO regime is very reliable. In addition, our method estimates both the coefficients of the spikes and the coefficients of the diffuse part of the ASF jointly. Extensive numerical simulations based on realistic channel emulator QuaDriGa under 3GPP communication scenarios show that the proposed methods are superior to several state-of-the-art algorithms in the literature in terms of different performance metrics, especially for a small number of samples, which is particularly relevant for the massive MIMO application.

Appendix A: Consistency of MUSIC

We provide an analysis of the consistency of proposed MUSIC algorithm based on the results in [62]. A more challenging scaling regime in [62] was studied, where the amplitude of the spikes decreases by increasing $M$ such that identifying them becomes more difficult. Specifically, recall in (6) that the ASF $\gamma(\xi)$ is decomposed into discrete part $\gamma_d(\xi) = \sum_{i=1}^r \epsilon_i \delta(\xi - \phi_i)$ and continuous parts $\gamma_c(\xi)$. In the pessimistic scaling regime

Figure 8: Comparison with Dirac delta and Gaussian dictionaries with $N/M = 0.125$ under various $G/M$ for scenario 3GPP-3D-UMa-NLOS.
studied in [62], the coefficients of spikes \( c_i^{(M)} : i \in [r] \) scale with \( M \) according to \( c_i^{(M)} = \frac{\nu_i}{M} \), where \( \{\nu_i : i \in [r]\} \) are positive constants encoding the relative strength of the spikes. The result in [62] shows that when the number of antennas \( M \) and the sample size \( N \) both approach to infinity at the same rate such that \( \frac{M}{N} \to \zeta > 0 \), where \( \zeta \) is a finite constant, if the separation condition \( \min_{i \in [r]} \nu_i > \omega_0(\zeta) \) holds, where \( \omega_0(\zeta) \geq \|\gamma_\infty\|_\infty + N_0 \) is a monotonically increasing function of \( \zeta \), then MUSIC is asymptotically consistent, i.e., \( M(\phi_i - \phi_j) \xrightarrow{\text{a.s.}} 0, \forall i \in [r] \), where a.s. stands for almost surely. In our wireless communication scenario, the spike coefficients \( \{\nu_i : i \in [r]\} \) remain the same independent of the number of BS antennas \( M \). We mimic this by assuming that the coefficients \( \{\nu_i\} \) are also growing proportionally to \( M \) like \( \nu_i = Mc_i \). Then, the separation condition would be satisfied for any finite \( \zeta \) and for any practically relevant \( \gamma_\infty \) provided that \( M \) is sufficiently large. It is worthwhile to emphasize that this result implies that no matter how small the spike amplitudes \( \{c_i : i \in [r]\} \) are and no matter how small the number of samples \( N \) is compared with \( M \) (of course provided that the asymptotic sampling ratio \( \zeta \) remains finite), and MUSIC will be able to recover all the spikes if \( M \) is sufficiently large. This provides a strong argument in favor of using MUSIC as the model order and spike location estimation for our application.

**Appendix B: Toeplitization of the sample covariance matrix**

The Toeplitized matrix \( \tilde{\Sigma}_h \) is obtained by solving the orthogonal projection problem

\[
\tilde{\Sigma}_h = \arg \min_\Sigma \| \Sigma - \tilde{\Sigma}_h \|_F^2, \quad \text{s.t. } \Sigma \text{ is Hermitian Toeplitz.}
\]  

(47)

Define \( \sigma \in \mathbb{C}^{M \times 1} \) as the first column of \( \Sigma \) and \( \mathcal{P} \) as the operation that projects \( \sigma \) and its conjugate to a Hermitian Toeplitz matrix as \( \Sigma = \mathcal{P}(\sigma) \). We further define \( \mathcal{K}_i = \{(r,c) : r \leq c \text{ with } [\Sigma]_{r,c} = [\sigma]_i, i \in [M] \} \) as the set of all those indices in \( \Sigma \), in which the \( i \)-th variable \( [\sigma]_i \), appears. Then, the objective function in (47) can be equivalently presented as

\[
f(\sigma) = \min_\sigma \| \mathcal{P}(\sigma) - \tilde{\Sigma}_h \|_F^2 = \min_\sigma \sum_{i=1}^M \sum_{(r,c) \in \mathcal{K}_i} \left| [\sigma]_i - [\tilde{\Sigma}_h]_{r,c} \right|^2.
\]  

(48)

By setting the derivative of \( [\sigma]_i \) to zero, the optimum \( \sigma^* \) is achieved as

\[
\frac{\partial f(\sigma)}{\partial [\sigma]_i} = \sum_{(r,c) \in \mathcal{K}_i} 2([\sigma]_i - [\tilde{\Sigma}_h]_{r,c}) \xrightarrow{\hat{=}} 0, \quad \forall i \in [M],
\]  

(49)

\[
\Rightarrow [\sigma^*]_i = \frac{\sum_{(r,c) \in \mathcal{K}_i} [\tilde{\Sigma}_h]_{r,c}}{|\mathcal{K}_i|}, \quad \forall i \in [M],
\]  

(50)

where \(|\mathcal{K}_i|\) denotes the number of elements in \( \mathcal{K}_i \). Correspondingly, \( \tilde{\Sigma}_h = \mathcal{P}(\sigma^*) \).
Appendix C: Computational complexity

We briefly provide an analysis of the floating-point operations (FLOP)-based computational complexity for proposed NNLS and ML-EM algorithms. We adopt the result of FLOP counting for different operations summarized in [63]. The matrix inverse operation for a positive definite matrix is based on Cholesky decomposition.

Complexity of NNLS algorithm

In general, the algorithms tackling NNLS can be roughly divided into active-set and projected gradient descent approaches [52]. The projected gradient descent approaches are known to require a very large number of iterations to converge to the optimum due to very small step size around the optimality to guaranty the convergence [64]. As an example, we have numerically tested the sequential coordinate-wise algorithm [65], which has a very simple update rule but unacceptable large required number of iterations to converge. Therefore, we adopt the standard active-set method [53], which is based on the observation that only a small subset of constraints are usually active at the solution. Specifically, considering the NNLS problem

\[ x^* = \arg \min_{x \geq 0} \| Ax - y \|, \]

the active-set method consists of two nested loops, where we count FLOP of the least squares \[ (A_P^T A_P)^{-1} A_P^T y \] in each inner and outer loop as well as the expression \[ A^T (y - Ax) \] in each outer loop, where \( A_P \) is a matrix associated with only the variables currently in the active set \( P \), please see [53] for more details. The input dimensions of our NNLS problem in (26) are \( A \in \mathbb{R}^{2M \times \hat{G}} \) and \( y \in \mathbb{R}^{2M \times 1} \), where \( \hat{G} = G + \hat{r} \) and \( 2M \) is due to the complex variables. Assuming that the total required number of outer loops is \( I \) and inner loops in the \( i \)-th outer iteration are \( J_i \) and ignoring the potential dimension reduction in inner loops, the upper bounded FLOP is given by

\[
\text{FLOP}_{\text{NNLS}} \leq 4(I + 1)\hat{G}M + \frac{1}{2} \sum_{i=1}^{I} \left( (1 + J_i) \left( 6Mi^2 + i^3 + 6Mi + 3i^2 \right) \right),
\]

\[
= 4(I + 1)\hat{G}M + 3M \sum_{i=1}^{I} i + \frac{3}{2} (2M + 1) \sum_{i=1}^{I} i^2 + \frac{1}{2} \sum_{i=1}^{I} i^3 + f(J_i),
\]

where \( f(J_i) := \frac{1}{2} \sum_{i=1}^{J_i} (6Mi^2 + i^3 + 6Mi + 3i^2) \) is the number of FLOP of inner loops. With a averaged number of inner iterations \( \tilde{J} \), the FLOP of NNLS can be approximated as

\[
\text{FLOP}_{\text{NNLS}} \approx 4(I + 1)\hat{G}M + \left( 1 + \tilde{J} \right) \left( \frac{1}{8} I^4 + (M + \frac{3}{4}) I^3 + (3M + \frac{7}{8}) I^2 + (2M + \frac{1}{4}) I \right).
\]

Complexity of ML-EM algorithm

We count the FLOP of calculating (36) in each E-step and (39) in each M-step. Note that the terms \( D^H Y \) and the gram \( D^H D \) in (36) only need to be calculated once and used in

---

11 Although in [63] a FLOP is assumed to be either a complex multiplication or complex summation, we only consider the complex multiplications since in practice the summation operations can be further optimized so that its complexity is much less than the complexity of multiplications.
each iteration. Additionally, $\hat{U}(\ell)$ is a diagonal matrix and its inverse in (36) can be easily obtained by taking the reciprocals of the main diagonal. Assuming that the total number of iterations is $I_{EM}$, the required number of FLOP is given as

$$\text{FLOP}_{EM} = \hat{G}M\left(N + \frac{1}{2}(\hat{G} + 1)\right) + I_{EM}\left(\frac{1}{2}\hat{G}^3 + \hat{G}^2\left(N + \frac{3}{2}\right) + \hat{G}N\right).$$

(54)

**Abbreviations**

AoA  Angle of arrival
ASF  Angular scattering function
AWGN  Additive white Gaussian noise
BS  Base station
CSI  Channel state information
DL  Downlink
EM  Expectation–maximization
FDD  Frequency division duplexing
LoS  Line of sight
MDL  Minimum description length
MIMO  Multiple input multiple output
ML  Maximum likelihood
MM  Majorization–minimization
MMSE  Minimum mean square error
MUSIC  Multiple signal classification
NNLS  Nonnegative least-squares
OFDM  Orthogonal frequency division multiplexing
PSD  Positive semi-definite
QP  Quadratic programming
QuaDriGa  QUAsi Deterministic RadIo channel GenerAtor
RB  Resource block
SNR  Signal-to-noise ratio
UL  Uplink
ULA  Uniform linear array
US  Uncorrelated scattering
WSS  Wide-sense stationary

**Author contributions**

All authors read and approved the final manuscript.

**Funding**

Open Access funding enabled and organized by Projekt DEAL. At the time of this research, the authors were all employed at the Technische Universität Berlin and this work has been not funded by external sources.

**Availability of data and materials**

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. All results can be fully replicated by implementing the described algorithms in MATLAB or in any other programming language capable of scientific computation, and generating via QuaDriGa or any other equivalent channel simulator the statistics of the channel vectors as correlated Gaussian random vectors as described in this work.

**Declarations**

**Competing interests**

This study is purely academic and theoretical and does not represent any corporate interest beyond the scientific interest of the authors.

**Received**: 14 November 2022  **Accepted**: 16 February 2023

**Published online**: 03 March 2023

**References**

1. M.B. Khalilsarai, T.Yang, S.Haghighatshoar, G. Caire, Structured channel covariance estimation from limited samples in massive MIMO, in ICC 2020-2020 IEEE International Conference on Communications (ICC). (IEEE, 2020), pp. 1–7

2. F. Boccardi, R.W. Heath, A. Lozano, T.L. Marzetta, P. Popovski, Five disruptive technology directions for 5G. IEEE Commun. Mag. 52(2), 74–80 (2014)

3. T.L. Marzetta, E.G. Larsson, H.Yang, H.Q. Ngo, Fundamentals of Massive MIMO. Cambridge University Press, 2016
4. Z. Chen, C. Yang, Pilot decontamination in wideband massive MIMO systems by exploiting channel sparsity. IEEE Trans. Wireless Commun. 15(7), 5087–5100 (2016)

5. H. Yin, L. Cottatellucci, D. Gesbert, R.R. Muller, G. He, Robust pilot decontamination based on joint angle and power domain discrimination. IEEE Trans. Signal Process. 64(11), 2990–3003 (2016)

6. S. Haghhighat, G. Caire, Massive MIMO pilot decontamination and channel interpolation via wideband sparse channel estimation. IEEE Trans. Wireless Commun. 16(12), 8316–8332 (2017)

7. M.B. Khalil, S. Haghhighat, X. Yi, G. Caire, Fdd massive mimo via ul/dl channel covariance extrapolation and active channel sparsification. IEEE Trans. Wireless Commun. 18(1), 121–135 (2018)

8. M.N. Boroujerdi, S. Haghhighat, G. Caire, Low-complexity statistically robust precoder/detector computation for massive MIMO systems. IEEE Trans. Wireless Commun. 17(10), 6516–6530 (2018)

9. S. Haghhighat, G. Caire, Low-complexity massive mimo subspace estimation and tracking from low-dimensional projections. IEEE Trans. Signal Process. 66(7), 1832–1844 (2018)

10. A. Adhikary, J. Nam, J.-Y. Ahn, G. Caire, Joint spatial division and multiplexing: the large-scale array regime. IEEE Trans. Inform. Theory 59(10), 6441–6463 (2013)

11. S. Qiu, D. Chen, D. Qiu, K. Luo, T. Jiang, Downlink precoding with mixed statistical and imperfect instantaneous csi for massive mimo systems. IEEE Trans. Veh. Technol. 67(4), 3028–3041 (2017)

12. L. You, J. Xiong, A. Zappone, W. Wang, X. Gao, Spectral efficiency and energy efficiency tradeoff in massive mimo downlink transmission with statistical csi. IEEE Trans. Signal Process. 68, 2645–2659 (2020)

13. H. Liu, X. Yuan, Y.J. Zhang, Statistical beamforming for fdd downlink massive mimo via spatial information extraction and beam selection. IEEE Trans. Wireless Commun. 19(7), 4617–4631 (2020)

14. V. Raghavan, A.M. Sayeed, Sublinear capacity scaling laws for sparse mimo channels. IEEE Trans. Inf. Theory 57(1), 345–364 (2010)

15. S. Haghhighat, G. Caire, Massive mimo channel subspace estimation from low-dimensional projections. IEEE Trans. Signal Process. 65(2), 303–318 (2016)

16. 3GPP Study on channel model for frequencies from 0.5 to 100 ghz (release 15), 3rd Generation Partnership Project (3GPP), Tech. Rep. TR 38.901 V15.0.0, (2018)

17. 3GPP Spatial channel model for multiple input multiple output (MIMO) simulations (release 16), 3rd Generation Partnership Project (3GPP), Tech. Rep. TR 25.996 V16.0.0, (2020)

18. S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials. IEEE Trans. Antennas Propag. 62(6), 3242–3256 (2014)

19. V.Ya, J. Choi, R.W. Heath, The impact of beamwidth on temporal channel variation in vehicular channels and its implications. IEEE Trans. Veh. Technol. 66(6), 5014–5029 (2016)

20. K. Mahler, W. Keusgen, F. Tufvesson, T. Zemen, G. Caire, Propagation of multipath components at an urban intersection, in IEEE 82nd Vehicular Technology Conference (VTC2015-Fall). (IEEE, 2015), pp. 1–5

21. L. Miretti, R. L. G. Cavalcante, S. Stanczak, Fdd massive mimo channel spatial covariance conversion using projection methods, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (IEEE, 2018), pp. 3609–3613

22. M. B. Khalilsarai, Y. Song, T. Yang, S. Haghhighat, and G. Caire, Uplink-downlink channel covariance transformations and precoding design for FDD massive MIMO, in 2019 53rd Asilomar Conference on Signals, Systems, and Computers. (IEEE, 2019), pp. 199–206

23. Y. Song, M. B. Khalilsarai, S. Haghhighat, G. Caire, Deep learning for geometrically-consistent angular power spread function estimation in massive MIMO, in GLOBECOM 2020-2020 IEEE Global Communications Conference. (IEEE, 2020), pp. 1–6

24. J.P. González-Coma, P. Suárez-Casal, P.M. Castro, L. Castedo, FDD channel estimation via covariance estimation in wideband massive MIMO systems. Sensors 20(3), 930 (2020)

25. A. Decurninge, M. Guillaud, and D. T. Slock, Channel covariance estimation in massive MIMO frequency division duplex systems, in Globecom Workshops (GC Wkshps), 2015 IEEE (IEEE, 2015), pp. 1–6

26. V.A. Marchenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices. Matematicheskii Sbornik (N.S.) 1(10), 6516–6530 (2018)

27. W. Hachem, P. Loubaton, J. Najim, The empirical eigenvalue distribution of a Gram matrix: From independence to stationarity, arXiv preprint math/0502535, (2005)

28. R. Couillet and M. Debbah, Random matrix methods for wireless communications. (Cambridge University Press, 2011)

29. M. Pourahmadi, High-Dimensional Covariance Estimation: With High-Dimensional Data. (Wiley, 2013), vol. 882

30. P. Ravikumar, M.J. Wainwright, G. Raskutti, B. Yu et al., High-dimensional covariance estimation by minimizing ℓ 1-penalized log-determinant divergence. Electron. J. Stat. 5, 935–980 (2011)

31. Y. Chen, A. Wiesel, A.O. Hero, Robust shrinkage estimation of high-dimensional covariance matrices. IEEE Trans. Signal Process. 59(9), 4097–4107 (2011)

32. J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3), 432–441 (2008)

33. P. Stoica, P Babu, J. Li, SPICE: a sparse covariance-based estimation method for array processing. IEEE Trans. Signal Process. 59(2), 629–638 (2011)

34. H. Xie, F. Gao, S. Jin, J. Fang, Y.-C. Liang, Channel estimation for tdd/fdd massive mimo systems with channel covari- ance computing. IEEE Trans. Wireless Commun. 17(6), 4206–4218 (2018)

35. S. Park, R.W. Heath, Spatial channel covariance estimation for the hybrid MIMO architecture: A compressive sensing-based approach. IEEE Trans. Wireless Commun. 17(12), 8047–8062 (2018)

36. P. Stoica, R.L. Moses et al., Spectral analysis of signals (Pearson Prentice Hall Upper Saddle River, NJ, 2005)

37. P. Chen, Z. Chen, Z. Cao, X. Wang, A new atomic norm for doa estimation with gain-phase errors. IEEE Trans. Signal Process. 68, 4293–4306 (2020)

38. A.M. Sayeed, Deconstructing multiantenna fading channels. IEEE Trans. Signal Process. 50(10), 2563–2579 (2002)

39. E. De Carvalho, A. Ali, A. Amiri, M. Angelichinoski, R.W. Heath, Non-stationarities in extra-large-scale massive mimo. IEEE Wirel. Commun. 27(4), 74–80 (2020)
40. M. J. Wainwright, *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. (Cambridge University Press, 2019), vol. 48
41. B. Hanssens, K. Saito, E. Tanghe, L. Martens, W. Joseph, J.-I. Takada, Modeling the power angular profile of dense multipath components using multiple clusters. IEEE Access 6, 56084–56098 (2018)
42. J. A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*. (Cambridge University Press, 2006)
43. A. Ali, N. González-Prelcic, R.W. Heath, Estimating millimeter wave channels using out-of-band measurements, in *Information Theory and Applications Workshop (ITA)*. (IEEE, 2016), pp. 1–6 (2016)
44. G. Schwarz, Estimating the dimension of a model, The annals of statistics, 461–464, (1978)
45. R.O. Schmidt, Multiple emitter location and signal parameter estimation. Antennas Propagat. IEEE Trans. 34(3), 276–280 (1986)
46. P. Stoica, A. Nehorai, MUSIC, maximum likelihood, and Cramer-Rao bound. IEEE Trans. Acoust. Speech Signal Process. 37(5), 720–744 (1989)
47. M. Wax, T. Kailath, Detection of signals by information theoretic criteria. IEEE Trans. Acoust. Speech Signal Process. 33(2), 387–392 (1985)
48. E. Fishler, M. Grosmann, H. Messer, Detection of signals by information theoretic criteria: General asymptotic performance analysis. IEEE Trans. Signal Process. 50(5), 1027–1036 (2002)
49. A. Barron, J. Rissanen, B. Yu, The minimum description length principle in coding and modeling. IEEE Trans. Inf. Theory 44(6), 2743–2760 (1998)
50. W. Xu, M. Kaveh, Analysis of the performance and sensitivity of eigendecomposition-based detectors. IEEE Trans. Signal Process. 46(6), 2743–2760 (1998)
51. A.P. Liavas, P.A. Regalia, On the behavior of information theoretic criteria for model order selection. IEEE Trans. Signal Process. 49(8), 1689–1695 (2001)
52. D. Chen, R. J. Plemmons, Nonnegativity constraints in numerical analysis, in *The Birth of Numerical Analysis*. (World Scientific, 2010), pp. 109–139
53. C.L. Lawson, R. Hanson, *Solving least squares problems*. pretice-hall Englewood Cliffs, NJ, 1974
54. D.R. Hunter, K. Lange, A tutorial on MM algorithms. Am. Stat. 58(1), 30–37 (2004)
55. Y. Sun, P. Babu, D.P. Palomar, Majorization-minimization algorithms in signal processing, communications, and machine learning. IEEE Trans. Signal Process. 65(3), 794–816 (2016)
56. A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B (Methodol.) 39(1), 1–22 (1977)
57. J. Nam, A. Adhikary, J.-Y. Ahn, G. Caire, Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling, IEEE J. Sel. Topics Sig. Proc. (JSTSP), 8(5), 876–890 (2014)
58. O. Najim, P. Vallet, G. Ferré, X. Mestre, On the statistical performance of music for distributed sources, in *IEEE Statistical Signal Processing Workshop (SSP)*. (IEEE, 2016), pp. 1–5
59. R. Hunger, *Floating point operations in matrix-vector calculus* (Munich University of Technology, Inst. for Circuit Theory and Signal Processing, 2005)
60. D. P. Bertsekas, A. Scientific, *Convex optimization algorithms*. (Athena Scientific Belmont, 2015)
61. V. Franc, V. Hlaváč, M. Navara, Sequential coordinate-wise algorithm for the non-negative least squares problem, in *International Conference on Computer Analysis of Images and Patterns*. (Springer, 2005), pp. 407–414

**Publisher’s Note**
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.