Spotting faults over the spectrum: Fast and online antenna array fault diagnosis for multi-carrier precoding systems

Sreekar Sai Ranganathan, Prajosh K P, Uday K Khankhoje, *Senior Member, IEEE*, Francesco Ferranti, *Senior Member, IEEE*

***Abstract*—We present a new online method to diagnose faulty elements in an antenna array quickly and reliably while a wireless communication system is under regular operation. Existing methods for fault diagnosis do not usually allow simultaneous operation and diagnosis, leading to downtime, which becomes a severe constraint for modern wireless systems. Our approach is based on exploiting an extra degree of freedom available in the frequency domain in modern multicarrier hybrid/digital precoding architectures by allocating a very small part of the available spectrum for diagnosis while the rest is used for regular operation. Measurements can be made in a single shot by using optimized inputs, thereby minimally disrupting regular operation. We address both fully and partially connected RF precoding architectures of hybrid precoders and introduce one and two-stage fault diagnosis procedures, respectively. While the faulty antennas for fully connected hybrid architectures or digital architectures are diagnosed employing ideas from sparse recovery, additional ideas from block-sparse recovery are used in the case of a partially connected architecture. Our results indicate that online array fault diagnosis is possible in both digital and hybrid precoding architectures by introducing simple digital operations into the system and by using a single fixed receiver measurement setup.**

***Index Terms*—Compressed sensing, Fault diagnosis, Antenna arrays, Precoding, OFDM, MIMO communication**

I. Introduction

With the rise of communication technologies such as massive MIMO [1], antenna systems are becoming more and more complex, and arrays with a large number of radiating elements are becoming commonplace. This also means the reliability of these systems comes into question since their performance is dependent on whether each of these antennas is working as expected. Undetected faults due to environmental exposure in a large system could prove costly, necessitating efficient techniques to detect and diagnose faulty elements.

Manuscript received Jul 24, 2022. (*Corresponding author: Uday Khankhoje*.)

S. S. Ranganathan, Prajosh K P, and U. K Khankhoje are with the Department of Electrical Engineering, Indian Institute of Technology (IIT) Madras, Chennai, India. F Ferranti is with the Brussels Photonics group, Department of Applied Physics and Photonics, Vrije Universiteit Brussel and Flanders Make, Pleinlaan 2, B-1050 Brussels, Belgium. F. Ferranti is also an Adjunct Professor at the Department of Electrical Engineering at IIT Madras. (e-mail: ee18b154@smail.iitm.ac.in; ee17d044@smail.iitm.ac.in; uday@ee.iitm.ac.in; Francesco.Ferranti@vub.be.)

Since the number of faulty antennas can be assumed to form a small fraction of the total number of elements in large arrays, compressed sensing-based techniques [2]–[4] for fault diagnosis using a few far-field measurements have been developed, mainly in the context of phased arrays. Later works [5]–[12] were aimed at simplifying the measurement method (e.g., using a fixed receiver probe instead of measurements at multiple locations), and overcoming modelling assumptions.

An important observation here is that while efforts have been made to improve the reliability of fault diagnosis under various conditions, all existing techniques involve making measurements while the system is not in regular operation. However, modern wireless systems such as 5G cannot afford disruptions in services for maintenance to perform diagnosis. Hence, it is imperative to develop techniques for online fault diagnosis while the system is still transmitting data to its users.

This paper proposes a novel fault diagnosis technique to address this problem by exploiting an extra degree of freedom in the frequency domain available in modern multicarrier precoding systems.

A. Motivation

Massive MIMO systems require spatial multiplexing capabilities and therefore employ precoding systems, either digital or hybrid. Hybrid architectures are more practical in the near term due to lower cost and hardware complexity, in contrast to fully digital ones that require dedicated RF chains for each antenna. Further, for mmWave communication, spatial multiplexing precoding algorithms for wideband mmWave systems become important. This involves multicarrier frequency-selective precoding in the digital domain [13]–[18].

These precoding systems open the avenue of digitally leveraging a small part of the spectrum for fault diagnosis. The core idea of our approach is to use a small part of the spectrum for fault diagnosis, while the rest is used for regular operation. This idea is analogous to allocating a part of the spectrum for a user – in this case, for fault diagnosis. This would result in minimal disruption in the services of antenna systems deployed on-field. Fig. 1 provides a visual summary of our fault diagnosis strategy. A single fixed receiver probe is used for measurements, while the spectral input to the RF chains is manipulated in the digital domain (shown in red). The part of the received spectrum used for diagnosis serves as our measurement for fault diagnosis.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3247727

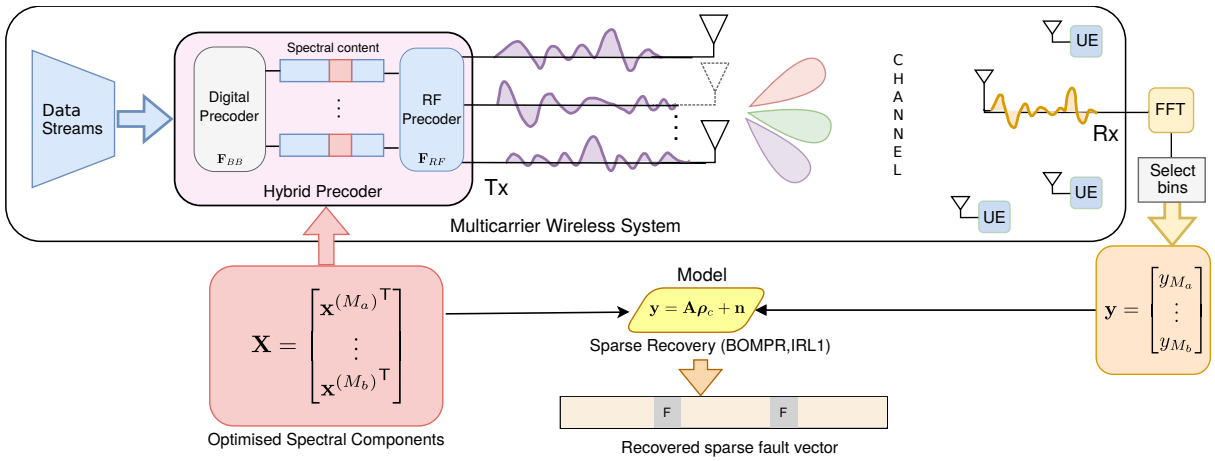


Fig. 1. FAST AND ONLINE ARRAY FAULT DIAGNOSIS - Fault Diagnosis scheme for a multicarrier wireless system being used for data transmission. The spectral content on M ($\{M_a, ..., M_b\}$) out of the K subcarriers on the RF chains is modified by 'injecting' optimized spectral components $\mathbf{X}$, that replace a part of the digitally precoded signals. This causes the antennas to carry a known spectral 'signature,' of which a few (shown dotted) are faulty. Observe that the system continues to serve the UEs through the rest of the spectrum, *while* fault diagnosis measurements are taken. The measurement by the single fixed probe, $\mathbf{y}$ is the corresponding M subcarrier components of the received signal, described by the forward model $\mathbf{A}$, where $\mathbf{n}$ represents noise, and $\boldsymbol{\rho}_c$ the complementary fault vector to be recovered. The engineered excitations $\mathbf{X}$ enable $\mathbf{A}$ (which is a function of $\mathbf{X}$) to be optimized for efficient sparse recovery. These replace only few of the available subcarriers, thereby drastically simplifying measurement without affecting communication services.

It is worth pointing out why this is not feasible using analog (RF)-based precoding systems. Since the RF precoding is frequency-flat, each spectral bin will provide the same projection of the underlying fault state of the antennas, providing only redundant information in the received spectrum. Instead, it will require a sequence of spectral measurements, each with a different RF precoding, to perform fault diagnosis. With digital/hybrid precoders, this is no longer a constraint since each spectral bin can provide a different projection of the underlying fault states, enabling measurements in parallel rather than sequentially.

B. Related Work

Compressed sensing has been used for fault diagnosis by several works [2]–[4], [19], [20], using measurement of the far-field at several points in space. The measurements made by a receiver antenna at each spatial position, when subtracted from the reference measurements (of a healthy, i.e. fault-free array) contains information that is used to recover the positions of faulty antennas using compressed sensing. Later works [5], [7], [21] performed fault diagnosis using a fixed receiver, by varying the excitations to each of the array elements, instead of the spatial location of the receiver, significantly simplifying the measurement process. Optimizing these excitations resulted in further improvements in sparse recovery, first demonstrated in [7], and subsequently generalized to include the effect of mutual coupling between antenna elements [12]. All these methods are targeted at analog precoding-based systems using a single frequency, use sequential measurement strategies for fault diagnosis, and do not address multicarrier precoding-based systems.

The methods proposed in this work employ optimized transmit signals to modify the spectral input to the antennas. This is similar, in principle, to the transmit of optimized sequences in communication, e.g. for peak-to-average-power ratio (PAPR) reduction [22], [23], and precoding techniques [24], [25]. What is different, however, is that our optimization objective is targeted at making our spectral measurement more suitable for efficient sparse recovery.

Our fault diagnosis method follows in line with the strategies of previous methods using reference measurements, and assumes the channel is known. There have also been innovative contributions towards removing the need for channel information or reference measurements of a healthy array [6], [9] and complete diagnosis of hybrid beamforming systems [8]. We do not try to tackle these issues with our approach and instead focus on making a step towards performing online fault diagnosis by exploiting the extra degree of freedom available in modern wireless systems.

C. Contributions

The following are the contributions of this paper:

- Firstly, we open the avenue of performing online fault diagnosis, i.e., performing fault diagnosis while the system is still in operation, allowing for continuous service with minimal disruptions.
- Secondly, by adapting the array fault diagnosis problem formulation to multicarrier systems, we introduce the idea of using a small part of the spectrum for diagnosis while the rest is used for regular operation. To our knowledge, this is the first work to exploit this extra degree of freedom available in modern wireless systems.
- We also introduce the paradigm of *two-stage* diagnosis for systems involving partially connected RF precoders, first identifying faults at the level of subarrays using block-sparsity, followed by pinning down the exact locations of faults within the subarray(s).

Though previous work proposed diagnosis during beam steering [21], it is dependent on the scanning angle used and is targeted at conventional phased array systems. While we

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3247727

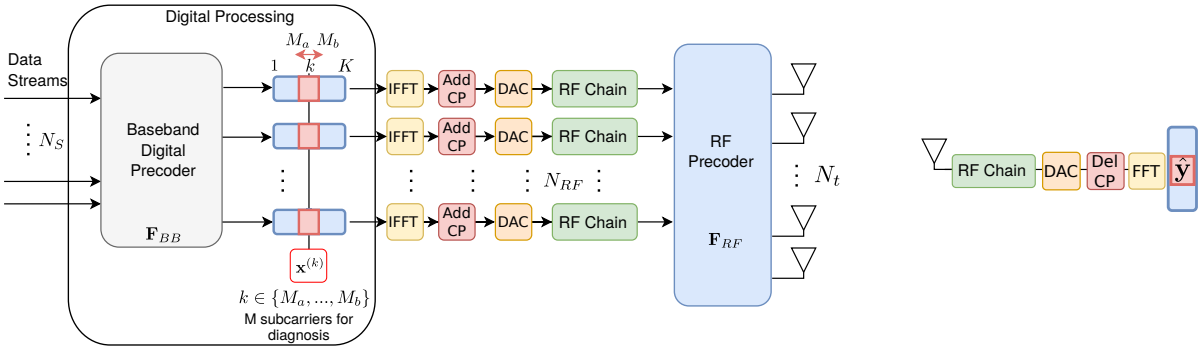


Fig. 2. A block diagram of the OFDM-MIMO based Hybrid Precoder Architecture with new injected input vectors, $\mathbf{x}^{(k)}$, for fault diagnosis. A total of N_S data streams enter the baseband digital precoder, which then feed N_{RF} RF chains, which are finally emitted by N_t broadband antennas after going through the RF precoder. A single receiver used for fault diagnosis is shown on the right.

address hybrid precoders (both partially [26], [27] and fully connected) in detail, our approach trivially works (as a special case) for fully digital architectures.

Our results are limited to multipath-free channels and require reference measurements of a healthy array. We expect that further work on this problem by combining ideas from [6], [9] could create a robust online fault diagnosis technique that can be implemented on-field, for which our work provides a strong base. The rest of this paper is organized as follows. Section II develops the forward measurement model of the diagnosis setup for a multicarrier hybrid precoding system. This is then used to solve for the faulty elements, as described in Section III, and matrix optimization methods described in Section IV. Finally, we show our simulations and analyses in Section V and discuss the outlook of the paper in Section VI.

Notation

$\mathbf{A}$ denotes a matrix, $\mathbf{a}$ a vector and a, A scalars. $\mathbf{A} \odot \mathbf{B}$ is the element-wise Hadamard product. $\mathbf{A}^\mathsf{T}$, $\mathbf{A}^\dagger$ and $\mathbf{A}^*$ denote the transpose, hermitian (conjugate-transpose) and element-wise complex conjugate respectively. $\mathbf{a}[i]$ denotes the ith 'block' of the vector $\mathbf{a} = \left[\mathbf{a}^\mathsf{T}[1]\ \mathbf{a}^\mathsf{T}[2]\ \ldots\ \mathbf{a}^\mathsf{T}[M]\right]^\mathsf{T}$, while $\mathbf{A}[i]$ denotes the ith block of the matrix $\mathbf{A}$, consisting of a contiguous set of columns from $\mathbf{A}$. a_i is the ith element of $\mathbf{a}$. $|\Gamma|$ refers to the size of a set Γ.

II. Fault Diagnosis Measurement Model

The system we consider (Fig. 2) is a base station (BS) antenna array with an orthogonal frequency-division multiplexing (OFDM) based hybrid precoding system, similar to the architecture described in [13], and a single fixed measurement probe.

The BS consists of N_t antennas and N_{RF} RF chains, communicating via N_S data symbol blocks, each of length K. For simplicity, let the number of subcarriers available also be equal to K; in other words, all the subcarriers are being used for transmission. The N_S ($\leq N_{\text{RF}} \leq N_t$) data symbols for the kth subcarrier $\mathbf{s}^{(k)}$ ($k \in \{1, .., K\}$) are first precoded by the $N_{\text{RF}} \times N_S$ subcarrier-dependent digital precoding matrix $\mathbf{F}_{\text{BB}}^{(k)}$, and transformed to the time domain using the K-point inverse fast Fourier transform (IFFT) blocks, and a cyclic prefix added.

The precoding in RF is performed using phase shifters, represented by the $N_t \times N_{\text{RF}}$ matrix $\mathbf{F}_{\text{RF}}$, which has all of its entries either of the form $e^{j\phi_{n,m}}$, with unit magnitude or zero. The RF precoder is the same for all the subcarriers, unlike the digital precoder.

The N_t equivalent complex baseband signals applied at the antennas for subcarrier k, $\mathbf{z}^{(k)}$, can therefore be written as [13]

$$\mathbf{z}^{(k)}_{N_t \times 1} = (\mathbf{F}_{\text{RF}})_{N_t \times N_{\text{RF}}} \underbrace{\left(\mathbf{F}^{(k)}_{\text{BB}}\right)_{N_{\text{RF}} \times N_S} \mathbf{s}^{(k)}}_{=\mathbf{x}^{(k)}}, \tag{1}$$

defining $\mathbf{x}^{(k)}$ to restrict our attention to the stages after the digital precoder.

To represents faults, we define a N_t length fault vector $\boldsymbol{\rho}$, multiplied element-wise with $\mathbf{z}^{(k)}$. An antenna that is nearly dead, for example, would have ρ_i close to zero, and a healthy one would have ρ_i equal to unity. If the $N_t \times 1$ channel vector for the kth subcarrier is represented by $\mathbf{h}^{(k)}$, then the kth subcarrier of the received spectrum [7] can be written as

$$\hat{y}^{(k)} = \mathbf{h}^{(k)^\dagger} \left(\boldsymbol{\rho} \odot \mathbf{z}^{(k)}\right) + \hat{n}^{(k)} \tag{2}$$

where $\hat{n}^{(k)}$ is the complex noise for the kth subcarrier.

A. Measurement using engineered inputs

Let us say that M subcarriers (shown in red in Fig. 2), $\Lambda = \{M_a,..,M_b\}$, out of the K available are used for diagnosis[1] ($M = M_b - M_a + 1$), while the rest ($\{1, ..., K\} - \Lambda$) are used for normal communication. By disengaging the subcarriers M_a to M_b, the block size of data is therefore reduced to $K - M$.

The core strategy of our approach is to replace the digitally precoded vectors $\mathbf{x}^{(k)}$ for the subcarriers M_a to M_b with our own engineered vectors $\mathbf{x}^{(M_a)}$ to $\mathbf{x}^{(M_b)}$. We elaborate on the choice of these vectors later, in Section IV. As we can see from Fig. 2, we are effectively substituting new intermediate 'symbols' for those specific subcarriers. Our 'measurement' in turn, becomes the M 'symbols' obtained on these subcarriers at the receiver probe.

[1]Not necessarily contiguous. These could, for example be one 'resource block' (12 subcarriers) that is allocated to a user.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3247727

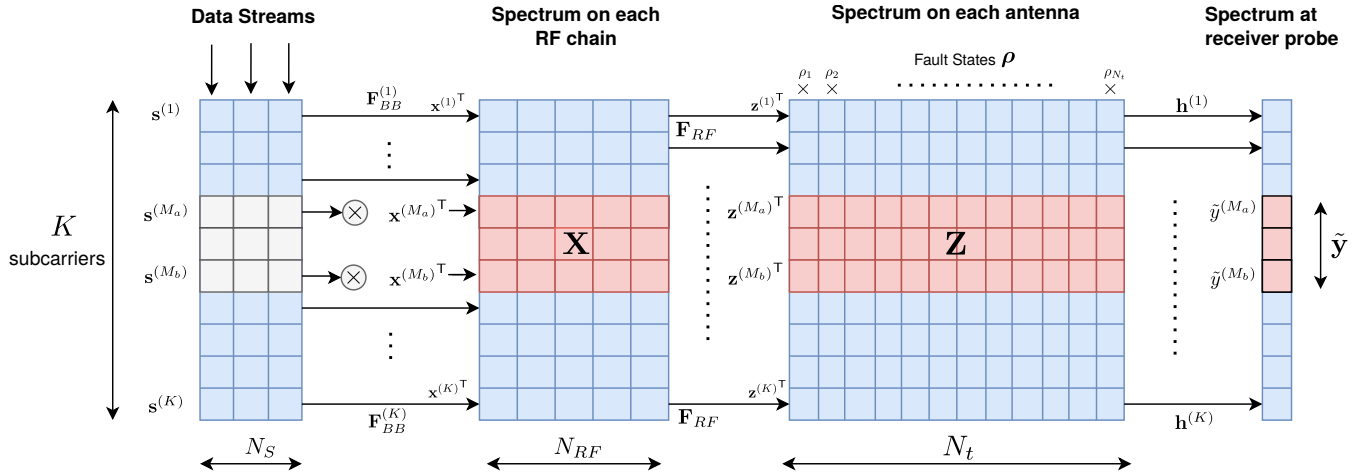


Fig. 3. Schematic showing the matrix-vector operations involved for fault diagnosis. The engineered inputs $\mathbf{x}^{(M_a)}$ to $\mathbf{x}^{(M_b)}$ (shown in red) occupy a small part of the spectrum while usual data transmission occurs on the rest of the subcarriers. The fault state of each antenna acts as a multiplying factor to the spectrum on the corresponding antenna. Therefore, the existence of faults influences the spectrum of the output $\mathbf{y}$, which we use to infer fault states.

For each of the subcarriers, the forward model (2) applies. We formulate this as a linear system for which each measurement (row) corresponds to each subcarrier. Let $\mathbf{X}$ and $\mathbf{Z}$ be the matrices with ${\mathbf{x}^{(k)}}^{\mathsf{T}}$ and ${\mathbf{z}^{(k)}}^{\mathsf{T}}$ respectively as its rows for the subcarriers used for diagnosis ($k \in \{M_a, ..., M_b\}$). Therefore, we have (using (1))

$$\mathbf{Z}_{M\times N_t} = \mathbf{X}_{M\times N_{\mathrm{RF}}} \left(\mathbf{F}_{\mathrm{RF}}^{\mathsf{T}}\right)_{N_{\mathrm{RF}}\times N_t}. \tag{3}$$

Similarly, using the channel vector defined for each of the subcarriers $\mathbf{h}^{(k)}$ we can construct a 'channel-frequency' matrix $\mathbf{H}$ with ${\mathbf{h}^{(k)}}^{\dagger}$ as its rows:

$$\mathbf{H} = \begin{bmatrix} {\mathbf{h}^{(M_a)}}^{\dagger} \\ \vdots \\ {\mathbf{h}^{(M_b)}}^{\dagger} \end{bmatrix}_{M\times N_t} \tag{4}$$

Note that this is not to be confused with the 'channel matrix' used in MIMO systems.

The regions marked in red in Fig. 3 schematically show the elements of $\mathbf{X}$ and $\mathbf{Z}$, which behave like 'injected' symbols. If we vertically stack the measurements $\hat{y}^{(k)}, k \in \{M_a, ...M_b\}$ from (2), we obtain our measurement vector, $\hat{\mathbf{y}}$ consisting of the M FFT components of the received signal. Using (3), we form the measurement vector as:

$$\begin{aligned} \hat{\mathbf{y}} &= (\mathbf{H} \odot \mathbf{Z})\boldsymbol{\rho} + \hat{\mathbf{n}} \\ &= (\mathbf{H} \odot (\mathbf{X}\mathbf{F}_{\mathrm{RF}}^{\mathsf{T}}))\boldsymbol{\rho} + \hat{\mathbf{n}} \\ &= \mathbf{A}\boldsymbol{\rho} + \hat{\mathbf{n}}, \end{aligned} \tag{5}$$

where $\mathbf{A}$ is referred to as the 'forward model' as also indicated in Fig. 1. Subtracting this from the reference measurement of the healthy array, $\tilde{\mathbf{y}}$, (where $\boldsymbol{\rho} = \mathbf{1}$), we have the measurement vector, $\mathbf{y} = \tilde{\mathbf{y}} - \hat{\mathbf{y}}$, for the system of equations to be solved, which has a sparse solution since $\mathbf{1} - \boldsymbol{\rho}$ has only few non-zero elements, corresponding to faults:

$$\mathbf{y} = \mathbf{A}(\mathbf{1} - \boldsymbol{\rho}) + \mathbf{n} \tag{6}$$

where $\mathbf{n}$ represents the total noise. The estimate of the fault vector can be expressed as the solution to a sparse recovery problem, formulated as an unconstrained non-convex minimization problem using the p quasi-norm ($0 < p \leq 1$) [28], [29], following the approach in [7], [21], [30], given by

$$\boldsymbol{\rho}_{c,est} = \arg\min_{\boldsymbol{\rho}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\rho}_c\|_2^2 + \lambda\|\boldsymbol{\rho}_c\|_p^p. \tag{7}$$

This is a relaxation of the theoretical ℓ_0 minimization, which is NP-hard, and often replaced by the convex ℓ_1 norm for compressed sensing [31]. [28] provides an empirical review of the superiority of the non-convex ℓ_p over ℓ_1 minimization. For clarity, $\mathbf{1}-\boldsymbol{\rho}$ has been replaced with $\boldsymbol{\rho_c}$, the complementary fault vector. We present our solution approach in section III. It is important to emphasize here that in general, we will only require $M \ll N_t$ subcarriers to perform sparse recovery. Since typically $K > N_t$, we will require only a small part of the available spectrum for diagnosis.

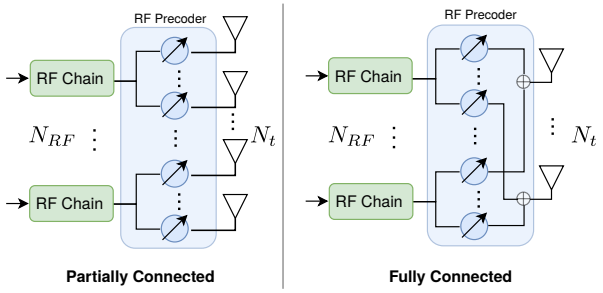


Fig. 4. Partially and Fully Connected architectures

B. Partial and Fully connected RF precoders

We address broadly two types of RF precoders (shown in Fig. 4) that determine the nature of $\mathbf{F}_{\mathrm{RF}}$, influencing our solution approach. In a partially connected structure (PCS), each antenna is connected to only one RF chain, whereas, in

Algorithm 1: Iterative Reweighted ℓ_1 (IRL1), [28]

Input: $\mathbf{A}, \mathbf{y}$
Output: $\boldsymbol{\rho}_{c,est}$
Let $l = 0$, $\boldsymbol{\rho_c}^{(l)} \leftarrow \mathbf{0}$
while $l < l_{max}$ **do**
  $\boldsymbol{\rho}_c^{(l+1)} = \arg\min_{\boldsymbol{\rho}_c} \frac{1}{2}||\mathbf{A}\boldsymbol{\rho}_c - \mathbf{y}||_2^2 + \lambda \sum_i v_i |\boldsymbol{\rho}_{c,i}|$,
  where $v_i = \frac{p}{(|\boldsymbol{\rho}_{c,i}^{(l)}| + \epsilon)^{(1-p)}}$
  ADMM implementation of weighted ℓ_1
  Let $k = 0$, initialize $\mathbf{w}^k \leftarrow \boldsymbol{\rho}_c^{(l)}$, $\mathbf{z}^k \leftarrow \mathbf{0}$, $\mathbf{u}^k \leftarrow \mathbf{0}$
  while $k < k_{max}$ **do**
    $\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{A}\mathbf{w} - \mathbf{y}||_2^2 + \frac{\rho}{2}||\mathbf{w} - \mathbf{z}^k + \mathbf{u}^k||_2^2$
    $\mathbf{z}^{k+1} = \arg\min_{\mathbf{z}} \lambda \sum_i v_i |\mathbf{z}_i| + \frac{\rho}{2}||\mathbf{w}^{k+1} - \mathbf{z} + \mathbf{u}^k||_2^2$
    $\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{w}^{k+1} - \mathbf{z}^{k+1}$
  end while
  $\boldsymbol{\rho}_c^{(l+1)} \leftarrow \mathbf{w}^k$; $l \leftarrow l + 1$
end while

Fig. 5. Iterative Reweighted ℓ_1 (IRL1) (based on [28, Algorithm 2])

a fully connected structure (FCS), every antenna is connected to all the RF chains through phase-shifters. $\mathbf{F}_{\text{RF}}$ for a PCS will have the specific structure as shown below [26]

$$\mathbf{F}_{\text{RF}} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{v}_2 & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{v}_{N_{\text{RF}}} \end{bmatrix}_{N_t \times N_{\text{RF}}}, \quad (8)$$

where $\mathbf{v}_n \in \mathbb{C}^{P\times 1}, n \in \{1, 2, \dots, N_{\text{RF}}\}$, $(P = N_t/N_{\text{RF}})$ and the mth element $\mathbf{v}_{n,m} = e^{j\phi_{(n,m)}}$, $m \in \{1, 2, \dots, P\}$. $\phi_{(n,m)}$ represent the corresponding phase shifts. In contrast, a FCS (Fig. 4), as in [13] corresponds to a dense matrix of the same size, with no zeroes (all unit-magnitude entries). Also note that fully digital architecture can be considered a special case of hybrid architectures where $\mathbf{F}_{\text{RF}}$ is an identity matrix, $\mathbf{I}_{N_t \times N_t}$.

III. Fault Diagnosis using Sparse Recovery

As evident from (5), the matrix $\mathbf{A}$ depends on the structure of $\mathbf{F}_{\text{RF}}$, which in turn is constrained by the RF precoding architecture used (see Section II-B). Depending on the architecture, we use two different solution approaches for these cases for the reasons described below.

A. Fault Diagnosis for Fully Connected Structures

For an FCS (and fully digital architectures), we solve for the complementary fault vector $\boldsymbol{\rho_c}$ in (7) using an iterative reweighted ℓ_1 (IRL1) approach, which solves an ℓ_1 optimization in each iteration (see Algorithm 1). We solve the ℓ_1 minimization step using the alternating direction method of multipliers (ADMM) [32].

To obtain good recovery, we desire the mutual coherence, $\mu(\mathbf{A})$ [33], [34] of $\mathbf{A}$ (the largest absolute normalized inner product among the columns of $\mathbf{A}$) to be small as possible. For an FCS, it is possible to engineer $\mathbf{x}^{(k)}$ for good mutual coherence properties of $\mathbf{A}$, as shown in Section IV.

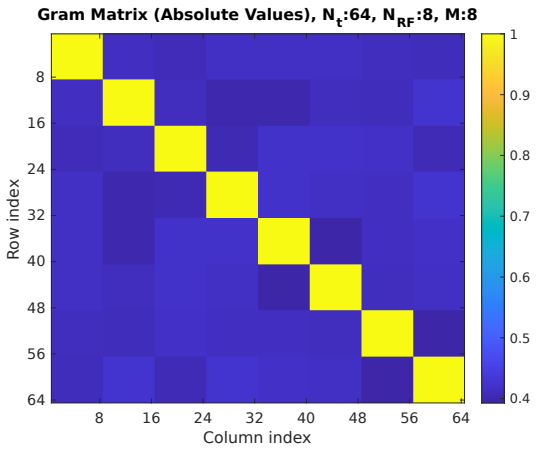


Fig. 6. Optimized Gram Matrix of $\mathbf{A}$, i.e., $\mathbf{A}^\dagger\mathbf{A}$ with a PCS

However, this runs into a fundamental problem in the case of PCS, which we deal with using an alternative strategy described next.

B. Fault Diagnosis for Partially Connected Structures

From (3) and (8) that define $\mathbf{Z}$ and $\mathbf{F}_{\text{RF}}$ respectively, it is easily inferred that for a PCS, multiplication of $\mathbf{X}$ by $\mathbf{F}_{\text{RF}}^T$ leads to $\mathbf{Z}$ having 'blockwise' coherent columns, i.e., the first P columns of $\mathbf{Z}$ are simply scaled versions of the first column of $\mathbf{X}$, similarly for the next P and so on. This results in $\mathbf{Z}$ having unit coherence, rendering $\mathbf{A}$ with high coherence, as illustrated in the example shown in Fig. 6. Whatever optimization we do on $\mathbf{X}$, the mutual coherence achievable for $\mathbf{A}$ is constrained by the nature of $\mathbf{F}_{\text{RF}}$. This is fundamentally due to the sub-connected nature of the array which is not a problem in the FCS and fully digital cases. It, therefore, does not appear to be useful for sparse recovery.

While it is true that sparse recovery of the exact position of faults using (7) is not a realistic proposition, identifying the subarray in which the fault(s) are present is possible. This bypasses the mutual coherence condition for an alternative criterion that is easily optimized for $\mathbf{A}$ with a PCS. In this way, we propose a 'two-stage' diagnosis method where the second stage, described in Section III-C identifies faulty elements within each subarray.

Rather than conceiving of $\boldsymbol{\rho_c}$ as a sparse vector, we can think of $\boldsymbol{\rho_c}$ to be 'block-sparse':

Definition III.1 (**Block-Sparsity** [35])**.** A vector $\mathbf{x}_B$ of length N is called *block S-sparse* over $\mathcal{I} = \{d_1, ..., d_B\}$ if $\mathbf{x}_B[i]$ is non-zero for at most S indices i where $N = \sum_{i=1}^{B} d_i$. The block-support of $\mathbf{x}_B$ is defined as $\text{supp}(\mathbf{x}_B) = \{i | \mathbf{x}_B \neq \mathbf{0}_{d_i \times 1}\}$. If $\Gamma = \text{supp}(\mathbf{x}_B)$, we have $|\Gamma| = S$.

In other words, we may assume the *number of subarrays that have faulty elements is much fewer than the total number of subarrays*. If only S_b of the subarrays (out of N_{RF}) contain faulty elements, then $\boldsymbol{\rho_c}$ will be S_b block sparse. This is a reasonable assumption since faulty antennas due to blockages, for instance, are more likely to be localized within a few subarrays.

Algorithm 2: Redundant-Block OMP (BOMP-R) [35]

Input: $\mathbf{A}, \mathbf{y}$
Output: Γ_t
Let $\mathbf{r}^0 \leftarrow \mathbf{y}, t = 1, \Gamma_0 = \emptyset,$ and $\mathbf{T}^0 = [\ ]$
while *"stopping criterion is not met"* **do**
  Choose the block index Λ_t that satisfies: $\Lambda_t = \arg\min_{i\in\Omega} \left\|\mathbf{r}^{t-1} - \text{Proj}\left(\mathbf{r}^{t-1}, \text{span}(\mathbf{A}[i])\right)\right\|_2$
  Let $\Gamma_t = \Gamma_{t-1} \cup \Lambda_t, \mathbf{T}^t = \left[\mathbf{T}^{t-1}, \mathbf{A}\left[\Lambda_t\right]\right]$, and calculate $\mathbf{y}^p = \text{Proj}\left(\mathbf{y}, \text{span}\left(\mathbf{T}^t\right)\right)$
  $\mathbf{r}^t = \mathbf{y} - \mathbf{y}^p$
  $t \leftarrow t + 1$
end while

Fig. 7. Redundant-Block OMP (BOMP-R)

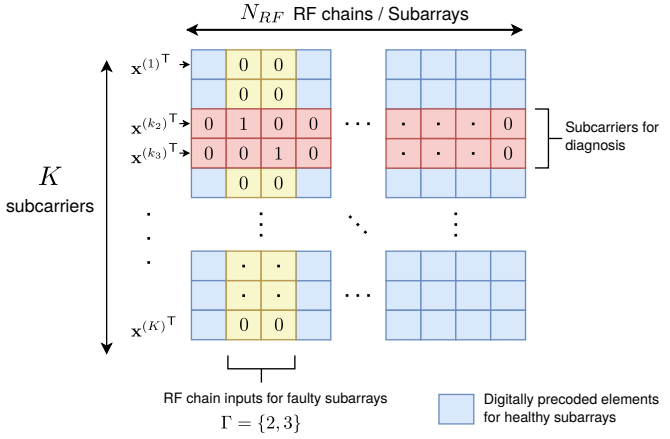


Fig. 8. Schematic showing spectrum of the inputs to the RF chains for the second stage of fault diagnosis for a PCS.

Block-sparse recovery algorithms such as BOMP [36] provide a solution when it is known that the solution has a block-sparse structure and provides the solution vector. In our case, however, obtaining a unique solution is impossible, as the columns within each block are redundant. To get the 'block-support' Γ of the $\boldsymbol{\rho}_c$ (i.e., which of the blocks in the solution vector have non-zero elements, or which subarrays have faults), we instead use Block OMP for redundant blocks (BOMP-R) [35], which accommodates the case where linearly dependent columns exist within each 'block' of the matrix and returns the block-support (in our case, the faulty subarrays) as the solution.

In our case, all of the blocks are of the same size, implying $d_i = P\ \forall\ i$, and there are N_{RF} such blocks. Each of the blocks of the fault vector $\boldsymbol{\rho}_C$ is represented as $\boldsymbol{\rho}_C[i] \in \mathbb{C}^P$, while the matrix $\mathbf{A}[i]$ refers to the ith block of the matrix (i.e. the corresponding ith block of columns). BOMP-R (see Algorithm 2) returns the estimate of the block support, Γ_{est}.

The condition that the matrix $\mathbf{A}$ must satisfy to guarantee recovery by BOMPR is the Block-RIP condition [35]. Similar to the RIP (Restricted Isometry Property), this condition is not easy to verify in practice. We instead use the *mutual subspace incoherence* defined in Section IV for a quantitative understanding of how 'good' the matrix $\mathbf{A}$ is for recovery, and optimize $\mathbf{X}$ for this metric.

C. Fault diagnosis within a subarray

The first stage for PCS described above plays two roles: One, of online detection of faults, as well as localizing faults to the subarray in which they are present. Next, to pinpoint faulty antennas within each of the subarrays found to be faulty, we require additional measurements. These measurements are made sequentially by changing the phase shifts within the faulty subarrays for each measurement, very similar to the method in [7]. It is important to note that controlling $\mathbf{x}^{(k)}$ is not useful here since it does not create any discriminative features among elements of a subarray. Measurements for this stage are made simultaneously for all of the subarrays $n \in \Gamma_{est}$ found to contain faults, each on its own subcarrier with index k_n. Therefore $S_b = |\Gamma_{est}|$ of the K subcarriers are disengaged from operation for diagnosis (inputs to those RF chains of healthy subarrays are set to zero at those subcarriers). We fix the inputs to the faulty subarray with index n as $x_n^{(k)} = 1$ for $k = k_n$ and $x_n^{(k)} = 0\ \forall\ k \neq k_n$. The faulty subarrays are not used for normal transmission ($x_n^{(k)}$ is the entry of $\mathbf{x}^{(k)}$ that corresponds to RF chain n, and therefore subarray n). Fig. 8 provides a visual description of the subcarriers used for diagnosis, and the subcarriers used for normal transmission.

For a PCS, $\mathbf{F}_{\text{RF}}$ will have the form as shown in (8). Each of the vectors $\mathbf{v}_n$ correspond to the phase-shifts for the nth subarray. For the lth of L measurements ($l \in \{1, ..., L\}$), let us say we apply the phase-shifts $\mathbf{v}_n^l$ to subarray n. We can then express the lth measurement $\hat{y}_l^{(k_n)}$ similar to (2), as

$$\begin{aligned}\hat{y}_l^{(k_n)} &= \mathbf{h}^{(k_n)^\dagger}\left(\mathbf{z}^{(k_n)}[n] \odot \boldsymbol{\rho}[n]\right) + \hat{n}_l^{(k_n)} \\ &= \mathbf{h}^{(k_n)^\dagger}\left(\mathbf{v}_n^l \odot \boldsymbol{\rho}[n]\right) + \hat{n}_l^{(k_n)}, \quad n \in \Gamma_{est}. \end{aligned} \tag{9}$$

Here, $\mathbf{z}^{(k)}[n]$ refers to the part (block) of $\mathbf{z}$ corresponding to the elements of subarray n. Stacking these L sequential measurements together as the vector $\hat{\mathbf{b}}_n = [\hat{y}_1^{(k_n)} \ldots \hat{y}_L^{(k_n)}]^\mathsf{T}$, we can write the measurement model as

$$\begin{aligned}\hat{\mathbf{b}}_n &= \begin{bmatrix}(\mathbf{v}_n^1)^\mathsf{T} \\ \vdots \\ (\mathbf{v}_n^L)^\mathsf{T}\end{bmatrix}\begin{bmatrix}\ddots & & 0 \\ & \mathbf{h}^{(k_n)^*} & \\ 0 & & \ddots\end{bmatrix}\boldsymbol{\rho}[n] + \hat{\mathbf{n}}^{(k_n)} \\ &= \mathbf{V}_n\,\text{diag}(\mathbf{h}^{(\mathbf{k_n})^*})\boldsymbol{\rho}[n] + \hat{\mathbf{n}}^{(k_n)} \\ &= \mathbf{B}_n\boldsymbol{\rho}[n] + \hat{\mathbf{n}}^{(k_n)}, \end{aligned} \tag{10}$$

We make measurements on their corresponding subcarriers k_n for all $n \in \Gamma_{est}$ (i.e., each of the faulty subarrays). The measurements for each subarray are then subtracted from reference measurements, resulting in $\mathbf{b}_n$. We then solve the following unconstrained optimization problem:

$$\min_{\boldsymbol{\rho}_c[\Gamma]} \frac{1}{2}\|\mathbf{b} - \mathbf{M}\boldsymbol{\rho}_c[\Gamma]\|_2^2 + \lambda\|\boldsymbol{\rho}_c[\Gamma]\|_p^p, \tag{11}$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3247727

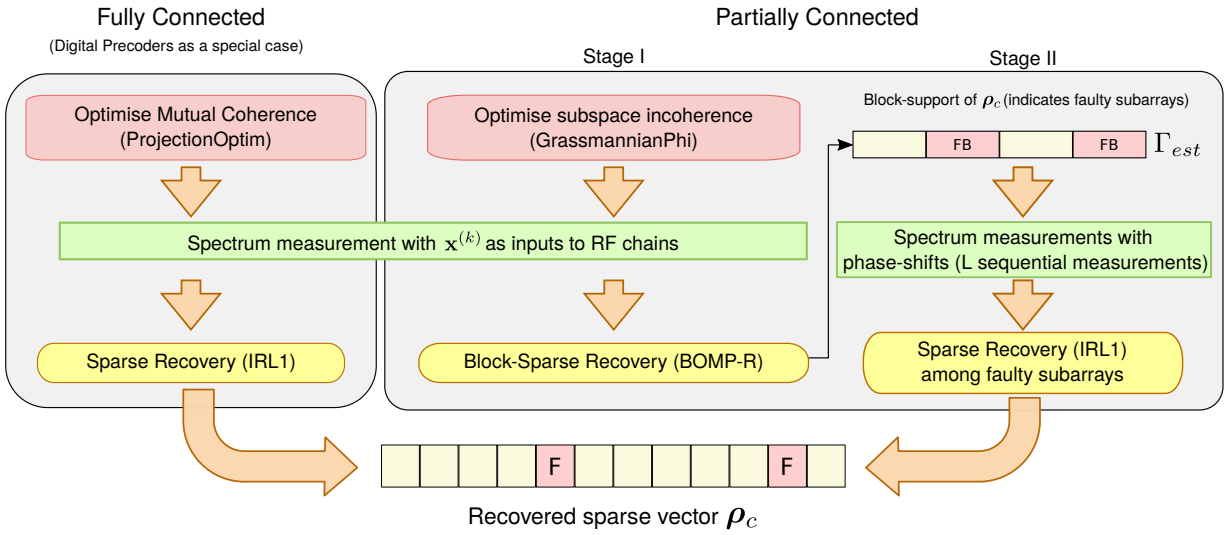


Fig. 9. **Algorithm flow for fault diagnosis**: For FCS, (and fully digital precoding systems, where $N_t = N_{\text{RF}}$ and $\mathbf{F}_{\text{RF}} = \mathbf{I}$) diagnosis is performed in a single step. The inputs $\mathbf{x}^{(k)}$ ($k \in \Lambda$) are optimized for the mutual coherence of the sensing matrix $\mathbf{A}$. Using these as inputs, and also the measured spectrum at the receiver, we obtain the sparse complementary fault vector. For PCS, diagnosis is performed in two steps - first, detecting and localizing faults to the level of subarrays ('fault blocks' (FB) of the block-sparse fault vector $\boldsymbol{\rho}_c$), followed by narrowing down the positions of faults with additional measurements. The first stage of diagnosis shares the same measurement modality as the FCS case, except that the inputs $\mathbf{x}^{(k)}$ are optimized for subspace incoherence rather than mutual coherence.

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{B}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{n_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{B}_{n_{S_b}} \end{bmatrix}_{LS_b \times PS_b} \quad \text{and}$$

$$\mathbf{b}^{\mathsf{T}} = \begin{bmatrix} \mathbf{b}_{n_1}^{\mathsf{T}} & \mathbf{b}_{n_2}^{\mathsf{T}} & \dots & \mathbf{b}_{n_{S_b}}^{\mathsf{T}} \end{bmatrix}_{1 \times LS_b}, \tag{12}$$

where S_b ($= |\Gamma|$) refers to the number of faulty subarrays, $\boldsymbol{\rho}_c[\Gamma]$ is the part of $\boldsymbol{\rho}_c$ indexed by the positions of faulty subarrays, and $\mathbf{0}$ is a $L \times P$ matrix of zeroes. Here, each of the $\mathbf{B}_n$ will have the same mutual coherence as $\mathbf{V}_n$, and the columns of different blocks (of the block-diagonal $\mathbf{B}$) are orthogonal. Therefore, the mutual coherence of $\mathbf{M}$ will simply be the highest mutual coherence among $\mathbf{V}_n$

We reiterate that (11) uses sparsity of faults among all the subarrays identified to be possibly faulty, rather than sparsity within each subarray, which is a flawed assumption, as faults can often be grouped within a subarray. The solution to (11) is obtained using IRL1, as it is in the FCS case.

It is worth pointing out here that this (second) stage is implemented online only if the system can operate using the rest of the subarrays (since the subarrays that are detected to contain faults are excluded from normal operation). We do not see this as a limitation, as the system cannot operate normally anyway due to the presence of possible faults that have been detected online by the first stage. In this way, the first stage of diagnosis acts as a *fault detection* scheme that can be periodically executed while minimally affecting transmission. A high-level summary of our complete method for systems with the FCS and PCS cases is shown in Fig. 9, with the optimization procedures described in the following section.

IV. Matrix Design

The use of excitations instead of different measurement positions as in [5], [7] for the purpose of fault diagnosis makes it possible to optimize the forward model sensing matrix, $\mathbf{A}$, to improve fault diagnosis performance. Here similarly, we perform optimization of the matrix $\mathbf{A}$ in terms of its mutual coherence properties, employing different strategies for the two cases of RF precoder architectures described earlier.

The matrix $\mathbf{A} = \mathbf{H} \odot (\mathbf{X}\mathbf{F}_{\text{RF}}^{\mathsf{T}})$ is not entirely under our control; the channel-frequency matrix, $\mathbf{H}$, is determined by the channel, while the RF precoder matrix $\mathbf{F}_{\text{RF}}$ is determined by the precoding algorithm being used for transmission. This is because, if we want to perform an online diagnosis, we are not allowed to affect the other subcarriers, and the RF precoder is frequency flat. Therefore, what remains in our control is only the matrix $\mathbf{X}$, which consists of the inputs to the RF chains on the particular subcarriers Λ being used for fault diagnosis (i.e., the partial frequency content).

According to compressed sensing theory, for equivalence of the ℓ_1 and ℓ_0 problems (i.e. (7) with the theoretical zero-norm and convex 1-norm, respectively, instead of the p-quasi norm), a Restricted Isometry Property (RIP) must be satisfied by $\mathbf{A}$ with constant $\delta_{2S} < \sqrt{2} - 1$ [31], [37]. Randomized sensing matrices (e.g. with normally distributed entries) which satisfy RIP with high likelihood are not implementable in our case because of the constraint described above (of having control only over the $\mathbf{X}$-component of the sensing matrix $\mathbf{A}$). Deterministic (as opposed to randomised) RIP matrix constructions are not straightforward and have been attempted previously [37], [38], but are not practical since they are often for fixed matrix sizes. The mutual coherence $\mu(\mathbf{A})$ is straightforward to compute, and provides a weak upper bound on δ_{2S}, with $\delta_{2S} < (2S-1)\mu$ [37]. Furthermore, the sparsity of the signal, S,

that can be recovered goes as per, $S < \frac{1}{2}(1+\frac{1}{\mu})$ [33], [39], and therefore reconstructing higher sparsity motivates optimizing for a smaller mutual coherence from a practical point of view.

A. Fully Connected

We aim to optimize the matrix $\mathbf{A}$ to have as minimum mutual coherence as possible while manipulating only the values in $\mathbf{X}$. This matrix $\mathbf{A}$, under some conditions (described below), can be factorized conveniently into a product of a projection matrix $\mathbf{\Phi}$ and a dictionary matrix $\mathbf{\Psi}$, i.e. $\mathbf{A} = \mathbf{\Phi\Psi}$, such that $\mathbf{\Phi}$ is dependent entirely on $\mathbf{X}$, allowing us to manipulate its values freely. A projection matrix optimization algorithm [40] (we refer to it as 'ProjectionOptim' in Fig. 9) generates the projection matrix given a dictionary matrix such that the overall sensing matrix $\mathbf{\Phi\Psi}$ is optimal for mutual coherence.

Considering the case where the channel is multipath-free, we can easily recast the expression for $\mathbf{A}$ from (5) in the form of $\mathbf{A} = \mathbf{\Phi\Psi}$, which we now show:

Proof: Consider a $d-delay$ multiple-input single-output (MISO) channel model, $\mathbf{h}_s[d]$, for a linear antenna array, represented as [13], [41]:

$$\mathbf{h}_s[d] = \sqrt{\frac{N_\mathrm{t}}{L_p}}\alpha_r p_\mathrm{rc}\left(dT_s - \tau\right)\mathbf{a}\left(\theta_r\right) \tag{13}$$

where α_r is the complex path gain, which depends on the angle of departure of the path θ_r to the receiver probe, T_s is the sampling interval, $\mathbf{a}(\theta_r)$ is the array vector, p_rc is the shaping pulse (e.g., raised cosine) used for transmission, and L_p is the free space path loss. Such a model applies in a single path, line-of-sight (LOS) scenario.

We then have the channel vector for subcarrier k, [13]

$$\mathbf{h}^{(k)} = \sum_{d=0}^{D-1}\mathbf{h}_s[d]e^{-j\frac{2\pi k}{K}d} = \mathbf{a}\left(\theta_r\right)g^{(k)}, \tag{14}$$

where

$$g^{(k)} = \alpha_r\sqrt{\frac{N_t}{L_p}}\sum_{d=0}^{D-1}p_{rc}\left(dT_s-\tau\right)e^{-j\frac{2\pi k}{K}d}$$

$$\mathbf{a}\left(\theta_r\right) = \frac{1}{\sqrt{N_t}}\left[\,1\;\; e^{j\frac{2\pi}{\lambda}d_s\sin\theta_r}\;\ldots\; e^{j(N_t-1)\frac{2\pi}{\lambda}d_s\sin\theta_r}\,\right]^\mathsf{T}. \tag{15}$$

The channel-frequency matrix $\mathbf{H}$ in (4), having $h^{(k)\dagger}$ as its rows becomes a single-rank matrix that can be expressed as the outer product:

$$\mathbf{H} = \underbrace{\left[g^{(M_a)}\quad\ldots\quad g^{(M_b)}\right]^\dagger}_{\mathbf{g}}\mathbf{a}^\dagger\left(\theta_r\right) \tag{16}$$

Now, it becomes straightforward to recast our expression for $\mathbf{A}$ as follows ($\mathbf{f}_n$ refers to the transpose of the nth row of F_RF, and $\mathbf{x}^{(k)^\mathsf{T}}$ the row of $\mathbf{X}$ corresponding to subcarrier k):

$$\begin{aligned}
\mathbf{A} &= \mathbf{H}\odot\left(\mathbf{X}\mathbf{F}_\mathrm{RF}^\mathsf{T}\right)\\
&= \left(\mathbf{g}\mathbf{a}^\dagger\left(\theta_r\right)\right)\odot\left(\mathbf{X}\mathbf{F}_\mathrm{RF}^\mathsf{T}\right)\\
&= \begin{bmatrix} g^{(M_a)^*}\mathbf{x}^{(\mathbf{M_a})^\mathsf{T}}\mathbf{f}_1 a_1^* & \ldots & g^{(M_a)^*}\mathbf{x}^{(\mathbf{M_a})^\mathsf{T}}\mathbf{f}_{N_t}a_{N_t}^* \\ & \vdots & \\ g^{(M_b)^*}\mathbf{x}^{(\mathbf{M_b})^\mathsf{T}}\mathbf{f}_1 a_1^* & \ldots & g^{(M_b)^*}\mathbf{x}^{(\mathbf{M_b})^\mathsf{T}}\mathbf{f}_{N_t}a_{N_t}^* \end{bmatrix}\\
&= \underbrace{\begin{bmatrix} g^{(M_a)^*}\mathbf{x}^{(\mathbf{M_a})^\mathsf{T}} \\ \vdots \\ g^{(M_b)^*}\mathbf{x}^{(\mathbf{M_b})^\mathsf{T}} \end{bmatrix}}_{\mathbf{\Phi}}\underbrace{\begin{bmatrix}\mathbf{f}_1 a_1^* & \ldots & \mathbf{f}_{N_t}a_{N_t}^*\end{bmatrix}}_{\mathbf{\Psi}},
\end{aligned} \tag{17}$$

thus enabling optimization as in [40]. □

Notice that $\mathbf{\Psi}$ is dependent on $\mathbf{F}_\mathrm{RF}$, which is determined by the precoding algorithm (and architecture) being used, and not $\mathbf{X}$, the effect of which has been isolated into $\mathbf{\Phi}$. Once we have optimized for $\mathbf{\Phi}$, by using $\mathbf{\Psi}$ as our 'dictionary matrix', we obtain the vectors $\mathbf{x}^{(k)}, k \in \{M_a, ..., M_b\}$ by dividing the rows of the optimized $\mathbf{\Phi}$ by the elements of $\mathbf{g}$. The values of $\mathbf{X}$ can be scaled to fit within the allowed dynamic range without affecting $\mu\left(\mathbf{A}\right)$. On the other hand, the multipath case results in a channel-frequency matrix that is no longer of unit rank and cannot be decomposed into the form shown above. This channel model is beyond the scope of this paper and will be investigated in future work. The mutual coherence values obtained using this method are shown in Fig. 10a, which is improved compared to randomized $\mathbf{X}$ (with entries from a complex normal distribution).

For the fully digital case, with $N_\mathrm{RF} = N_t$ and $\mathbf{F}_\mathrm{RF} = \mathbf{I}_{N_t\times N_t}$, optimization of $\mathbf{A}$ for mutual coherence becomes trivial. Since $\mathbf{\Psi}$ becomes a diagonal matrix it does not affect the mutual coherence of $\mathbf{A}$. We can directly optimize for the mutual coherence of $\mu(\mathbf{\Phi})$ by generating a Grassmannian matrix [33], whose mutual coherence approaches the welch bound, $\sqrt{\frac{N_t-M}{M(N_t-1)}}$, the theoretical lower limit for mutual coherence of a matrix of size $M \times N_t$.

B. Partially Connected

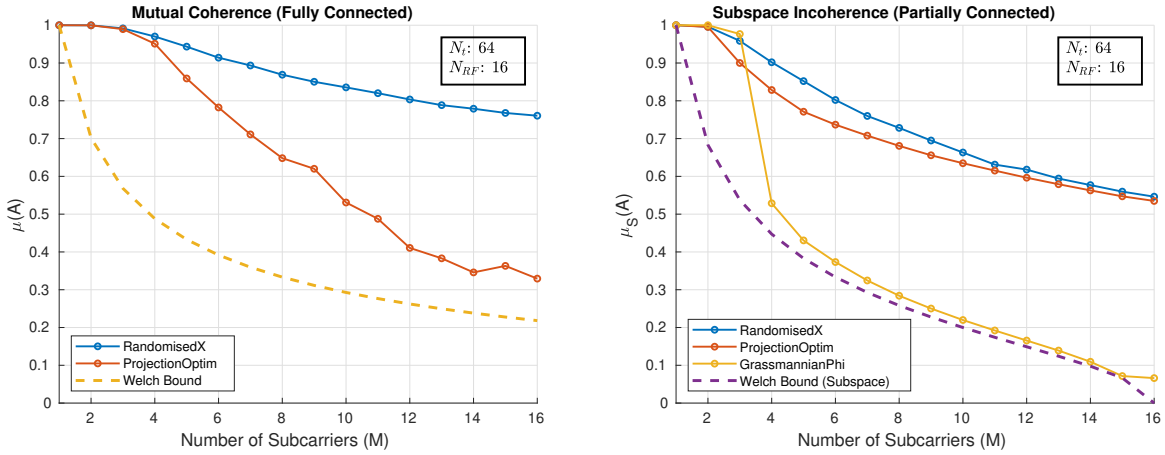
We now describe our methods to optimize the sensing matrix for each of the two stages of diagnosis in the PCS case - the first stage is block-sparse recovery.

a) Stage I: While the matrix $\mathbf{A}$ can still be decomposed as in (17), our objective for efficient *block-sparse* recovery is to optimize for a metric known as mutual *subspace* incoherence [34], [42], rather than mutual coherence.

In the matrix $\mathbf{A}$, each of the blocks $\mathbf{A}[i]$ span a subspace $\mathcal{S}_i$. The mutual subspace incoherence [34], [42], μ_S among the blocks of $\mathbf{A}$ is defined a measure of the smallest angle between any two subspaces from the given set,

$$\mu_S = \max_{i\neq j}\max_{\mathbf{p}\in\mathcal{S}_i,\mathbf{q}\in\mathcal{S}_j}\frac{\left|\mathbf{p}^\dagger\mathbf{q}\right|}{\|\mathbf{p}\|_2\|\mathbf{q}\|_2}. \tag{18}$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3247727



(a) (b)

Fig. 10. **Matrix design using coherence optimization**: (a) FCS case showing the effect of optimization for mutual coherence using [40]. (b) Subspace Incoherence (for partially connected) plotted using different matrix design techniques. The welch bound shown corresponding to a matrix of size $M \times N_{\text{RF}}$. In both plots, each data point is averaged over 400 realisations of $\mathbf{A}$.

Optimizing the mutual subspace incoherence of $\mathbf{A}$ will help us obtain a 'good' matrix to perform BOMP-R on (see Algorithm 2). Minimizing μ_S makes sense because the block sparsity, S_b is related to μ_S as $S_b < \frac{1}{2}\left(1 + \frac{1}{\mu_S}\right)$ [34], [42] just as in the case with conventional sparsity, thus enabling larger block-sparsities (more faulty subarrays) to be recovered.

Owing to the partially connected nature, $\mathbf{F}_{\text{RF}}^{\text{T}}$ (refer Section II-B), and hence $\mathbf{\Phi}$ have a special form which causes $\mathbf{\Phi\Psi}$ to have repeated/scaled columns of $\mathbf{\Phi}$. Therefore, if $\mathbf{\Phi}_i$ refers to the ith column of $\mathbf{\Phi}$, it is easy to see that $\text{span}\{\mathbf{\Phi}_i\} = \mathcal{S}_i$.

The problem of optimizing $\mathbf{A}$ for subspace incoherence, therefore, reduces to the problem of optimizing $\mathbf{\Phi}$ for mutual coherence. As the elements of $\mathbf{\Phi}$ are unconstrained, we simply generate a Grassmannian matrix [33], which closely approaches the welch bound, the theoretical lower limit for mutual coherence. The resulting mutual subspace incoherence is plotted in Fig. 10b (labeled as GrassmannianPhi), showing that it closely approaches the welch bound (minimum possible subspace incoherence) for subspace incoherence. It clearly achieves a much better sensing matrix for block-sparse recovery than projection matrix optimization [40], which tries to minimize the mutual coherence of $\mathbf{A}$, rather than mutual subspace incoherence.

b) Stage II: For this stage, we optimize the phase-shifts $\mathbf{V}_n$ applied to each subarray during the measurements sequence. As stated in Section III-C, the overall block-diagonal sensing matrix $\mathbf{M}$ in (11) will have the same mutual coherence as the maximum among those of $\mathbf{B}_n$. An essential constraint to optimization is that the RF precoder only has phase shifts; therefore, the phase shifts for subarray n, $\mathbf{V}_n$ will consist of elements with unit magnitude. This means that we cannot arbitrarily optimize $\mathbf{B}_n$.

As $\text{diag}(\mathbf{h}^{(k_n)^*})$ simply multiplies each of the columns of $\mathbf{V}_n$ by a factor, it will not affect the mutual coherence of $\mathbf{B}_n$. Hence, optimizing the matrix $\mathbf{V}_n$ for mutual coherence, which contains all unit magnitude elements is sufficient. From CS theory, we know that a random partial Fourier matrix (RPFM) satisfies a RIP with high probability [43]. We, therefore, choose $\mathbf{V}_n$ to be an RPF matrix.

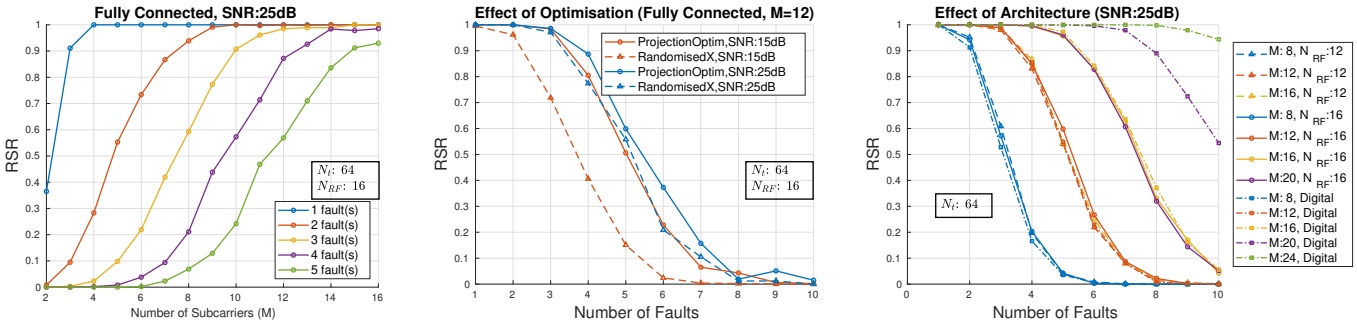
V. Simulations and Analyses

In this section, we evaluate our proposed approach for the various precoding configurations discussed, validate their effectiveness, and highlight their limitations.

For all of the results shown (unless specified otherwise), we consider a half-wavelength spaced linear array with $N_t = 64$ antennas and the number of RF chains $N_{\text{RF}} = 16$ or 8, as specified using a boxed annotation in each plot. In general, for each data point, we average performance over 20 random realizations of F_{RF}, and for each of these, the faults' positions are selected at random, averaging over 50 Monte Carlo (MC) simulations, thus resulting in an average over 1000 realizations per data point. The y-axis is the rate of successful recovery (RSR), defined as the fraction of MC trials where the diagnosis recovers all elements' state (faulty or healthy block/antenna) correctly. For evaluation purposes, we consider binary faults, i.e., an element is assumed to have a value of 1 when working correctly and a value of 0 when faulty. To obtain the simulated measurement vector $\mathbf{y}$, we used the forward model described in Section II, with a LOS channel model as in (16), and do not consider mutual coupling effects. The SNR for each simulation is given by $10\log\left(\frac{\|\mathbf{y}\|_2^2}{M\sigma^2}\right)$ dB (for FCS), with the noise vector $\mathbf{n}$ assumed to follow $\mathcal{CN}(0, \sigma^2 I)$. A similar procedure is adopted for both stages of the partially connected case.

The RF precoding matrix $\mathbf{F}_{\text{RF}}$ is taken to be composed of N_{RF} random columns from a DFT matrix of size N_t, mirroring the precoding algorithms described in [44], [45] for the FCS case. For the partially connected case, we fill the non-zero elements of $\mathbf{F}_{\text{RF}}$ (see (8)) with phase shifts uniformly at random in $[0, \pi]$ rad. A system with $K = 2048$ subcarriers and a central frequency of 3.5 GHz (to mimic a C-band 5G carrier frequency) with a subcarrier spacing of 15 kHz is used. For the sparse recovery optimization algorithm of (7), the

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3247727



(a) (b) (c)

Fig. 11. **Fault diagnosis performance for FCS**: (a) Progressive improvement in fault diagnosis with increasing number of subcarriers M used for diagnosis. (b) Plot showing the improved reliability of fault diagnosis with optimized inputs. (c) Influence of the number of RF chains on the maximum achievable diagnosis for $N_{\text{RF}} = 12, 16$. This also shows the improvement in recovery for larger number of RF chains for a given number of subcarriers M used.

hyperparameters $p \in \{0.01, 0.05, 0.1, 0.5\}$ and $\lambda \in [10^{-3}, 10]$ were chosen using a grid-search for different values of SNR, using the average RSR over varying number of faults from 1 to 5 as the metric.

A. Fully Connected RF precoder

For the FCS case, we calculate the inputs $\mathbf{X}$ using the procedure described in Section IV-A, and solve the minimization problem (7) for the complementary fault vector $\boldsymbol{\rho}_c$ using IRL1. In Fig. 11, we show how fault diagnosis improves with the number of subcarriers M used, the advantage of optimizing the inputs, and how the precoder configuration may affect the performance of diagnosis. The hyperparameter pair (p, λ) chosen using a grid-search to solve (7) in all of the plots shown in Fig. 11 are $(0.05, 0.063)$ and $(0.01, 0.32)$ for 25 dB and 15 dB respectively. For the fully digital case in Fig. 11c, the hyperparameter pair used is $(0.01, 3.2 \times 10^{-4})$.

Again, we emphasize that all these M subcarrier measurements can be obtained at one go *while* the system is still in operation on the rest of the spectrum (other subcarriers). On an OFDM-based system, one may allocate a single resource block (RB), made of 12 subcarriers [46]. The essential features of the plots in Fig. 11 are highlighted below:

1) In Fig. 11a, we show how fault diagnosis improves with the number of subcarriers M used for diagnosis. Our method comfortably detects and diagnoses up to 3 faults in a 64-antenna array by simply allocating one RB ($M = 12$) from the time-frequency resource grid for fault diagnosis. The minimum number of subcarriers required for diagnosis increases gradually with the number of faults.
2) Fig. 11b shows that engineering the injected $\mathbf{x}^{(k)}$ using the projection matrix optimization algorithm (ProjectionOptim) proves advantageous for diagnosis. The diagnosis is more resilient to noise when using optimized inputs, than when using randomized values in $\mathbf{X}$. At 15 dB, the optimized version recovers up to 3 faults with near certainty, which is achievable only at 25 dB when using randomized inputs.
3) Since our technique hinges on using engineered inputs to the RF chains, the number of RF chains in the system N_{RF} can have an influence on the diagnosis performance. From Fig. 11c, we compare performance of diagnosis with increasing number of RF chains, also including the fully digital case ($N_{\text{RF}} = N_t$, and $\mathbf{F}_{\text{RF}} = \mathbf{I}_{N_{\text{RF}} \times N_t}$):
   a) For a given number of RF chains, the improvement in diagnosis performance saturates once the number of subcarriers $M \geq N_{\text{RF}}$, meaning that there is not much useful information that can be gained from additional measurements. We can intuitively understand this in the light of the fact that when $M > N_{\text{RF}}$, the projection matrix $\boldsymbol{\Phi}$ becomes a tall matrix. Measuring more projections would not lead to more information since the measurements have been optimized.
   b) For a given number of subcarriers M, the average performance of diagnosis remains the same irrespective of the number of RF chains (N_{RF}), as long as $M \leq N_{\text{RF}}$. In fact, the performance of diagnosis with a hybrid architecture *matches with the performance of diagnosis for a fully digital architecture*. The optimization allows us to obtain the maximum information possible from the measurements for each value of M.

B. Partially Connected - Two-Stage Diagnosis

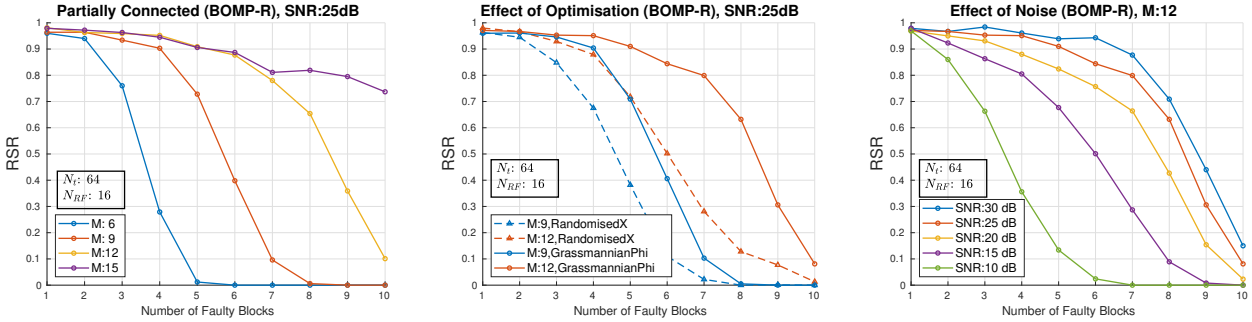
a) Stage I (BOMP-R): An implementation detail in Algorithm 2 is the stopping criterion used to stop BOMP-R iterations without the knowledge of the block-sparsity (number of faulty subarrays). The iterations are continued as long as the inequality

$$\frac{\|\mathbf{r}^t\|_2}{\sqrt{M}} \leq \kappa\sigma \tag{19}$$

holds, where the residue $\mathbf{r}^t = \mathbf{y} - \mathbf{y}^p$ (see Algorithm 2), σ is the standard deviation of noise, and $\kappa = \sqrt{3}$ specifies the confidence region of 95%, as the noise is assumed to follow $\mathcal{CN}(0, \sigma^2 I)$.

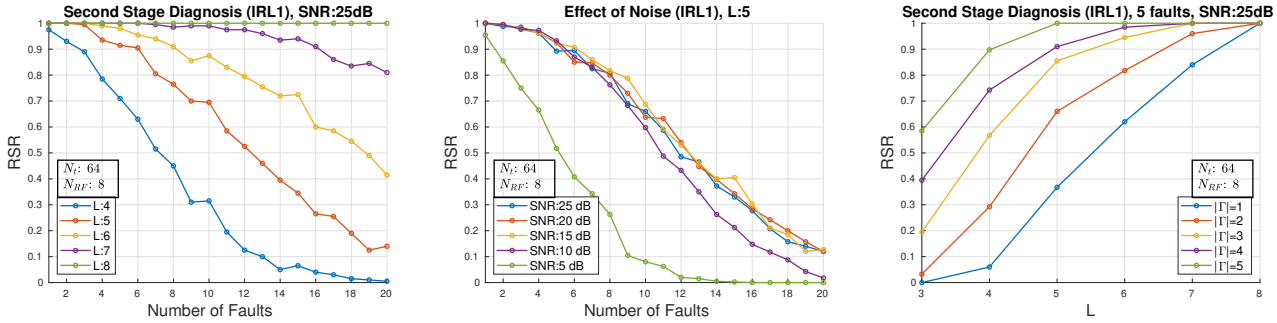
Similar to the previous performance metric, the RSR here is the fraction of trials where the existence of faults in all the subarrays that contain faults are identified exactly. Fig. 12

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3247727



(a) (b) (c)

Fig. 12. **First-stage diagnosis (BOMP-R) for PCS:** The total number of faults among the antennas was taken to be double the number of faulty subarrays (blocks), while the minimum number of faults in each faulty block was restricted to be at least one, randomly distributing the faults among the subarrays that contain faults. (a) Diagnosis performance for different value of M (subcarriers used for diagnosis). (b) Plot shows an improved diagnosis of subarrays by optimizing mutual subspace incoherence using $\mathbf{\Phi}$ as a Grassmannian matrix. (c) Degradation of the reliability of diagnosis with decreasing SNR of the online measurements.



(a) (b) (c)

Fig. 13. **Second-stage diagnosis with BlockGenie for PCS**: In these plots, the set of indices of faulty subarrays (Γ) is assumed to be known apriori. (a) RSR plotted as a function of total number of faults in the array (across all subarrays). (b) Sensitivity to noise in measurements, plotted for number of second-stage measurements $L = 5$. (c) Dependence of fault diagnosis performance on the distribution of faults among subarrays. $|\Gamma| = 1$ means that all the faults are concentrated in a single subarray.

shows the performance of first-stage diagnosis using BOMP-R:

1) In Fig. 12a, we show the number of faults diagnosed for a given number of subcarriers M. With a single RB ($M = 12$), BOMP-R can detect up to 4 faulty subarrays with near-certainty at an SNR of 25 dB.
2) Similar to the FCS case, using engineered inputs $\mathbf{x}^{(k)}$ from a Grassmannian $\mathbf{\Phi}$ is clearly beneficial, as illustrated in Fig. 12b, validating our understanding that minimising the mutual subspace incoherence $\mu_S(\mathbf{A})$ (from Section IV-B) helps BOMP-R perform better.
3) At higher noise levels (SNR below 20 dB), when using one RB ($M = 12$) performance of BOMP-R degrades (see Fig. 12c).

b) Stage II (IRL1): To isolate our evaluation of the second stage diagnosis from the first, we first assume that the indices of the faulty subarrays Γ are known exactly, provided by an oracle/genie (hereafter referred to as *BlockGenie*). In the plots of Fig. 13, N_{RF} has been chosen as 8, instead of 16 to show a broader range of values (since the number of measurements, $L \leq P$, the number of antennas in a subarray). We also randomly distribute the faulty elements at random across all the subarrays for the results in Figs. 13a and 13b, meaning that each subarray could have any number of faulty elements, while the total number of faults is as shown on the x-axis. The hyperparameter pair (p, λ) used for (11) in all of Fig. 13 is $(0.01, 0.01)$. The performance is not affected by the specific hyperparameters as long as λ is less than a certain threshold (different for each p). Some of the main features shown in Fig. 13 are described below:

1) Fig. 13a shows the RSR for increasing number of faults. With about 5-6 additional measurements (L), about 4 faulty antennas (distributed anywhere among the subarrays) can be detected with near certainty on average.
2) The diagnosis is very robust to noise, with performance remaining nearly the same for 10 dB and above (see Fig. 13b). This may be be attributed to the fact that (11) is a much easier problem to solve due to its block-diagonal structure (implying that the faulty elements in different subarrays cannot be confused among each other) composed of RPF matrices.
3) Since the sensing matrix $\mathbf{M}$ in (11) has a unique block-diagonal structure, faults spread across many subarrays (corresponding to different blocks of the matrix) will be easier to distinguish than faults concentrated within a

single subarray. This is demonstrated in Fig. 13c. For example, when $|\Gamma| = 1$, i.e., all the faults are concentrated within a single subarray, second-stage diagnosis is not as good. It requires 8 measurements to achieve $RSR > 90$, which is equal to the number of antennas within the subarray. Diagnosis improves as the faults are spread across subarrays. This can be attributed to the fact that since sparsity is not satisfied within the subarray, the 'block' columns in $\mathbf{M}$ are more easily confused among each other. In contrast, when all the faults are spread across the subarrays such that only one fault exists in each faulty subarray ($|\Gamma| = 5$), they are easily distinguishable from one another.

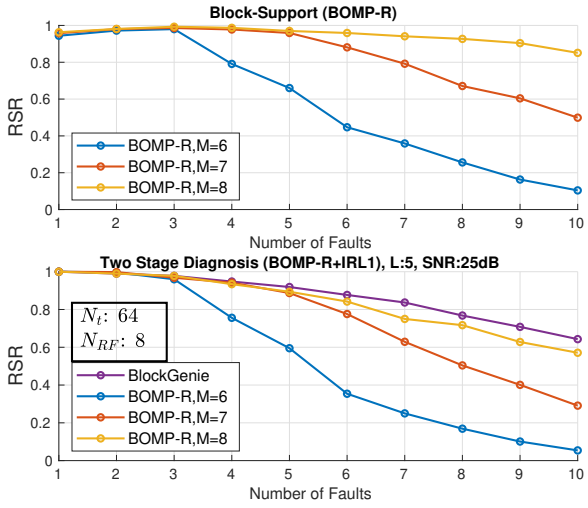


Fig. 14. **Overall two-stage diagnosis for PCS**: (Top) The success rate of recovering the correct subarray (block-support of the solution) using BOMP-R, the first stage of diagnosis. (Bottom) The success rate of second-stage diagnosis using $L = 5$ measurements given the result of BOMP-R from the first stage. Also shown for reference is the performance of BlockGenie+IRL1 (assuming a priori knowledge of the faulty subarrays). In both, the x-axis is the total number of faults in the array. This shows the susceptibility of overall diagnosis to errors in the first stage (BOMP-R).

Finally, we show the overall performance of the two-stage diagnosis and the effect of inaccuracies of BOMP-R on second-stage diagnosis in Fig. 14. When BOMP-R exactly recovers the indices of faulty subarrays, the overall performance of the two-stage diagnosis coincides with the one obtained using BlockGenie. As the number of measurements decreases and BOMP-R becomes more inaccurate for a higher number of faults, the performance of the second-stage diagnosis is naturally affected. Our results show that using just a tiny part of the available spectrum ($M \ll K$), our method can reliably perform detection and diagnosis of faults in a large antenna array while minimally affecting normal operation.

VI. Discussion

In this work, we have proposed a new method for fault diagnosis that has the unique advantage of being implementable while the wireless system is still in operation by making use of a small subset of subcarriers for fault diagnosis while usual data transmission simultaneously happens through the rest of the subcarriers. The measurement can be obtained in a single shot (for fully connected and the first stage for partially connected) and is directly implementable on practical systems by using only digital operations. By engineering the inputs for optimal measurement, we enable diagnosis using very few subcarriers, thus minimally affecting services. Since the method is online, it can be used to run diagnoses as an automated self-diagnosis system periodically.

Some significant challenges to solve include the requirement for reference measurements and the handling of multi-path channels. This could potentially pave the way for a robust method of fault diagnosis that caters to the needs of real-world communication systems by performing quick online diagnoses and avoiding interruption of services. Our novel methodology and results provide a basis for such exciting and interesting work in the future.

References

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless. Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] M. D. Migliore, "Array diagnosis from far-field data using the theory of random partial fourier matrices," *IEEE Antennas Wireless Propag. Lett.*, vol. 12, pp. 745–748, Jun. 2013.

[3] G. Oliveri, P. Rocca, and A. Massa, "Reliable diagnosis of large linear arrays—a bayesian compressive sensing approach," *IEEE Trans. Antennas Propag.*, vol. 60, no. 10, pp. 4627–4636, Oct. 2012.

[4] A. F. Morabito, R. Palmeri, and T. Isernia, "A compressive-sensing-inspired procedure for array antenna diagnostics by a small number of phaseless measurements," *IEEE Trans. Antennas Propag.*, vol. 64, no. 7, pp. 3260–3265, Jul. 2016.

[5] M. E. Eltayeb, T. Y. Al-Naffouri, and R. W. Heath, "Compressive sensing for millimeter wave antenna array diagnosis," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2708–2721, Jun. 2018.

[6] G. Medina, A. S. Jida, S. Pulipali, R. Talwar, N. A. J, T. Y. Al-Naffouri, A. Madanayake, and M. E. Eltayeb, "Millimeter-wave antenna array diagnosis with partial channel state information," in *ICC 2021 - IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–5.

[7] K. P. Prajosh, U. K. Khankhoje, and F. Ferranti, "Element excitation optimization for phased array fault diagnosis," *J. Electromagn. Waves and Appl.*, vol. 35, no. 1, pp. 39–50, Sep. 2021.

[8] R. Sun, W. Wang, L. Chen, G. Wei, and W. Zhang, "Hybrid beamforming system diagnosis: Failure modeling and identification," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3415–3429, Jun. 2021.

[9] ——, "Blind diagnosis for millimeter-wave large-scale antenna systems," *IEEE Comm. Lett.*, vol. 25, no. 7, pp. 2390–2394, Jul. 2021.

[10] S. Ma, W. Shen, J. An, and L. Hanzo, "Antenna array diagnosis for millimeter-wave MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4585–4589, Apr. 2020.

[11] C. Xiong and G. Xiao, "A compressed sensing based diagnosing method for array antenna element failures by sampling at different frequencies," in *2020 IEEE MTT-S International Wireless Symposium (IWS)*, Sep. 2020, pp. 1–3.

[12] K. P. Prajosh, S. S. Ranganathan, F. Ferranti, and U. K. Khankhoje, "Efficient mutual-coupling aware fault diagnosis of phased array antennas using optimized excitations," *IEEE Antennas Wireless Propag. Lett.*, 2022.

[13] A. Alkhateeb and R. W. Heath, "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1801–1818, May 2016.

[14] S. Park and R. W. Heath, "Frequency selective hybrid precoding in millimeter wave OFDMA systems," in *2015 IEEE Global Commun. Conf. (GLOBECOM)*, 06–10 Dec. 2015, pp. 1–6.

[15] L. Kong, S. Han, and C. Yang, "Wideband hybrid precoder for massive mimo systems," in *2015 IEEE Global Conf. Signal and Inf. Process. (GlobalSIP)*, 14–16 Dec. 2015, pp. 305–309.

[16] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for OFDM-based large-scale MIMO systems," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 03-06 Jul. 2016, pp. 1–6.

[17] H. Yuan, J. An, N. Yang, K. Yang, and T. Q. Duong, "Low Complexity Hybrid Precoding for Multiuser Millimeter Wave Systems Over Frequency Selective Channels," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 983–987, Jan. 2019.

[18] K. Yang, M. Shi, H. Yuan, and Z. Ni, "Hybrid precoding for wideband multiuser mmwave systems," in *Millimeter-Wave Communication Systems: Network Analysis and Hybrid Precoding Design*. Springer, 2022, pp. 133–154.

[19] M. D. Migliore, "A compressed sensing approach for array diagnosis from a small set of near-field measurements," *IEEE Trans. Antennas Propag.*, vol. 59, no. 6, pp. 2127–2133, Jun. 2011.

[20] B. Fuchs, L. L. Coq, and M. D. Migliore, "Fast antenna array diagnosis from a small number of far-field measurements," *IEEE Transactions on Antennas and Propagation*, vol. 64, no. 6, pp. 2227–2235, Jun. 2016.

[21] C. Xiong, G. Xiao, Y. Hou, and M. Hameed, "A Compressed Sensing-Based Element Failure Diagnosis Method for Phased Array Antenna During Beam Steering," *IEEE Antennas Wireless Propag. Lett.*, vol. 18, no. 9, pp. 1756–1760, Sep. 2019.

[22] R. W. Bäuml, R. F. H. Fischer, and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *Electronics Lett.*, vol. 32, no. 22, p. 2056, Jan. 1996.

[23] H. Breiling, S. H. Muller-Weinfurtner, and J. B. Huber, "Slm peak-power reduction without explicit side information," *IEEE Comm. Lett.*, vol. 5, no. 6, pp. 239–241, 2001.

[24] A. Bereyhi, M. A. Sedaghat, R. R. Müller, and G. Fischer, "Glse precoders for massive mimo systems: Analysis and applications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4450–4465, 2019.

[25] A. Bereyhi, M. A. Sedaghat, and R. R. Müller, "Asymptotics of nonlinear lse precoders with applications to transmit antenna selection," in *2017 IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2017, pp. 81–85.

[26] N. Li, Z. Wei, H. Yang, X. Zhang, and D. Yang, "Hybrid precoding for mmwave massive mimo systems with partially connected structure," *IEEE Access*, vol. 5, pp. 15 142–15 151, 2017.

[27] M. Majidzadeh, A. Moilanen, N. Tervo, H. Pennanen, A. Tolli, and M. Latva-aho, "Partially connected hybrid beamforming for large antenna arrays in multi-user MISO systems," in *2017 IEEE 28th Annu. Int. Symp. Pers., Indoor, and Mobile Radio Commun. (PIMRC)*, 08–13 Oct. 2017, pp. 1–6.

[28] Q. Lyu, Z. Lin, Y. She, and C. Zhang, "A comparison of typical ℓp minimization algorithms," *Neurocomputing*, vol. 119, pp. 413–424, Nov. 2013.

[29] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓq-minimization for $0<q\leq 1$," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 395–407, May 2009.

[30] T. Ince and G. Ögücü, "Array failure diagnosis using nonconvex compressed sensing," *IEEE Antennas Wireless Propag. Lett.*, vol. 15, pp. 992–995, 2016.

[31] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[33] M. Elad, "Uniqueness and uncertainty," in *Sparse and Redundant Representations*. Springer, 2010, pp. 17–33.

[34] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012.

[35] Y. Fu, H. Li, Q. Zhang, and J. Zou, "Block-sparse recovery via redundant block OMP," *Signal Process.*, vol. 97, pp. 162–171, 2014.

[36] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.

[37] A. S. Bandeira, M. Fickus, D. G. Mixon, and P. Wong, "The road to deterministic matrices with the restricted isometry property," *J. Fourier Anal. Appl.*, vol. 19, no. 6, pp. 1123–1149, 2013.

[38] R. R. Naidu, P. Jampana, and C. S. Sastry, "Deterministic compressed sensing matrices: Construction via euler squares and applications," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3566–3575, 2016.

[39] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[40] T. Hong, H. Bai, S. Li, and Z. Zhu, "An efficient algorithm for designing projection matrix in compressive sensing based on alternating optimization," *Signal Process.*, vol. 125, pp. 9–20, 2016.

[41] D. Tse and P. Viswanath, "MIMO I: Spatial multiplexing and channel modeling," in *Fundamentals of Wireless Communication*. Cambridge Univ. Press, 2005, pp. 290–331.

[42] A. Ganesh, Z. Zhou, and Y. Ma, "Separation of a subspace-sparse signal: Algorithms and conditions," in *2009 IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2009, pp. 3141–3144.

[43] M. Fornasier and H. Rauhut, "Compressive sensing," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 1–48.

[44] Y. Huang, C. Liu, Y. Song, and X. Yu, "DFT codebook-based hybrid precoding for multiuser mmWave massive MIMO systems," *EURASIP J. Adv. Signal Process.*, vol. 2020, no. 11, pp. 1–13, Mar 2020.

[45] Y. Han, S. Jin, J. Zhang, J. Zhang, and K.-K. Wong, "DFT-based hybrid beamforming multiuser systems: Rate analysis and beam selection," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 514–528, Jun. 2018.

[46] J. Zyren and W. McCoy, "Overview of the 3GPP long term evolution physical layer," Freescale Semiconductor, Inc., White Paper, 2007.