Hierarchical Sparse Channel Estimation for Massive MIMO

7 authors, including:

Gerhard Wunder
Freie Universität Berlin
224 PUBLICATIONS   **3,971** CITATIONS

SEE PROFILE

Axel Flinth
University of Gothenburg
34 PUBLICATIONS   **205** CITATIONS

SEE PROFILE

Mahdi Barzegar
Technische Universität Berlin
34 PUBLICATIONS   **160** CITATIONS

SEE PROFILE

Saeid Haghighatshoar
Technische Universität Berlin
67 PUBLICATIONS   **886** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   H2020 PhasmaFOOD (Grant No. 732541) View project

Project   H2020 (ICT-7-2016) ONE5G View project

Hierarchical Sparse Channel Estimation for Massive MIMO

Gerhard Wunder[1], Ingo Roth[2], Axel Flinth[3], Mahdi Barzegar[4], Saeid Haghighatshoar[4], Giuseppe Caire[4], Gitta Kutyniok[3]

[1]Heisenberg Communications and Information Theory Group, [2]Quantum Information Theory Group, FU Berlin

[3]Applied Functional Analysis Group, [4]Communications and Information Theory Group, TU Berlin

Abstract—The problem of wideband massive MIMO channel estimation is considered. Targeting for low complexity algorithms as well as small training overhead, a compressive sensing (CS) approach is pursued. Unfortunately, due to the Kronecker-type sensing (measurement) matrix corresponding to this setup, application of standard CS algorithms and analysis methodology does not apply. By recognizing that the channel possesses a special structure, termed hierarchical sparsity, we propose an efficient algorithm that explicitly takes into account this property. In addition, by extending the standard CS analysis methodology to hierarchical sparse vectors, we provide a rigorous analysis of the algorithm performance in terms of estimation error as well as number of pilot subcarriers required to achieve it. Small training overhead, in turn, means higher number of supported users in a cell and potentially improved pilot decontamination. We believe, that this is the first paper that draws a rigorous connection between the hierarchical framework and Kronecker measurements. Numerical results verify the advantage of employing the proposed approach in this setting instead of standard CS algorithms.

I. INTRODUCTION

Massive MIMO, i.e. deploying large number of antennas at the base station, is a key technology for 5G [1]. Although its benefits are by now well understood and documented, the critical bottleneck for massive MIMO deployment is still the acquisition of channel state information (CSI), with designs attempting to balance the conflicting requirements of training overhead and co-pilot contamination reduction, in addition to the standard requirement of computational efficiency. This problem is even more important in massive machine type communications (MTC) where accurate CSI with low-training overhead is of critical importance [2], [3].

Towards addressing these issues, one major line of works applies compressed sensing (CS) techniques in order to account for and exploit the sparsity properties of the wireless channel. The typical approach comes in two stages: first the (spatial) covariance matrix of the received signal is estimated and, second, the dedicated user channels are estimated within the estimated spatial subspaces. This approach has been mostly applied to the narrowband signaling case, exploiting the channel sparsity in the, so called, angle domain (see [4] for an overview). Extensions of CS techniques to the wideband (OFDM) massive MIMO channel have been recently considered in [5], [6], [7]. One intuitive motivation for such an approach is that, in the wideband setting, there are more degrees of freedom available so that pilot contamination induced by co-pilots in other cells is effectively combated.

Initial work has been provided in [5], arguing that with large enough resolution in angular and time domain, controlled by the number of antennas and subcarriers, the user channels become approximately mutually orthogonal so that pilot contamination diminishes. Notably, the subspace estimates are obtained through pilot coordination over multiple slots, but not primarily through sparse channel estimation. Moreover, the claimed orthogonality is not rigorous and crucially depends on the parameter setting, let alone that bandwidth (spanned by the system subcarriers) is not a free design parameter that can be set arbitrarily large. In [6], a computationally efficient CS-based CSI acquisition approach is proposed, utilizing observations from only a limited number of subcarriers and antennas. Moreover, a fine resolution of angle and time domain together with a sparse subspace tracking algorithm is proposed. The resulting algorithm shows good performance and is exploited for pilot decontamination purposes. A similar setting is also considered in [7], where a minimimum mean squared error (MMSE) channel estimator is proposed, however, requiring prior knowledge of the channel second order statistics.

Generally, a two stage approach may require excessive observations in time in order to obtain an accurate estimate of the received signal covariance matrix, which may be unacceptable, e.g., in massive MTC setting. Another major problem is that the proposed CS algorithms lack rigorous performance analysis in terms of estimation error and, equally important, number of utilized subcarriers in order to achieve it. This is mainly due to the specific Kronecker-like measurement structure of the equivalent CS problem, where, as we will show, the classical CS assumptions fail to hold, entailing convergence issues and leakage effects. However, targeting low complexity algorithms as well as small training overhead, say, in future MTC applications, it is imperative to understand such structures and to derive a tailored algorithmic framework.

Contributions. We propose an efficient "one stage" uplink massive MIMO wideband (OFDM) channel estimation taking into account only the sparsity of the wideband channel into account without any additional prior knowledge. In particular, we identify that the channel estimation problem can be posed as the identification of a vector that is *hierarchically sparse*, i.e., it is not only sparse but its support possesses certain structural properties. This property, together with the Kronecker-like measurement, is taken into account for the design of an efficient CS-inspired algorithm tailored for this particular setup. In addition, by extending the standard CS analysis methodology to hierarchical sparse vectors, we provide a

rigorous analysis of the algorithm performance in terms of estimation error as well as number of pilot subcarriers required to achieve it. We believe, that this is the first paper that draws a rigorous connection between the hierarchical framework and Kronecker measurements. Preliminary simulations show that exploitation of the hierarchical sparsity property leads to improved estimation performance compared to standard CS algorithms that ignore this property. Even worse, standard algorithms may completely fail in some extreme parameter settings. Obviously, this might have some profound impact on system parameters such as pilot signal design, user capacity per cell etc.

Basic notations. $\|x\|_{\ell_q} := (\sum_i |x_i|^q)^{1/q}$, $q > 0$, is the usual notion of ℓ_q-norms, $\|x\| := \|x\|_{\ell_2}$ and $\|X\|$ is the Frobenius norm of matrix X. $\mathcal{A}$ denotes a set of cardinality $|\mathcal{A}|$ and $[N]$ denotes $\{0, 1, ..., N-1\}$. The elements of a vector/sequence x are denoted as $(x)_i$ (or simply x_i if it clear from the context). Vector $x_{\mathcal{A}}$ (matrix $X_{\mathcal{A}}$) is the projection of elements (rows) of vector x (matrix X) onto $\mathbb{C}^{\mathcal{A}}$. $\odot$ means point-wise (Hadamard) product, I_n is the $n \times n$ identity matrix, $\text{diag}(x)$ is the diagonal matrix with $x \in \mathbb{C}^n$ on its diagonal. $A^{H/T}$ is the Hermitian/transpose of matrix A. The $N \times D$ ($N \times N$) DFT matrix is denoted as $F_{N,D}$ ($F_{N,N} = F_N$) with $(F_{N,D})_{m,n} := e^{-j2\pi mn/N}$, $m \in [N], n \in [D]$. $\mathcal{CN}\left(0, \sigma^2 I_n\right)$ denotes the multivariate complex Gaussian distribution of zero mean and covariance matrix $\sigma^2 I_n$. A vector $x \in \mathbb{C}^N$ is called s-sparse if it consists of at most s non-zero elements. The set of non-zero elements (support) of $x \in \mathbb{C}^N$ is denoted as $\text{supp}(x)$.

Organization of the paper: First, we describe our signal model and formulate the channel recovery problem. Next, we briefly present the recent framework of hierarchical sparsity, which was developed by two of the authors together with co-authors in [8]. This framework is applied to the channel estimation problem under the assumption of, so called, on-grid channel parameters, and an efficient estimation algorithm is proposed. Numerical results on the performance of the algorithm are presented, demonstrating that highly accurate channel estimation can be achieved with a very small training overhead. The more general (and practical) case of off-grid channel parameters is then considered where it is shown that the framework also applies with certain modifications that take into account basis mismatch effects.

II. WIDEBAND SIGNAL MODEL

We consider an uplink massive MIMO OFDM wideband channel with single-antenna users and $M \gg 1$ antenna elements at the base station corresponding to an array manifold $a\left(\cdot\right) : [0, \pi) \to \mathbb{C}^M$, which maps angular to spatial domain. Considering a uniform linear array (ULA), the array manifold is given by $a\left(\phi\right) = (1, e^{-j2\pi d \sin \phi}, \ldots, e^{-j2\pi d(M-1)\sin \phi})^T$. Here, d is normalized spatial separation of the ULA, which without loss of generality (w.l.o.g.) is assumed equal to 1 in the following. As is routinely done, we perform the change of variable $\theta = \sin(\phi) \in [0, 1)$ and, with a slight abuse of notation, we write the array manifold as a function of $2\pi\theta$.

Considering a discretized approximation of the interval $[0, 2\pi)$ by the M points $\{k2\pi/M\}_{k=0}^{M-1}$, yields the steering matrix $A_\theta := [a(0), a(1/M)\ldots, a((M-1)/M)] = F_M \in \mathbb{C}^{M \times M}$.

Further, suppose there are $N \gg 1$ OFDM subcarriers located at the (angular) frequencies $\omega_0, \omega_1, \ldots, \omega_{N-1}$, where $\omega_k := 2\pi k/T_s$, with T_s being the OFDM symbol duration. Assuming that the maximum delay spread of the channel for all antennas is no greater than a fraction $\alpha T_s, \alpha \leq 1$, which is the case in any reasonable OFDM design, the "delay manifold" $b\left(\cdot\right) : [0, \alpha T_s] \to \mathbb{C}^N$ is defined as $b\left(\tau\right) := (e^{-j\omega_0\tau} \, e^{-j\omega_1\tau}\ldots e^{-j\omega_{N-1}\tau})^T$, which maps the delay to the frequency domain. Considering a discretized approximation of $[0, T_s]$ by the N points $\{kT_s/N\}_{k=0}^{N-1}$, yields the steering matrix $A_\tau := [b(0), b(T_s/N), \ldots, b((D-1)T_s/N)] = F_{N,D} \in \mathbb{C}^{N \times D}$ where sample number $D \leq N$ is the discrete delay spread[1].

The channel of any user is a superposition of a small number L of impinging wavefronts (paths) characterized by their delay/angle pairs $\{(\tau_p, \theta_p)\}_{p=0}^{L-1}$, with $\tau_p \in [0, \alpha T_s]$, $\theta_p \in [0, 1)$, whose values are assumed to remain constant during the considered transmission interval. The channel "spatial-frequency" transfer matrix $H\left(t\right) \in \mathbb{C}^{N \times M}$ of an arbitrary user corresponding the OFDM symbol (slot) index $t \in \mathbb{Z}$ can be then be written as [5]

$$H\left(t\right) = \sum_{p=0}^{L-1} \rho_p\left(t\right) b\left(\tau_p\right) a^H\left(\theta_p\right), \quad (1)$$

where $\rho_p(t) \in \mathbb{C}$ is the time-varying gain of the p-th path at slot t.

Assuming that the base station observes T consecutive (pilot) OFDM slots dedicated for channel estimation, the overall received signal equals

$$Y\left(t\right) = \text{diag}(c(t))H(t) + Z\left(t\right), \, t \in [T]. \quad (2)$$

The matrix $Z\left(t\right) \in \mathbb{C}^{N \times M}$ represents noise with independent, identically distributed elements as $\mathcal{CN}(0, \sigma^2)$. Vector $c(t) := (c\left(t, \omega_0\right), ..., c\left(t, \omega_{N-1}\right)) \in \mathbb{C}^N$ contains the pilot symbols transmitted over slot t. For the single-cell case considered in this paper, it is reasonable to assume that users are assigned non-overlapping sets of subcarriers to transmit their pilots on, which, w.l.o.g., are set to unity, i.e., $\text{diag}(c(t)) = I_N$. Leveraging the sparse nature of the wideband massive MIMO channel, estimation of the channel of an arbitrary user can be achieved in principle by considering only the observations from the $O_\tau \leq N$ pilot subcarriers the user in consideration utilizes and $O_\theta \leq M$ antennas. Let $P_\tau \in \{0, 1\}^{O_\tau \times N}$ and $P_\theta \in \{0, 1\}^{O_\theta \times M}$ denote the corresponding *sampling* matrices in frequency and space dimensions, respectively. We consider a random sampling in frequency and space, i.e., the O_τ pilot subcarriers are selected uniformly from the N available subcarriers and similarly for the sampled antennas.

---

[1]In general, a denser discretized approximation for the angle and delay domains could be employed. We leave investigation of this case for future work.

In principle, identification of the continuous-valued channel parameters $\{(\rho_p, \tau_p, \theta_p)\}_{p=0}^{L-1}$ can be obtained from the low-dimensional sketches $\{P_\tau Y(t) P_\theta^T\}_{t=0}^{T-1}$ using the, so called, (two-dimensional) super-resolution approach (see, e.g., [9]), which is an extension of the one dimensional super-resolution approaches (see, e.g., [10], [11]). Unfortunately, the numerical resolution of the two-dimensional problem is computationally intensive, rendering this approach practical only for scenarios with O_τ and O_θ up to about 10 each.

Targeting low-complexity channel estimation, we consider a discretized representation of the channel matrix, which translates the physical channel sparsity to sparsity of an appropriately defined matrix that is to be identified by the estimator. In particular, let us first assume that each delay/angle pair lies exactly on the delay/angle grid corresponding to the steering matrices A_θ and A_τ, i.e., it holds $(\tau_p, \theta_p) = (k_p T_s/N, l_p 2\pi/M)$ for some $k_p \in [N]$ and $l_p \in [M]$, for all $p \in [L]$. It is easy to see that, in this case, $H(t)$ can be written as

$$H(t) = A_\tau W(t) A_\theta^H, \qquad (3)$$

where

$$W(t) := \sum_{p=0}^{L-1} \rho_p(t) e_{k_p} e_{l_p}^T \in \mathbb{C}^{D \times M}, \qquad (4)$$

with e_n denoting the canonical basis vector of appropriate dimension with the n-th element equal to 1. Matrix $W(t)$ is the *delay-angular representation* of the channel at time t, which is a sparse matrix, with only L nonzero elements out of a total MD. Note that the set of non-zero elements (support) of $W(t)$ is the same for each t, although $W(t_1) \neq W(t_2)$, for $t_1 \neq t_2$ due to the time-varying path gains.

The general case, i.e. when the delay/angle pairs are off the delay/angle grid, is a bit more subtle as the discretized model leads to *basis mismatch* effects [12]. In particular, even though the representation of (3) is still valid, the delay-angular representation matrix now equals

$$W(t) = \sum_{p=0}^{L-1} \rho_p(t) u_{\tau_p} u_{\theta_p}^H \in \mathbb{C}^{D \times M}, \qquad (5)$$

where $u_{\tau_p} \in \mathbb{C}^{D \times 1}, u_{\theta_p} \in \mathbb{C}^{M \times 1}$ are defined by the equations $b(\tau_p) = A_\tau u_{\tau_p}, a(\theta_p) = A_\theta u_{\theta_p}$, respectively, for all $p \in [L]$. In general, $W(t)$ no longer consists of only L non-zero elements due to energy leakage over the grid points [12] (see Fig. 1). As we will see, it can however (under a regularity condition) be approximated by a matrix with a so-called *hierarchical sparsity pattern*, for which recovery methods have recently been developed in [8].

III. PROBLEM STATEMENT

Considering the estimation of the channel of a single user, the problem can be formulated as the estimation of the set of matrices $\{W(t)\}_{t=0}^{T-1}$ given the low-dimensional sketches $\{P_\tau Y(t) P_\theta^T\}_{t=0}^{T-1}$. For technical reasons, the observed matrices $\{Y(t)\}_{t=0}^{T-1}$ are normalized by $1/\sqrt{O_\tau O_\theta}$, which, of
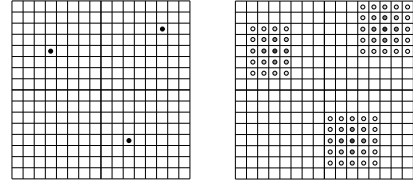

Fig. 1. Example of support of the delay-angular channel representation $W(t)$ for $L = 3$ impinging wavefronts. Left: on grid case, Right: off-grid case.

course, occurs no information loss, and, for convenience, are vectorized, resulting in

$$\begin{aligned} X(t) &:= \frac{1}{\sqrt{O_\tau O_\theta}} \text{vec}\left(P_\tau Y(t) P_\theta^T\right) \\ &= \Psi \text{vec}(W(t)) + \Phi \text{vec}(Z(t)), t \in [T], \end{aligned} \qquad (6)$$

where $\Phi := (1/\sqrt{O_\tau O_\theta}) P_\theta \otimes P_\tau \in \{0,1\}^{O \times MN}$ with $O := O_\tau O_\theta$ and

$$\Psi := \left(\sqrt{\frac{1}{O_\theta}} P_\theta A_\theta^*\right) \otimes \left(\sqrt{\frac{1}{O_\tau}} P_\tau A_\tau\right) \in \mathbb{C}^{O \times MD} \qquad (7)$$

is the, so called, *sensing matrix* [13] for the observations (measurements). By repeating this vectorization procedure over the slot dimension as well, the problem can be formulated as follows.

Problem 1. *Find a computationally efficient estimator of*

$$\bar{W} := \text{vec}\left([\text{vec}(W(0)), \dots, \text{vec}(W(T-1))]^T\right) \in \mathbb{C}^{MDT}$$

given the vector consisting of multiple measurements

$$\bar{X} := \text{vec}\left([\text{vec}(X(0)), \dots, \text{vec}(X(T-1))]^T\right) \in \mathbb{C}^{OT},$$

under the linear model

$$\bar{X} = \bar{\Psi}\bar{W} + \bar{Z},$$

where $\bar{Z} := \text{vec}\left([\Phi \text{vec}(Z(0)), \dots, \Phi \text{vec}(Z(T-1))]^T\right)$, $\bar{\Psi} := \Psi \otimes I_T \in \mathbb{C}^{OT \times MNT}$ *with* Ψ *as given in (7) and with the sampling matrices* $P_\tau \in \{0,1\}^{O_\tau \times N}$ *and* $P_\theta \in \{0,1\}^{O_\theta \times M}$ *generated by randomly and independently selecting* O_τ *and* O_θ *rows from the identity matrices* I_N *and* I_M, *respectively. We also ask for the scaling of required antennas* O_θ *and subcarriers* O_τ *for reliable (in a specific approximate sense that we will made precise later) recovery.*

Note that we have assumed random subsampling matrices in frequency and antenna spaces, which, in turn, imply random sets of pilot subcarriers. Even though not necessarily optimal, this random subsampling approach simplifies analysis and is actually shown to achieve good performance.

A naive application of standard results from compressive sensing theory [13] suggests that recovery is guaranteed as soon as

$$TO \gtrsim LT \log(TNM). \qquad (8)$$

However, this result only holds as long as the sensing matrix of the problem satisfies certain properties, e.g. the restricted
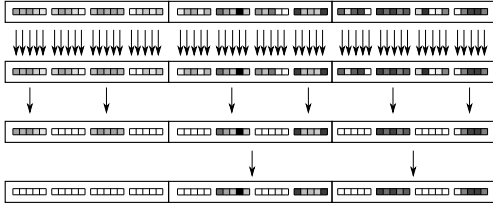
Fig. 2. Calculation of the action of the $\mathcal{T}_{2,2,5}$-operator on a vector $X \in \mathbb{C}^{3 \cdot 4 \cdot 5}$

isometry property (RIP) [13]. As we will see, the Kronecker-like sensing matrix of this problem *might not* possess such properties in general, rendering a different algorithm design and analysis methodology necessary. To this end, the appropriate mathematical framework is developed in the next section by exploiting the specific sparsity structure of $\bar{W}$.

IV. ALGORITHM DESIGN EXPLOITING STRUCTURAL PROPERTIES OF SPARSITY

This section identifies important structural properties of the channel matrices $\{W(t)\}_{t=0}^{T-1}$, which are taken into account for designing efficient and accurate algorithms, as well as obtaining rigorous performance guarantees. For this section, the case where the path/delay values of each path lie exactly on the sampling grid is considered, i.e., $W(t), t \in [T]$, consists of exactly L non-zero elements. The general case will be treated in the next section.

A. Hierarchical Sparsity and Algorithm Design

The fundamental observation towards an efficient channel estimation algorithm is that the delay/angle channel representation is not only sparse, but possesses certain structural properties as well, that fall within the notion of *hierarhical sparsity*.

Definition 1 (Hierarchical sparsity). Let $\mathbf{s} = (s_1, \ldots, s_\ell)$ be ℓ-tuples of natural numbers and consider a vector $\tilde{X} \in \mathbb{C}^{N_1 N_2 \cdots N_l}$, with integer $N_i \geq s_i, i \in \{1, 2, \ldots, l\}$. We define the hierarchical $\mathbf{s}$-sparsity of $\tilde{X}$ inductively as follows: For $l = 1$ it is exactly the same as the standard notion of sparsity (i.e., at most s_1 out of N_1 elements are non-zero). For $l > 1$, $\tilde{X}$ is called $\mathbf{s}$-sparse if it consists of N_1 blocks out of which at most s_1 are non-zero and each of the non-zero blocks is $(s_2, \ldots, s_\ell)$-sparse.

It is easy to see that the unknown vector $\bar{W} \in \mathbb{C}^{MDT}$, defined in Sec. III, is a level $l = 3$ compound vector that is (L, L, T)-sparse. Actually, for sufficiently large M, one may also expect that no more that K ($1 \leq K \leq L$) paths have the same angle, implying further restricting $\bar{W}$ as an (L, K, T)-sparse vector.

Clearly, the hierarchically sparse structure $\bar{W}$ *should be exploited* in algorithm design and analysis as it provides significant restrictions on the support of $\bar{W}$, compared to the standard notion of sparsity (which would characterize $\bar{W}$ simply as TL-sparse). Towards this end, the low-complexity, hierarchical hard thresholding pursuit (HiHTP) algorithm is considered

here, originally proposed in [8] for the estimation of general hierarchically sparse vectors from linear measurements. For the channel estimation problem, it takes the following form.

Algorithm 1 HiHTP

Require: measurement $\bar{X}$, sensing matrix $\bar{\Psi}$, (L, K, T) hierarchical sparsity for $\bar{W}$

1: $\hat{\bar{W}}^{(i)} = 0$
2: **repeat**
3:    $\mathcal{A}^{(i+1)} = \mathcal{T}_{(L,K,T)} \left(\hat{\bar{W}}^{(i)} + \bar{\Psi}^H \left(\bar{X} - \bar{\Psi} \hat{\bar{W}}^{(i)} \right) \right)$
4:    $\hat{\bar{W}}^{(i+1)} = \arg \min_{\substack{M \in \mathbb{C}^{MNT}, \\ \mathrm{supp}(M) \subseteq \mathcal{A}^{(i+1)}}} \left\{ \| \bar{X} - \bar{\Psi} M \| \right\}$
5: **until** stopping criterion is met at $i = i^*$

Ensure: (L, K, T)-sparse matrix $\hat{\bar{W}}^{(i^*)}$

The algorithm follows the philosophy of model-based compressed sensing [14]: In each iteration, it first makes a gradient descent step towards minimizing a least square objective, then projects the resulting signal onto the (L, K, T)-sparse support containing the most of its energy via application of operator $\mathcal{T}_{(L,K,T)}(\cdot)$, and subsequently solves a least squares problem restricted to that support to find the next iterate.

Utilization of the projection (or thresholding) operator $\mathcal{T}_{(L,K,T)}$ is the main differentiator of the HiHTP algorithm compared to the standard HTP algorithm [13]. In particular, for any compound vector $\tilde{X} \in \mathbb{C}^{N_1 N_2 \cdots N_l}$ and any $\mathbf{s} := (s_1, s_2, \ldots, s_l)$, $\mathcal{T}_{\mathbf{s}}(\tilde{X})$ is defined as

$$\mathcal{T}_{\mathbf{s}}(\tilde{X}) := \mathrm{supp} \arg_{\mathbf{s}\text{-sparse } \tilde{Z}} \| \tilde{X} - \tilde{Z} \|. \tag{9}$$

The action of this operator can be computed very efficiently. First, at the lowest level, it selects the s_l largest-magnitude entries (out of a total N_l entries) for each sub-block. Then, iteratively, at level $k < l$, it selects the s_k sub-blocks (out of a total N_k sub-blocks) whose best $(s_{k+1}, \ldots, s_l)$-sparse approximation are largest in l_2-norm. As an example, the iterative calculation of $\mathcal{T}_{2,2,5}$ applied on a vector $X \in \mathbb{C}^{3 \cdot 4 \cdot 5}$ is illustrated in Fig. 2.

B. Performance analysis

Towards characterizing the algorithm performance, the authors of [8] introduced the concept of Hierarchical RIP (HiRIP) constant.

Definition 2 (HiRIP). Given a matrix $\tilde{\Psi} \in \mathbb{C}^{O \times N_1 \cdots N_\ell}$ and a vector $\mathbf{s} = (s_1, s_2, \ldots, s_l)$, we denote by $\delta_{\mathbf{s}}$ the smallest $\delta \geq 0$ such that

$$(1 - \delta)\|\tilde{X}\|^2 \leq \|\Psi \tilde{X}\|^2 \leq (1 + \delta)\|\tilde{X}\|^2, \tag{10}$$

for all $\mathbf{s}$-sparse vectors $\tilde{X} \in C^{N_1 \cdots N_\ell}$. $\tilde{\Psi}$ is said to have the hierarchical RIP (HiRIP).

Note the definition of HiRIP is less general than standard RIP: Since $\mathbf{s}$-sparse vectors in particular are $s_1 \cdots s_\ell$-sparse, $s_1 \cdots s_\ell$-RIP implies $\mathbf{s}$-HiRIP, whereas $\mathbf{s}$-HiRIP does not necessary imply $s_1 \cdots s_\ell$-RIP. However, this more restricted

notion of RIP is well justified here as it explicitly takes into account the hierarchical sparsity property of the vectors that are considered in our problem.

Using the HiRIP concept, the following recovery guarantee for the HiHTP algorithm can be stated.

Theorem 1. *(Recovery guarantee [8]) Let* $3 \times (L, K, T) := (\min(3L, M), \min(3K, D), T)$ *and suppose that the sensing matrix* $\bar{\Psi}$ *in Algorithm 1 has a HiRIP constant*

$$\delta_{3 \times (L,K,T)} < \frac{1}{\sqrt{3}}. \tag{11}$$

Then, the sequence of estimates $\{\hat{\bar{W}}^{(i)}\}$ *of the HiHTP algorithm satisfies*

$$\|\bar{W} - \hat{\bar{W}}^{(i+1)}\| \le \kappa^i \|\bar{W} - \hat{\bar{W}}^{(0)}\| + \tau \|\bar{Z}\|, \tag{12}$$

for any $i \ge 0$, *with*

$$\kappa := \left(\frac{2\delta_{3 \times (L,K,T)}}{1 - \delta_{3 \times (L,K,T)}^2} \right)^{1/2} < 1 \tag{13}$$

and $\tau \le 5.15/(1 - \kappa)$.

It follows that, in order to ensure accurate recovery of $\bar{W}$ via the HiHTP algorithm and assuming $M > 3L$, $N > 3K$, we need to estimate the $(3L, 3K, T)$-HiRIP constants for $\bar{\Psi}$. We will use the following bound for general Kronecker-type sensing matrices [**?**].

Theorem 2. *Suppose* $\tilde{\Psi} := M_1 \otimes M_2 \otimes \cdots \otimes M_\ell$, *where* $M_k \in \mathbb{C}^{O_k \times N_k}$ *for all* k. *Further suppose that, for each* k, M_k *has* s_k-*sparse RIP with constant* δ_{s_k}. *Then, with* $\mathbf{s} = (s_1, \dots, s_k)$, $\tilde{\Psi}$ *satisfies (10) with a HiRIP constant*

$$\delta_{\mathbf{s}} \le \prod_{k=1}^{l} (1 + \delta_{s_k}) - 1. \tag{14}$$

The above theorem implies that sensing matrices resulting by Kronecker products will have HiRIP, provided each of the constituent matrices has the (standard) RIP. Notably, it is important to emphasize that while HiRIP is attainable, RIP may actually not: For example, it can be shown [15] that the (standard RIP-properties of a Kronecker product $M_1 \otimes \cdots \otimes M_\ell$ cannot be better than the RIP-properties of the weakest matrix (with respect to the total sparsity!) M_i, since $\delta_{\sum_k s_k}(M_1 \otimes \cdots \otimes M_\ell) \gtrsim \max_{i=1}^{\ell} \delta_{\sum_k s_k}(M_i)$. Hence, to consider the HiRIP / HiHTP framework, instead of the standard RIP framework, is *inevitable* to obtain recovery guarantees when using sensing matrices we consider in this publication.

C. Numerical Results

This section demonstrates the effectiveness of the HiHTP in obtaining highly accurate channel estimates with limited pilot overhead. In all cases, an OFDM system with $N = 64$ subcarriers and $M = 16$ antennas at the base station is considered. The "spatial-frequency" transfer matrix of the user in consideration is represented as in (3) and (4), i.e., with "on grid"

angle/delay values for each path, with $L = 3$. The angle/delay values of each path remain constant for the T slots considered in the estimation procedure, whereas the channel gains are independent and identically distributed (i.i.d.) over slots. For each slot, the channel gains are generated as i.i.d. complex Gaussian variables with a total power $\sum_{p=0}^{L-1} \mathbb{E}(|\rho_p(t)|^2) = 1$. The path angles $\{\theta_p\}_{p=0}^{L-1}$ are generated independently and uniformly over the angle sampling grid, however, no two paths are allowed to have the same angle. The path delays are independent and uniformly distributed over the delay sampling grid with $D = 16$. For this moderate number of antennas it is reasonable to consider all of them for channel estimation purposes, i.e., $O_\theta = M$, however, the number of (randomly) selected pilot subcarriers O_θ is a design variable.

Figure 3 depicts the (average) mean squared error (MSE), $\frac{1}{NM} \mathbb{E}\|H(t) - \hat{H}(t)\|^2$, of the space-frequency transfer matrix estimate obtained as $\hat{H}(t) := A_\tau \hat{W}(t) A_\theta^H$ where $\hat{W}(t)$ is the estimate of the delay-angular channel representation at slot t provided by HiHTP. The MSE is depicted as a function of the training overhead O_τ/N and for various values of observed slots T. The signal-to-noise ratio (SNR) was set equal to $1/\sigma^2 = 0$ dB. It can be seen that HiHTP offers excellent estimation accuracy for a very limited training overhead. For example, for $T = 1$, a training overhead of around 0.08 is sufficient to achieve a MSE that is almost one order of magnitude less than the noise level. This overhead should be compared with conventional estimation approaches which would require a pilot overhead in the order of $D/N = 0.25$. As expected, jointly considering $T > 1$ slots improves performance as the algorithm incorporates the common support of the delay-angular representation of the channel over slots. The gain of increasing T is more visible in the low training overhead regime, with $T = 4$ sufficient to achieve almost all of the possible gain. As a comparison, the performance of the standard HTP algorithm [13] is depicted. It can be seen that, for small training overhead, HTP performance is significant worse than HiHTP, exactly due to not taking into account the hierarchical sparsity of the channel. For sufficiently large training overhead, the performance of both algorithms are the same, implying that knowledge of the sparsity structure plays no role, exactly analogous to standard estimation theory where a priori information becomes irrelevant once sufficiently many observations are obtained.

Figure 4 shows the performance of HiHTP for a training overhead $O_\tau/N = 0.125$ as a function of SNR and for various T. As expected MSE performance improves with SNR. Utilizing more than one slots is beneficial for further improving performance in the low SNR regime.

V. HIERARCHICAL SPARSITY AND HIHTP PERFORMANCE UNDER BASIS MISMATCH

The previous section identified the structural properties of the delay/angular channel representation which were exploited in order to obtain an efficient estimation algorithm (HiHTP) and obtain performance guarantees using the concept of HiRIP. However, the analysis considered the on-grid case, i.e., with
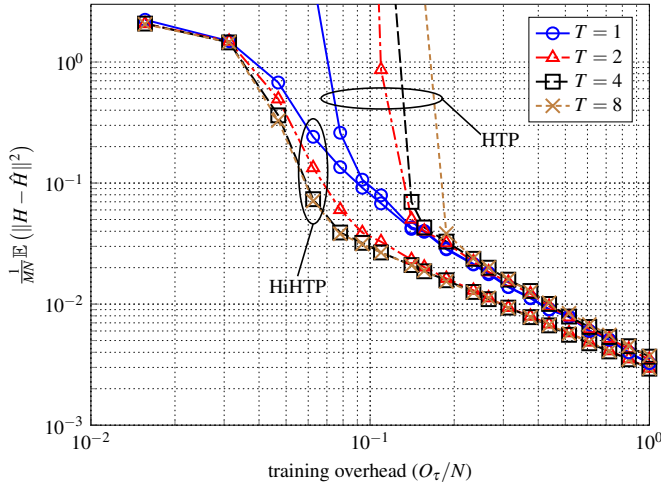
Fig. 3. MSE of HiHTP/HTP as a function of training overhead, for various values of T ($N = 64, M = 16, D = 16, L = 3, \text{SNR} = 10$ dB$, O_\theta = M$).
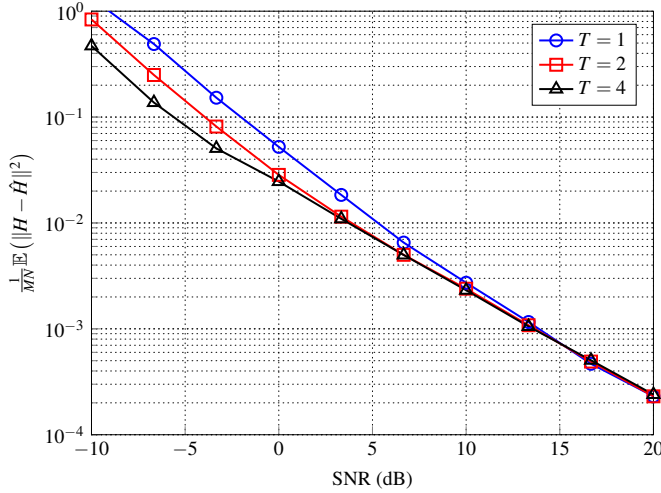

Fig. 4. MSE of HiHTP as a function of SNR ($1/\sigma^2$) for various values of T ($N = 64, M = 16, D = 16, L = 3, O_\tau = 0.125N, O_\theta = M$).

the delay/angle values lying on the grid assumed by the steering matrices A_τ and A_θ. As mentioned in Sec. II, this is not the case in general leading to basis mismatch effects. This section considers this case with the analysis split into two parts: First, we show that the matrix $\bar{W}$ can be approximated by a hierarchically sparse matrix even under basis mismatch. Then, we provide conditions in terms of number of subcarriers (O_τ) and number of antennas (O_θ) that should be considered in order to achieve (approximate) recovery of W for a single user.

A. Sparse Approximation Analysis

As has been argued in the introduction, the matrix $W(t), t \in [T]$ is not exactly sparse in general. In order to get a concrete recovery result for the application of the HiHTP algorithm for this case as well, we need to quantify how well it can be approximated as (hierarchically) sparse. In order to do this,

we first analyze the sparse approximation error for the vectors u_{τ_p} and $u_{\theta_p}, p \in [L]$ appearing in (5). For simplicity, the case $T = 1$ will be considered with the results trivially extending to the multiple measurements case.

The vector u_θ, for θ arbitrary, is easily seen to be equal to $(1/M)F_M^H a(\theta)$. As shown in [5], the i-th element of u_θ equals

$$(u_\theta)_i = \frac{\sin(\pi M \xi_i(\theta))}{M \sin(\pi \xi_i(\theta))} e^{-j\pi(M-1)\xi_i(\theta)}, i \in [M], \quad (15)$$

where $\xi_i(\theta) := \theta - i/M$ is a translation of the angle θ. In the exact same fashion and for any value of τ, it holds

$$(u_\tau)_i = \frac{\sin(\pi N \eta_i(\tau))}{N \sin(\pi \eta_i)} e^{-j\pi(N-1)\eta_i(\tau)}, i \in [N], \quad (16)$$

where $\eta_i(\tau) := \tau - i/N$. The sparse approximation properties of u_τ and u_θ are identified in the following lemma.

Lemma 1. *For any value of θ, there exists a $(2K_\theta+1)$-sparse vector $u_{\theta,K_\theta} \in \mathbb{C}^M$, with integer $K_\theta \geq 1$ independent of θ and M, such that*

$$\|u_{\theta,K_\theta} - u_\theta\|_2 \lesssim \frac{1}{\sqrt{K_\theta}}.$$

Similarly, for any value of τ, there exists a $(2K_\tau+1)$-sparse vector $u_{\tau,K_\tau} \in \mathbb{C}^D$, with integer $K_\tau \geq 1$ independent of τ and D, with

$$\|u_{\tau,K_\tau} - u_\tau\|_2 \lesssim \frac{1}{\sqrt{K_\tau}}.$$

Proof. The proof is omitted, and deferred to [16]. $\square$

The error estimates derived in Lemma 1 are independent of the design parameters M and N, which means that increasing them results in an increased *relative sparsity* (number of non-zero to total elements) for the sparse approximations of u_θ and u_τ, respectively. However, increasing N requires a proportional bandwith increase, which is not always available. On the other hand, there are no such limitation on increasing M, which can safely be assumed arbitrarily large.

We now use Lemma 1 to quantify the error of approximating $W(t)$ as having a hierarchical sparse structure. Before proceeding, we will assume the following two properties for the delay/angle pairs of the channel paths. The first property imposes a separation among distinct path angles.

Condition 1 (Angular separation)**.** *For any two distinct path angles θ_i, θ_j ($\theta_i \neq \theta_j$) in the set of delay/angle pairs, it holds*

$$|\theta_i - \theta_j| \geq 2K_\theta/M,$$

whereby $|\cdot|$ is meant in a wrap-around sense over the interval $[0, 2\pi)$.

Note that angular separations is a standard assumption in this area of research, see for instance [17], [10], [9] and can be safely assumed to hold for sufficiently large M.

The second assumption essentially excludes the possibility of a channel with excessively many paths having the same angle.

Condition 2 (Limited delays-per-angle). For each distinct angle θ_i in the set of paths angle/delay values , there exists at most K delays τ_k such that (τ_k, θ_i) is in the set of delay/angle pairs.

The delays-per-angle condition is furthermore very reasonable on physical grounds: multipath components with different delays travel over different paths, hence the probability of arriving at the ULA with the same angle is very small. By the same argument one excepts that a choice $K = 1$ is reasonable for typical propagation conditions.

We can now characterize the error of approximating $W(t)$ (for any t) by a matrix with a hierarchically sparse structure.

Proposition 1. *Consider an arbitrary slot $t \in [T]$. Under the angular separation and delays-per-angle conditions, there exists a matrix $W_\Omega(t) \in \mathbb{C}^{D \times M}$ whose vectorized version is $(L(2K_\theta + 1), K(2K_\tau + 1))$-sparse and satisfies*

$$\|W(t) - W_\Omega(t)\| \lesssim \left(K_\theta^{-1} + K_\tau^{-1}\right) \sum_{p=0}^{L} \rho_p(t_i)$$

Proof. Please see Appendix A. □

Clearly, $W(t), t \in [T]$, can be well approximated by a hierarchical sparse matrix $W_\Omega(t)$ by choosing K_θ and K_τ sufficiently large. This, in turn, suggests incorporation of the HiHTP algorithm, treating $W(t), t \in [T]$ as $(L(2K_\theta + 1), K(2K_\tau + 1))$-sparse.

B. Error/overhead tradeoff of HiHTP algorithm

We now want to identify the tradeoff between estimation accuracy and number of subcarriers and antennas that should be considered in order to achieve it. The following result provides a rigorous characterization of the HiHTP recovery guarantees, taking into account the probabilistic nature of the sampling matrices and the error of approximating $\bar{W}(t)$ as hierarchically sparse.

Theorem 3. *Let the number of sampled antennas and subcarriers satisfy*

$$O_\theta \gtrsim \delta^{-2} L(2K_\theta + 1) \log^4(M),$$
$$O_\tau \gtrsim \delta^{-2} K(2K_\tau + 1) \log^4(N),$$

respectively, for some $\delta < 1/\sqrt{3}$. Then, with a probability larger than $1 - M^{-\log^3(M)} - N^{-\log^3(N)}$, the sequence of estimates $\hat{\bar{W}}^{(i)}$ generated by the HiHTP algorithm (Algorithm 1) treating $\text{vec}(\bar{W})$ as $(L(2K_\theta + 1), K(2K_\tau + 1), T)$-sparse, will obey the error bound

$$\|\bar{W} - \hat{\bar{W}}^{(i+1)}\| \leq \kappa^i \|\bar{W} - \hat{\bar{W}}^{(i)}\| + \tau \|Z\|$$
$$+ C(K_\theta^{-1} + K_\tau^{-1}) \sum_{i=1}^{T} \sum_{p=0}^{L} \rho_p(t_i)$$

where ρ, τ are as in Theorem 1, and C is a universal constant.

Proof. Please see Appendix B. □

VI. Conclusion

In this paper, we explored hierarchical sparse estimation framework for massive MIMO channel estimation. The framework can be used to design appropriate algorithms exploiting the sparse nature of massive MIMO channel in joint angular and delay domain. In numerical simulations we show the benefit in terms of reduced pilot overhead, particular for small overhead numbers. We also extend our analysis to the general 'off-grid' case and derive the sufficient pilot overhead scaling for perfect recovery for large signal dimension.

APPENDIX A
PROOF OF PROPOSITION 1

Let us drop the index t_i. For each delay/angle pair (τ_p, θ_p), we define $\imath_p$ and $\jmath_p$ through $\imath_p := \text{argmin}_i |i/QN - \tau_p|, \jmath_p := \text{argmin}_j |j/M - \theta_p|$, and a rectangle $\Omega_p := [\imath_p - K_\theta, \imath_p + K_\theta] \times [\jmath_p - K_\tau, \jmath_p + K_\tau] := I_p \times J_p \subseteq [M] \times [QN]$. Finally define Ω through

$$\Omega = \bigcup_{p=0}^{L} \Omega_p.$$

Due to the angular separation and delay-per angle condition, Ω is an $(L(2K_\theta + 1), K(2K_\tau + 1))$-sparse support (see Figure 5.)

Now we estimate

$$\|W(t_i) - W_\Omega(t_i)\| \leq \sum_{p=0}^{L} \rho_p(t_i) \| u_{\tau_p} u_{\theta_p}^H - \left(u_{\tau_p} u_{\theta_p}^H\right)_{\Omega_p} \|.$$

For each p, we estimate

$$\| u_{\tau_p} u_{\theta_p}^H - (u_{\tau_p}(u_{\theta_p})^H)_{\Omega_p} \|$$
$$\leq \| u_{\tau_p} u_{\theta_p}^H - (u_{\tau_p})_{J_p}(u_{\theta_p})^H \|$$
$$+ \| (u_{\tau_p})_{J_p}(u_{\theta_p})^H - (u_{\tau_p})_{I_p}(u_{\theta_p})_{I_p}^H \|$$
$$= \| u_{\tau_p} - (u_{\tau_p})_{J_p} \| \| u_{\theta_p}^H \|$$
$$+ \| (u_{\tau_p})_{J_p} \| \| (u_{\theta_p})^H - (u_{\theta_p})_{I_p})^H \|$$
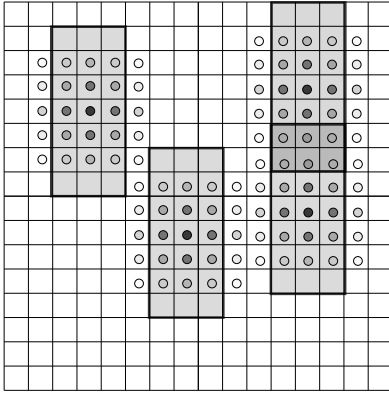$$\lesssim K_\theta^{-1} + K_\tau^{-1}.$$

Fig. 5. The angular separation condition prevents the rectangles Ω_p to intersect when they are associated to different values of θ. It does not prevent intersection for different values of τ, but this does not prevent $K(K_\tau + 1)$-sparsity in each θ-block.

We used that $\|u_\tau\| = \|u_\theta\| = 1$ for each τ and θ, together with Lemma 1. The claim follows.

APPENDIX B
PROOF OF THEOREM 3

Let Ω be the $(L(2K_\theta + 1), K(2K_\tau + 1))$-sparse support set defined in Proposition 1. Define the set $\widehat{\Omega} = [0, T - 1] \times \Omega$. $\widehat{\Omega}$ is then $(T, L(2K_\theta + 1), K(2K_\tau + 1))$-sparse, and further

$$\|\bar{W} - W_{\widehat{\Omega}}\| \lesssim (K_\theta^{-1} + K_\tau^{-1}) \sum_{i=1}^{T} \sum_{p=0}^{L} \rho_p(t_i). \quad (17)$$

Now, we utilize that the matrices A_θ and A_τ are quadratic DFT-matrices (set $D = N$). Applying Theorem 12.31 from [13, p.405] together with the assumptions on O_θ and O_τ yields that

- with probability larger than $1 - M^{-\log^3(M)}$

$$\delta_{3L(2K_\theta + 1)} \left(\frac{1}{\sqrt{O_\theta}} P_\theta A_\theta^* \right) \leq \widetilde{\delta},$$

- with probability larger than $1 - N^{-\log^3(N)}$

$$\delta_{3K(2K_\tau + 1)} \left(\frac{1}{\sqrt{O_\tau}} P_\tau A_\tau \right) \leq \widetilde{\delta},$$

where $\widetilde{\delta}$ is defined through $\delta = \widetilde{\delta}(2 + \widetilde{\delta})$, which in particular implies that $\widetilde{\delta} \leq \delta/2 \leq 1/(2\sqrt{3})$, so that

$$\delta \gtrsim \widetilde{\delta}.$$

Hence, an estimate of the form $\gtrsim \delta^{-2}$ implies an estimate of the form $\gtrsim \widetilde{\delta}^{-2}$ (with another constant), whence the theorem is applicable.

Since further I_T trivially obeys $\delta_T = 0$, Theorem 2 implies that

$$\delta_{(3 \cdot L(2K_\theta + 1), 3 \cdot (L_2(2K_\tau + 1), T)}(\Psi) \leq (1 + \delta)^2 - 1 = \delta(\delta + 2)$$
$$\leq \frac{1}{\sqrt{3}}.$$

Theorem 1 implies that

$$\|W_{\widehat{\Omega}} - W_{\widehat{\Omega}}^{(t+1)}\| \leq \rho^t \|W_{\widehat{\Omega}} - W_{\widehat{\Omega}}^{(t)}\| + \tau \|Z\|.$$

The claim now follows from

$$\|W - W^{(r)}\| \leq \|W_{\widehat{\Omega}} - W^{(r)}\| + \|W - W_{\widehat{\Omega}}\|$$

for $r = 0$ and $r = (t + 1)$ and the error bound (17).

REFERENCES

[1] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[2] G. Wunder, P. Jung, and M. Ramadan, "Compressive Random Access Using A Common Overloaded Control Channel," in *IEEE Global Communications Conference (Globecom'14) – Workshop on 5G & Beyond*, San Diego, USA, December 2015. [Online]. Available: www.arxiv.com/1504.05318

[3] G. Wunder, I. Roth, R. Fritschek, and J. Eisert, "HiHTP: A Custom-Tailored Hierarchical Sparse Detector for Massive MTC," in *Proc. Asilomar Conference on Signals, Systems, and Computers (Asilomar'17)*, Pacific Grove, USA, November 2017.

[4] S. Haghighatshoar and G. Caire, "Massive mimo channel subspace estimation from low-dimensional projections," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 303–318, 2017.

[5] Z. Chen and C. Yang, "Pilot decontamination in wideband massive mimo systems by exploiting channel sparsity," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 5087–5100, July 2016.

[6] S. Haghighatshoar and G. Caire, "Massive MIMO Pilot Decontamination and Channel Interpolation via Wideband Sparse Channel Estimation," *IEEE Trans. on Wireless Communications*, Januar 2017, submitted. [Online]. Available: arXivpreprintarXiv:1702.07207

[7] L. You, X. Gao, A. L. Swindlehurst, and W. Zhong, "Channel acquisition for massive mimo-ofdm with adjustable phase shift pilots," *IEEE Transactions on Signal Processing*, vol. 64, no. 6, pp. 1461–1476, March 2016.

[8] I. Roth, M. Kliesch, G. Wunder, and J. Eisert, "Reliable recovery of hierarchically sparse signals and application in machine-type communications," *IEEE Trans. on Signal Processing*, December 2016, submitted. [Online]. Available: https://arxiv.org/abs/1612.07806v2

[9] R. Heckel, V. I. Morgenshtern, and M. Soltanolkotabi, "Super-resolution radar," *CoRR*, vol. abs/1411.6272, 2014. [Online]. Available: http://arxiv.org/abs/1411.6272

[10] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *IEEE Trans. Sign. Proc.*, vol. 62, no. 21, pp. 5565–5576, 2014.

[11] M. Barzegar, G. Caire, A. Flinth, S. Haghighatshoar, G. Kutyniok, and G. Wunder, "Estimation of angles of arrival through superresolution a soft recovery approach for general antenna geometries," *arXiv*, 2017. [Online]. Available: https://export.arxiv.org/pdf/1711.03996

[12] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, 2011.

[13] S. Foucart and H. Rauhut, *A mathematical introduction to Compressed Sensing*. Birkhäuser, 2013.

[14] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Th.*, vol. 56, no. 4, pp. 1982–2001, April 2010.

[15] S. Jokar and V. Mehrmann, "Sparse solutions to underdetermined kronecker product systems." *Lin. Alg. Appl.*, vol. 431, no. 12, pp. 2437–2447, 2009.

[16] Gerhard Wunder et al, "Hierarchical sparse channel estimation and pilot decontamination for massive mimo," *to appear on arXiv*, 2018.

[17] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Comm. Pur. Appl. Math.*, vol. 67, no. 6, pp. 906–956, 2014.