FDD Massive MIMO Downlink Channel Estimation via Selective Sparse Coding over AoA/AoD Cluster Dictionaries

Mahmoud Nazzal
Istanbul Medipol University
Email: mahmoud.nazzal@ieee.org

Haji Muhammad Furqan
Istanbul Medipol University
Email: hamadni@st.medipol.edu.tr

Huseyin Arslan
University of South Florida
Email: arslan@usf.edu

***Abstract*—Sparse coding over a redundant dictionary has recently been used as a framework for downlink channel estimation in frequency division duplex massive multiple-input multiple-output antenna systems. This usage allows for efficiently reducing the inherently high training and feedback overheads. We present an algorithm for downlink channel estimation via selective sparse coding over multiple cluster dictionaries. A channel training set is divided into clusters based on the angle of the arrival/departure of the majority physical subpaths corresponding to each channel tap. Then, a compact dictionary is trained in each cluster. Channel estimation is done by first identifying the channel cluster and then using its dictionary for reconstruction. This selective sparse coding allows for adaptive regularization via sparse model selection, thereby offering additional regularization to the ill-posed channel estimation problem. We empirically validate the selectivity of the cluster dictionaries. Simulation results show the advantage of the proposed algorithm in achieving better estimation quality at lower computational cost, as compared the case of using standard sparse coding.**

I. Introduction

Massive multiple-input-multiple-output (MIMO) is reported as a key enabler for the fifth generation (5G) communication standard. However, reaping the advantages of massive MIMO requires the knowledge of the channel impulse response (CIR). This can be achieved either by frequency division duplex (FDD) or time division duplex (TDD). FDD has several advantages over TDD. Still, its underlying training and feedback overhead forms the bottleneck against utilizing such advantages [1].

A massive MIMO channel is known to have correlations [2]. Thus, it can be sparsely represented with a few low-dimensional measurements. In a compressive sensing context, this suggests sub-Nyquist channel sampling and reduced-dimensional processing. Consequently, the number of training pilots becomes proportional to the assumed sparsity, rather than the number of antennas. Besides, CIR estimation becomes a sparse recovery problem [3] where sparsity is exploited as a natural regularizer.

The early works utilizing channel sparsity considered the discrete Fourier transform (DFT) as a sparsifying basis [4], [5], [6]. However, channel sparsity with a DFT basis is valid only under the conditions of extremely poor scattering and infinitely many transmitting antennas at the base station [2]. Subsequently, Ding and Rao proposed a dictionary learning channel model (DLCM) [7], [8], [9] where a sparsifying dictionary is obtained by training. Despite efficiently using sparsity as a regularizer, DLCM [7], [8], [9] does not consider discriminative channel properties such as spatial directionality characterized by the angle of arrival/departure (AoA/AoD) [10], [11].

This paper presents an algorithm for FDD massive MIMO downlink CIR estimation based on selective sparse coding over cluster dictionaries. We divide training data into several clusters based on the AoA/AoD of their respective physical subpaths, and train a compact dictionary for each cluster. The result is improved CIR estimation with reduced computational complexity. We show that each dictionary is well-suited for reconstructing the CIRs of its own cluster, exclusively. Besides, sparsity minimization is empirically shown to point to the best cluster. Experiments validate a performance improvement in terms of the normalized mean-squared error (NMSE) measure.

Notation: Lower-case plain, lower-case bold-faced and upper-case bold-faced letters represent scalars, vectors and matrices, respectively. $\|.\|_2$ and $\|.\|_0$ represent the 2-norm and the number of nonzero elements, respectively.

II. Background and Related Work

A. System Model

This work considers a single-cell FDD massive MIMO system. The base station (BS) has a uniform linear antenna array (ULA) of N antennas serving a single-antenna user equipment (UE) as illustrated in Fig. 1. The downlink channel is a narrow-band block flat-fading channel, and its CIR is denoted by $\boldsymbol{h} \in \mathbb{C}^N$. We model $\boldsymbol{h}$ using the geometry-based stochastic channel model (GSCM) [12], where the channel measurement at the BS consists of the effects of signals from both; far scatterers in the cell and local scatterers around the UE.

978-1-5386-6009-6/18/$31.00 ©2018 IEEE

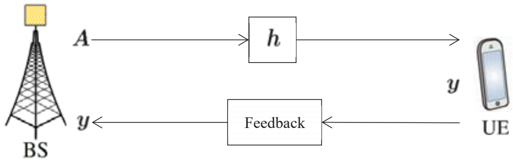

Fig. 1. Pilot downlink and feedback setup for CIR estimation.

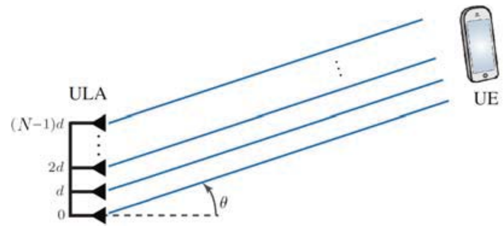

Fig. 2. A top view showing the AoA/AoD for a ULA [1].

The CIR $\boldsymbol{h}$ can be modeled as follows [13].

$$\boldsymbol{h} = \sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \boldsymbol{\alpha}_{il} \boldsymbol{\beta}(\theta_{il}), \tag{1}$$

where $\boldsymbol{\alpha}_{il}$ is the complex gain of the l-th subpath in the i-th scattering cluster, N_c is the number of clusters, and N_s is the number of subpaths in each cluster. The symbol θ_{il} denotes the AoA/AoD of the l-th subpath in the i-th scattering cluster, as depicted in Fig. 2. The steering vector $\boldsymbol{\beta}(\theta_{il})$ represents the normalized array response at the UE.

For a ULA, $\boldsymbol{\beta}(\theta_{il})$ can be modeled as follows [14].

$$\boldsymbol{\beta}(\theta_{il}) = \frac{1}{\sqrt{N}}[1,\ e^{jc\sin(\theta_{il})},\ \ldots\ , e^{jc\sin(\theta_{il})(N-1)}]^{tr}, \tag{2}$$

where $c = 2\pi\frac{d}{\lambda_d}$, with d and λ_d denoting the antenna spacing and the propagation wavelength, respectively, and tr denoting the transpose operator.

Fig. 1 shows the setup for downlink CIR estimation. The BS transmits training pilots to the UE through $\boldsymbol{h}$. Each pilot is a vector in $\mathbb{C}^{1\times N}$. The BS sends T pilots, where T is referred to as the pilot period. So the pilot matrix is $\boldsymbol{A} \in \mathbb{C}^{T\times N}$. The signal received at the UE is.

$$\boldsymbol{y} = \sqrt{\rho}\boldsymbol{A}\boldsymbol{h} + \boldsymbol{n}, \tag{3}$$

where ρ is the signal power and $\boldsymbol{n}$ is additive white Gaussian noise. Then, the UE feeds back $\boldsymbol{y}$ to the BS through the uplink channel. To this end, the BS estimates $\boldsymbol{h}$ based on $\boldsymbol{A}$ and $\boldsymbol{y}$. While classical solutions such as least-squares lack robustness due to insufficient priors, sparsity is shown to be an efficient regularizer.

B. CIR Estimation as a Sparse Recovery Problem

If a signal $\boldsymbol{x} \in \mathbb{R}^N$ admits sparse coding over a dictionary $\boldsymbol{D} \in \mathbb{R}^{N\times K}$, then $\boldsymbol{x} \approx \boldsymbol{D}\boldsymbol{w}$, where $\boldsymbol{w} \in \mathbb{R}^K$ is said to be a sparse coding coefficient vector. For a given $\boldsymbol{x}$ and $\boldsymbol{D}$, $\boldsymbol{w}$ can be obtained through the following sparse coding process.

$$\arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0 \ s.t.\ \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{w}\|_2^2 < \epsilon, \tag{4}$$

where ϵ is the error tolerance.

A collection of predefined basis function such as the DFT can be used as a dictionary. However, learning a redundant dictionary over a set of training data points $\boldsymbol{X} \in R^{N\times m}$ is a better alternative [15]. This is referred to as dictionary learning (DL), formulated as.

$$\arg\min_{\boldsymbol{W},\boldsymbol{D}} \|\boldsymbol{W}_i\|_0 \ s.t.\ \|\boldsymbol{X}_i - \boldsymbol{D}\boldsymbol{W}_i\|_2^2 < \epsilon \ \forall\ i, \tag{5}$$

where i indicates the i-th column in the matrix.

DLCM [7], [8], [9] reformulates CIR estimation as a sparse coding problem. This is based on assuming that $\boldsymbol{h}$ admits a sparse coding over a dictionary $\boldsymbol{D}$ trained over a training set of example CIR realizations $\mathbf{H}$. This means that $\boldsymbol{h} \approx \boldsymbol{D}\boldsymbol{w}$. Since $\boldsymbol{y} = \sqrt{\rho}\boldsymbol{A}\boldsymbol{h}$, then $\boldsymbol{h} - \boldsymbol{D}\boldsymbol{w}$ corresponds to $\sqrt{\rho}\boldsymbol{A}\boldsymbol{h} - \sqrt{\rho}\boldsymbol{A}\boldsymbol{D}\boldsymbol{w} = \boldsymbol{y} - \sqrt{\rho}\boldsymbol{A}\boldsymbol{D}\boldsymbol{w}$. Accordingly, $\boldsymbol{w}$ is calculated based only on $\boldsymbol{y}$, $\boldsymbol{A}$ and $\boldsymbol{D}$ through the sparse recovery process in (6). Finally, a CIR estimate is obtained as $\hat{\boldsymbol{h}} = \boldsymbol{D}\boldsymbol{w}$.

$$\arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0 \ s.t.\ \|\boldsymbol{y} - \sqrt{\rho}\boldsymbol{A}\boldsymbol{D}\boldsymbol{w}\|_2^2 < \epsilon. \tag{6}$$

III. Selective Sparse Coding over AoA/AoD Cluster Dictionaries for CIR Estimation

A. Motivation for Clustered Sparse Coding

The early used DFT basis does not promote sparsity, and its basis vectors have an inherent directional mismatch with channel subpaths. To address these drawbacks, DLCM uses a learned dictionary that promotes sparsity and provides a denser sampling grid thereby relatively reducing the mismatch. Still, a significant mismatch reduction requires a very dense grid, and thus a highly redundant dictionary. In a sparse coding context, high redundancy facilitates the recovery search space. However, this comes at the cost of dramatically increasing the search computational cost and the likelihood of instabilities [16] and degradation [17]. Furthermore, it necessitates using very large training sets for the DL process [18]. To this end, a compact dictionary selected from a set of cluster dictionaries is promising to achieve finer sampling at reduced computational cost compared to a single highly-redundant dictionary.

In the massive MIMO setting, the number of scattering clusters is typically small [12]. Also, the effective subpaths associated with a scattering cluster are likely to concentrate in a small angular spread around the line-of-sight scattering direction [18]. Thus, a compact dictionary dedicated for this directionality would provide a sufficient number of relevant atoms (sampling grid points). This directionality is characterized by the propagation AoA/AoD as shown in (2).

B. The Proposed Algorithm

The proposed algorithm is composed of the following two stages.

Algorithm 1 Cluster DL Stage.

Input: Error tolerance ϵ and the number of clusters M.
Output: Cluster dictioanries $\boldsymbol{D}^i, i = 1, 2, \ldots, M$.

1: Generate $\boldsymbol{H}$, and record the value of θ for each $\boldsymbol{H}_i$.
2: Cluster $\boldsymbol{H}$, into $\boldsymbol{H}^i, i = 1, 2, \ldots, M$ using the recorded angles.
3: **for** $i = 1, 2, \ldots, M$,
   Use a DL algorithm to solve for
   $\underset{\boldsymbol{W}^i, \boldsymbol{D}^i}{\operatorname{argmin}} \|\boldsymbol{W}_k^i\|_0 \ s.t. \ \|\boldsymbol{H}_k^i - \boldsymbol{D}^i \boldsymbol{W}_k^i\|_2^2 < \epsilon \ \forall \ k$
4: **end for**

Algorithm 2 CIR Estimation Stage.

Input: cluster dictionaries $\boldsymbol{D}^i, i = 1, 2, \ldots, M$, training pilots $\boldsymbol{A}$, and error tolerance ϵ.
Output: a CIR estimate $\hat{\boldsymbol{h}}$.

1: Send $\boldsymbol{A}$ over the channel to receive $\boldsymbol{y}$.
2: **for** $i = 1, 2, \ldots, M$,
   Solve $\underset{\boldsymbol{w}^i}{\operatorname{argmin}} \|\boldsymbol{w}^i\|_0 \ s.t. \ \|\boldsymbol{y} - \sqrt{\rho} \boldsymbol{A} \boldsymbol{D}^i \boldsymbol{w}^i\|_2^2 < \epsilon$
3: **end for**
4: Identify the cluster $s = \underset{i}{\operatorname{argmin}} \|\boldsymbol{w}^i\|_0$
5: Reconstruct: $\hat{\boldsymbol{h}} = \boldsymbol{D}^s \boldsymbol{w}^s$

1) Training Stage: This stage trains for M cluster dictionaries $\boldsymbol{D}^i, i = 1, 2, \ldots, M$, where the superscript denotes the cluster index. The AoA/AoD is used as a clustering criterion to split a training set $\boldsymbol{H}$ into cluster datasets $\boldsymbol{H}^i, i = 1, 2, \ldots, M$. Then, a compact dictionary is trained for each cluster over its own data $\boldsymbol{H}^i$ using any standard DL algorithm. This stage is illustrated in Algorithm 1. It is noted that the number of clusters and their bounds can be empirically set. In this context, one may balance the trade-off between dictionary selectivity and the accuracy of model selection.

2) Testing Stage: This stage uses the fed-back received signal $\boldsymbol{y}$ along with cluster dictionaries to obtain a CIR estimate $\hat{\boldsymbol{h}}$. First, model selection is applied to identify the correct channel cluster, as in (7). Motivated by the selectivity of the cluster dictionaries, it is intuitively expected that the most appropriate model is the one that yields the sparsest solution. This means that one can select the dictionary that minimizes the sparsity. The proposed testing stage is outlined in Algorithm 2.

$$\begin{gathered} \underset{\boldsymbol{w}^i}{\arg\min} \|\boldsymbol{w}^i\|_0 \ s.t. \ \|\boldsymbol{y} - \sqrt{\rho} \boldsymbol{A} \boldsymbol{D}^i \boldsymbol{w}^i\|_2^2 < \epsilon \\ s = \underset{i}{\arg\min} \|\boldsymbol{w}^i\|_0. \end{gathered} \tag{7}$$

C. CIR Estimation Error with Cluster Dictionaries

It is interesting to analyze the impact of the proposed cluster regularization on the CIR estimation error. Let δ denote the maximum mismatch between the sine function of the estimated θ_{il}, and that of the true θ_{il}^t, so $\sin\theta_{il} = \sin\theta_{il}^t + \delta$. For simplicity, let us assume unity complex gains $\boldsymbol{\alpha}_{il}$, and perfect model selection. Recalling (1), the error between an estimate $\boldsymbol{h}$ and the true $\boldsymbol{h}^t$ can be expressed as follows.

$$e = ||\boldsymbol{h}^t - \boldsymbol{h}||_2^2 = ||\sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \boldsymbol{\beta}(\theta_{il}^t) - \boldsymbol{\beta}(\theta_{il})||_2^2. \tag{8}$$

Let us define the following difference vector.

$$\boldsymbol{d}(\theta_{il}) = \boldsymbol{\beta}(\theta_{il}^t) - \boldsymbol{\beta}(\theta_{il}) \tag{9}$$

Using (2), the k-th element in $\boldsymbol{d}(\theta_{il})$ is as follows.

$$\boldsymbol{d}_k(\theta_{il}) = e^{jc(k-1)\sin(\theta_{il}^t)} - e^{jc(k-1)\sin(\theta_{il})} \tag{10}$$

For simplicity, let us denote the quantity $c(k-1)$ by x. With some simplification, (10) reduces to (11).

$$\begin{aligned} \boldsymbol{d}_k(\theta_{il}) &= \cos(x \ \sin \ \theta_{il}^t) - \cos(x \ \sin \ \theta_{il}^t + x\delta) \\ &+ j(\sin(x \ \sin \ \theta_{il}^t) - \sin(x \ \sin \ \theta_{il}^t + x\delta)). \end{aligned} \tag{11}$$

Doing a few further trigonometric and algebraic steps, the square of the k-th element in $\boldsymbol{d}(\theta_{il})$ is as follows.

$$\boldsymbol{d}_k(\theta_{il})^2 = 2 \ \sin(\frac{x\delta}{2}). \tag{12}$$

The energy of $\boldsymbol{d}(\theta_{il})$ is thus the summation of the terms in (12). Using the triangle inequality, we can write.

$$||\boldsymbol{d}(\theta_{il})||^2 \leq ||\sum_{k=1}^{N-1} \boldsymbol{d}_k(\theta_{il})||^2 \tag{13}$$

From (8) and (13), we can write.

$$e = ||\sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \boldsymbol{d}(\theta_{il})||^2 \leq \sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \sum_{k=1}^{N-1} \boldsymbol{d}_k(\theta_{il})^2 \tag{14}$$

Substituting (13) into (14) reveals.

$$e \leq \sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \sum_{k=1}^{N-1} 2 \ \sin(\frac{c(k-1)\delta}{2}) \tag{15}$$

From (15), it is clear that the error upper bound is directly dependent on the sine function error δ which is equal to 2 for the case of standard reconstruction where as it is $2/M$ for the case of using M clusters.

D. Computational Complexity Discussion

Sparse coding forms the bottle neck in CIR estimation computational cost. So, we can roughly model this cost in terms of that of sparse coding. Considering basis pursuit denoising (BPDN) [19] as an example sparse coding technique, its computational cost working on $\boldsymbol{D} \in \mathbb{C}^{N \times K}$ is approximately $\mathcal{O}((NK)^3)$ [20]. If one employs M dictionaries, each being γ times smaller than the standard universal dictionary, the overall computation of CIR estimation is approximately $\mathcal{O}(\frac{M(NK)^3}{\gamma^3})$. Therefore, the computational complexity will be reduced by a factor of $(\frac{M}{\gamma^3})$. The same argument holds for the computational complexity of DL. This is because the computational burden of the DL process is mainly caused by sparse coding.

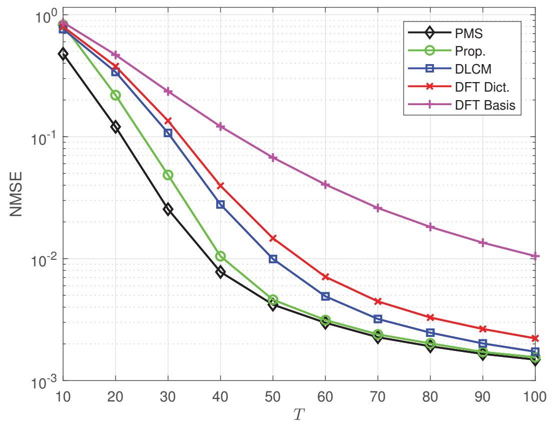

Fig. 3. NMSE of downlink channel estimation versus training period.

IV. Experimental Validation

We compare the proposed algorithm to the DLCM algorithm with: a learned dictionary, an overcomplete DFT dictionary, and a DFT basis. These are denoted by (Prop.), (DLCM), (DFT Dict.), and (DFT Basis), respectively. We also include the proposed algorithm with perfect model selection (denoted by PMS). This is the case where a CIR estimate is obtained using each dictionary, and the best estimate to approximate the ground-truth CIR is chosen. This scenario is impractical, and is included only to investigate the impact of model selection on the performance of the proposed algorithm.

We adopt the experimental setup of the DLCM algorithm reported in [9]. This contains a single urban macro cell with a radius of 1200 meters, centered at the BS. The BS has a ULA of N=100 antennas and the UE has one antenna. The principles of the GSCM channel model [12] are used to generate the channel coefficients for training and testing. The channel parameters are set according to the spatial channel model [21]. The azimuth angle θ ranges between $-\pi/2$ and $\pi/2$. The cell contains seven fixed-location scattering clusters. The locations of these clusters are randomly selected to range between 300 meters and 800 meters at the beginning of the simulation, and are kept unchanged afterwards. Each channel is modeled using four scattering clusters; one is at the UE location, and the remaining three are the closest to the user from the aforementioned seven scattering clusters. The UE location is drawn uniformly to be between 500 meters and 1200 meters. Each scattering cluster has 20 effective subpaths with a 4-degree angular spread. We generate 10^4 downlink CIR realizations for the DL processes. As done in [9], we use k-svd [15] and BPDN [19] for DL and sparse coding, respectively. The signal-to-noise ratio is 30 dB.

The proposed algorithm defines 8 AoA/AoD clusters C_1 through C_8 with θ bounds of -90°,-67.8°, -35.5°, -16.8°, 0°, 16.8°, 35.5°, 67.8°, and 90°, respectively.

TABLE I
Cluster dictionary reconstruction selectivity in the NMSE sense. The best two estimates are in bold-face.

	Cluster Dictionary								D^u
Set	D^1	D^2	D^3	D^4	D^5	D^6	D^7	D^8	
H^1	**0.0037**	0.4027	2.5971	2.1432	2.4528	7.1593	11.4142	2.6698	**0.0161**
H^2	0.1952	**0.0091**	1.1032	1.5056	1.8737	7.1348	11.6273	3.119	**0.0188**
H^3	1.4208	1.1108	**0.017**	0.1793	0.9362	5.0664	8.4372	3.1578	**0.0449**
H^4	1.9346	2.8687	0.3600	**0.0226**	0.1888	3.3529	6.4138	3.0551	**0.0490**
H^5	2.8546	6.5632	3.1024	0.1736	**0.0215**	0.3006	2.7644	2.0817	**0.0464**
H^6	2.8726	8.3330	4.5589	0.7382	0.2293	**0.0152**	1.2721	1.5800	**0.0439**
H^7	2.8225	11.1688	6.300	1.6675	1.6574	1.4051	**0.0071**	0.2015	**0.0198**
H^8	2.4423	11.2631	6.793	2.3085	2.4536	3.2952	0.3578	**0.0037**	**0.0153**

These bound are chosen to quantize the trigonometric sine function range of -1 to 1 into 8 fair ranges. Then, a compact 100×100 dictionary is trained for each cluster, following the steps in Algorithm 1. The DLCM algorithm uses a single 100×400 dictionary.

With the above specifications, we randomly generate a test set of 10^3 CIR vectors for the testing part of this experiment. For each test CIR, we generate random pilots $\boldsymbol{A}$ with periods T of 10, 20, …, 100. For each T value, a channel estimate $\hat{\boldsymbol{h}}$ is obtained via the aforementioned methods and compared to the true CIR in the NMSE sense. Then, we average the NMSE values. The results of this experiment are presented in Fig. 3.

In view of Fig. 3, the advantage of basis redundancy is clear as the over-complete DFT is superior to the orthogonal DFT. Moreover, DLCM is consistently superior to the two DFT scenarios indicating the the advantage of a learned dictionary over predefined bases. Besides, the proposed algorithm is superior to DLCM, indicating the added benefit of selective sparse coding. Moreover, the proposed algorithm with actual model selection coincides with its pms variant except for small T values, where model selection is less accurate.

The selectivity of a cluster dictionary is seen in its suitability to exclusively reconstruct the channels in its cluster. To investigate the selectivity of the designed dictionaries, the following experiment is conducted. For each cluster, we randomly select 10^3 training CIR vectors as a cluster testing set $\boldsymbol{H}^i, i = 1$ through 8. For each testing CIR vector $\boldsymbol{h}$, we generate random pilots $\boldsymbol{A}$ and send them over this channel. Next, based on received signal $\boldsymbol{y}$, we obtain a channel estimate $\hat{\boldsymbol{h}}$ using each of $\boldsymbol{D}^i, i = 1, 2, \ldots 8$, and the DLCM dictionary $\boldsymbol{D}^u$. We calculate the NMSE between the ground-truth CIR vector, and each of these reconstructions. Finally, we average the reconstruction NMSE for the 1000 CIR vectors of each cluster. The results are listed in Table I.

In view of Table I, one can make the following conclusions. First, the designed cluster dictionaries are selective as the correct cluster dictionary results in the best reconstruction of its cluster channels. Second, using the "*best*" dictionary is consistently better than using $\boldsymbol{D}^u$. This indicates the advantage of using cluster dictionaries over a universal dictionary. Third, a given

$\boldsymbol{y}$ can be used to identify the best cluster to reconstruct its underlying channel by selecting the "*best*" dictionary that minimizes the reconstruction NMSE.

TABLE II
AVERAGE SPARSITY OF RECEIVED SIGNALS IN CLUSTERS. THE MINIMAL SPARSITY IS IN BOLD-FACE.

	Average Sparsity with Dictionary							
CIR's in	$\boldsymbol{D}^1$	$\boldsymbol{D}^2$	$\boldsymbol{D}^3$	$\boldsymbol{D}^4$	$\boldsymbol{D}^5$	$\boldsymbol{D}^6$	$\boldsymbol{D}^7$	$\boldsymbol{D}^8$
$\boldsymbol{H}^1$	**31.1**	54.9	78.7	64.8	65.7	89.0	99.6	100
$\boldsymbol{H}^2$	57.1	**34.3**	67.0	62.7	64.4	89.0	99.6	100
$\boldsymbol{H}^3$	99.9	68.3	**42.5**	51.9	60.6	86.7	99.6	100
$\boldsymbol{H}^4$	100	92.1	52.5	**44.1**	52.9	82.5	99.4	100
$\boldsymbol{H}^5$	100	99.3	79.5	53.2	**44.7**	51.4	93.2	100
$\boldsymbol{H}^6$	100	99.3	83.1	59.4	53.3	**41.6**	69.9	100
$\boldsymbol{H}^7$	100	99.3	85.8	62.6	64.2	69.4	**34.5**	56.5
$\boldsymbol{H}^8$	100	99.2	85.9	64.0	66.8	83.3	56.3	**31.2**

It is intuitively expected that the best dictionary is the one that minimizes the sparsity of coding $\boldsymbol{y}$. To investigate this expectation, we repeated the previous experiment calculating the sparsity of the sparse coding vector $\boldsymbol{w}$ over each cluster dictionary. The average sparsity of each test dataset $\boldsymbol{H}^i, i = 1$ through 8 over each dictionary $\boldsymbol{D}^i$ is reported in Table II. It is noted that best cluster dictionary for each given test set, is the one that results in minimal sparsity of coding $\boldsymbol{y}$. This motivates confidently depending on the sparsity of $\boldsymbol{w}$ for model selection.

V. CONCLUSION

This work shows the advantage of clustered sparse coding in FDD massive MIMO downlink channel estimation. The AoA/AoD of effective channel subpaths is used as a clustering criterion. For each cluster, a compact dictionary is trained. The designed cluster dictionaries are shown to be selective to channels in their clusters. The minimal sparsity measure is employed as a model selection criterion. We analytically show that using several compact dictionaries leads to reducing the maximal channel reconstruction error, as compared to the case of standard sparse coding. The proposed algorithm is shown to improve the channel estimation quality. It is also shown to reduce the computational complexity of the underlying sparse coding and DL processes, owing to the compactness of the dictionaries.

ACKNOWLEDGMENT

During this work, Dr. Mahmoud Nazzal was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the 2221 program.

REFERENCES

[1] E. Bjrnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[2] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct 2002.

[3] G. Wunder, H. Boche, T. Strohmer, and P. Jung, "Sparse signal processing concepts for efficient 5g system design," *IEEE Access*, vol. 3, pp. 195–208, 2015.

[4] X. Rao and V. K. N. Lau, "Distributed compressive csit estimation and feedback for fdd multi-user massive mimo systems," *IEEE Trans. Signal Process*, vol. 62, no. 12, pp. 3261–3271, June 2014.

[5] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, June 2010.

[6] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave mimo systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, March 2014.

[7] Y. Ding and B. D. Rao, "Compressed downlink channel estimation based on dictionary learning in FDD massive MIMO systems," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[8] ——, "Channel estimation using joint dictionary learning in FDD massive MIMO systems," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2015, pp. 185–189.

[9] ——, "Dictionary learning based sparse channel representation and estimation for FDD massive MIMO systems," *arXiv preprint arXiv:1612.06553*, 2016.

[10] A. A. Esswie, M. El-Absi, O. A. Dobre, S. Ikki, and T. Kaiser, "Spatial channel estimation-based fdd-mimo interference alignment systems," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 254–257, April 2017.

[11] ——, "A novel fdd massive mimo system based on downlink spatial channel estimation without csit," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[12] A. F. Molisch, A. Kuchar, J. Laurila, K. Hugl, and R. Schmalenberger, "Geometry-based directional model for mobile radio channelsprinciples and implementation," *Transactions on Emerging Telecommunications Technologies*, vol. 14, no. 4, pp. 351–359, 2003.

[13] R. B. Ertel, P. Cardieri, K. W. Sowerby, T. S. Rappaport, and J. H. Reed, "Overview of spatial channel models for antenna array communication systems," *IEEE Pers. Commun.*, vol. 5, no. 1, pp. 10–22, Feb 1998.

[14] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct 2002.

[15] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[16] M. Elad and I. Yavneh, "A plurality of sparse representations is better than the sparsest one alone," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4701–4714, Oct 2009.

[17] M. Protter, I. Yavneh, and M. Elad, "Closed-form MMSE estimation for signal denoising under sparse representation modeling over a unitary dictionary," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3471–3484, July 2010.

[18] J. Dai, A. Liu, and V. K. N. Lau, "Fdd massive mimo channel estimation with arbitrary 2d-array geometry," *IEEE Trans. Signal Process.*, vol. PP, no. 99, pp. 1–1, 2018.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, Jan. 2001.

[20] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug 2005.

[21] "Universal mobile telecommunications system (UMTS) ; spatial channel model for multiple input multiple output (MIMO) simulations 3gpp tr 25.996 version 12.0.0 release 12.)," 2014.