Fast signal recovery from quadratic measurements

Miguel Moscoso, Alexei Novikov, George Papanicolaou and Chrysoula Tsogka

***Abstract*—We present a novel approach for recovering a sparse signal from cross-correlated data. Cross-correlations naturally arise in many fields of imaging, such as optics, holography and seismic interferometry. Compared to the sparse signal recovery problem that uses linear measurements, the unknown is now a matrix formed by the cross correlation of the unknown signal. Hence, the bottleneck for inversion is the number of unknowns that grows quadratically. The main idea of our proposed approach is to reduce the dimensionality of the problem by recovering only the diagonal of the unknown matrix, whose dimension grows linearly with the size of the problem. The keystone of the methodology is the use of an efficient *Noise Collector* that absorbs the data that come from the off-diagonal elements of the unknown matrix and that do not carry extra information about the support of the signal. This results in a linear problem whose cost is similar to the one that uses linear measurements. Our theory shows that the proposed approach provides exact support recovery when the data is not too noisy, and that there are no false positives for any level of noise. Moreover, our theory also demonstrates that when using cross-correlated data, the level of sparsity that can be recovered increases, scaling almost linearly with the number of data. The numerical experiments presented in the paper corroborate these findings.**

***Index Terms*—quadratic data, ℓ_1-minimization, noise, model reduction**

I. Introduction

Reconstruction of signals from cross correlations has interesting applications in many fields of science and engineering such as optics, quantum mechanics, electron microscopy, antenna testing, seismic interferometry, or imaging in general [9], [12], [21], [19]. Using cross correlations of measurements collected at different locations presents several advantages since the inversion does not require knowledge of the emitter positions, or the probing pulses shapes as only time differences matter. Cross correlations have been used, for example, when imaging is carried out with opportunistic sources whose properties are mainly unknown [7], [8], [6], [13].

In many applications, we seek information about an object or a signal $\boldsymbol{\rho} \in \mathcal{C}^K$ given data $\boldsymbol{b} \in \mathcal{C}^N$ most often related through a linear transformation

$$\mathcal{A}\,\boldsymbol{\rho} = \boldsymbol{b}\,, \tag{1}$$

where $\mathcal{A} \in \mathcal{C}^{N\times K}$ is the measurement or model matrix. When the signal $\boldsymbol{\rho}$ is compressed or when the data is scarce, $N < K$, in which case (1) is underdetermined and infinitely many signals or objects match the data. However, if the signal $\boldsymbol{\rho}$ is sparse so only $M \ll K$ components are different than zero, ℓ_1-minimization algorithms that solve

$$\boldsymbol{\rho}_{\ell_1} = \text{argmin}\,\|\boldsymbol{\rho}\|_{\ell_1},\ \text{subject to}\ \mathcal{A}\boldsymbol{\rho} = \boldsymbol{b} \tag{2}$$

can recover the true signal efficiently even when $N \ll K$.

On the other hand, there are situations in which it is difficult or impossible to record high quality data, $\boldsymbol{b}$, and it is more convenient to use the cross-correlated data contained in the matrix

$$B = \boldsymbol{b}\,\boldsymbol{b}^* \in \mathcal{C}^{N\times N} \tag{3}$$

to find the desired information about the object or signal $\boldsymbol{\rho}$ (see [7] and references therein).

One way to address this problem is to lift it to the matrix level and reformulate it as a low-rank matrix linear system, which can be solved by using nuclear norm minimization as it was suggested in [4], [3] for imaging with intensities-only. This makes the problem convex over the appropriate matrix vector space and, thus, the unique true solution can be found using well established algorithms with increased storage requirements as they involve the SVD of the iterate matrix unknown [1]. Furthermore, the big caveat is that the computational cost rapidly becomes prohibitive because the dimension of the problem increases quadratically with K, making its solution infeasible when the problem is large

In this paper we suggest a different approach. We propose to consider the linear matrix equation

$$\mathcal{A}X\mathcal{A}^* = B \tag{4}$$

for the correlated signal $X = \boldsymbol{\rho}\,\boldsymbol{\rho}^* \in \mathcal{C}^{K\times K}$, vectorize both sides so

$$\text{vec}(\mathcal{A}X\mathcal{A}^*) = \text{vec}(B)\,, \tag{5}$$

and use the Kronecker product $\otimes$, and its property $\text{vec}(PQR) = (R^T \otimes P)\text{vec}(Q)$, to express the matrix multiplications as the linear transformation

$$(\bar{\mathcal{A}} \otimes \mathcal{A})\,\text{vec}(X) = \text{vec}(B)\,. \tag{6}$$

Thus, we can promote the sparsity of the sought signal using ℓ_1-minimization algorithms that are more efficient than nuclear norm minimization ones, as they do not require heavy operations that involve matrix factorizations. However, the dimension of the unknown $\text{vec}(X)$ in (6) also increases quadratically with K, so this approach by itself would still be impractical when K is not small.

Hence, we propose to use a *Noise Collector* to reduce the dimensionality of problem (6). The *Noise Collector* was introduced in [17] to eliminate the clutter in the recovered signals when the data are contaminated by additive noise. In

M. Moscoso is with the Department of Mathematics, Universidad Carlos III de Madrid, Leganes, Madrid 28911, Spain e-mail: moscoso@math.uc3m.es.

A. Novikov is with the Mathematics Department, Penn State University, University Park, PA 16802, USA e-mail: novikov@psu.edu

G. Papanicolaou is with the Department of Mathematics, Stanford University, Stanford, CA 94305 USA e-mail: papanicolaou@stanford.edu

C. Tsogka is with the Applied Math Unit, University of California, Merced, 5200 North Lake Road, Merced, CA 95343 USA e-mail: ctsogka@ucmerced.edu

this paper, we use the Noise Collector to absorb part of the signal instead. Specifically, we treat the signal that corresponds to the $K^2 - K$ off-diagonal entries in the matrix X as noise. Using the *Noise Collector* allows us to ignore these entries and construct a linear system with the same number of unknowns as the original problem (1) that uses linear data. Thus, a dimension reduction from K^2 to K unknowns is achieved with almost no extra computational cost (see Section IV-A), as the number of operations used by the *Noise Collector* is of the same order as the ones needed to solve the original problem with linear measurements. In other words, the cost of solving the sparse signal recovery problem using cross-correlated data also grows linearly with K using the proposed approach.

The main result of this paper is Theorem 3 which says that under certain decoherence conditions on the matrix $\mathcal{A}$, we can find the support of an M-sparse signal exactly if the data is noise-free or the noise is low enough. Furthermore, Theorem 3 shows that the level of sparsity M that can be recovered increases from $O(\sqrt{N}/\sqrt{\ln N})$ to $O(N/\sqrt{\ln N})$ when quadratic cross correlation data are used instead of the linear ones.

The numerical experiments included in this paper support the results of Theorem 3. They show that the support of a signal can be found exactly if the noise in the data is not too large with almost no extra computational cost with respect to the original problem (1) that considers linear data with no correlations. Once the support has been found, a trivial second step allows us to find the signal, including its phases. The reconstruction is exact when there is no noise in the data and the results are very satisfactory even for noisy data with low signal to noise ratios. That is, our numerical experiments suggest that the approach presented here is robust with respect to additive noise. Additional properties of this approach are that for any level of noise the solution has no false positives, and that the algorithm is parameter-free, so it does not require an estimation of the energy of the *off-diagonal signal* that we need to absorb, or of the level of noise in the data.

The paper is organized as follows. In Section II, we summarize the model used to generate the signals to be recovered, which in our case are images. In Section III, we present the theory that supports the proposed strategy for dimension reduction when correlated data are used to recover the signals. Section IV explains the algorithm for carrying out the inversion efficiently. Section V shows the numerical experiments. Section VI summarizes our conclusions. The proofs of the theorems are given in A.

II. Passive array imaging

We consider processing of passive array signals where the object to be imaged is a set of point sources at positions $\vec{\boldsymbol{z}}_j$ and (complex) amplitudes α_j, $j = 1, \ldots, M$. The data used to image the object are collected at several sensors on an array; see Figure 1. The imaging system is characterized by the array aperture a, the distance L to the sources, the bandwidth B and the central wavelength λ_0 of the signals.

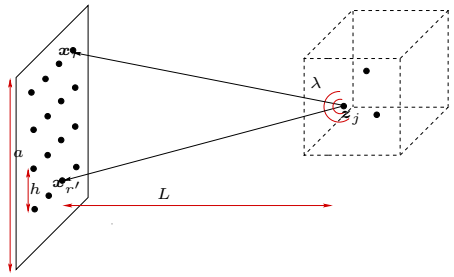

Fig. 1. General setup for passive array imaging. The source at $\vec{\boldsymbol{z}}_j$ emits a signal that is recorded at all array elements $\vec{\boldsymbol{x}}_r$, $r = 1, \ldots, N_r$.

The sources are located inside an image window IW discretized with a uniform grid of points $\vec{\boldsymbol{y}}_k$, $k = 1, \ldots, K$. Thus, the signal to be recovered is the source vector

$$\tilde{\boldsymbol{\rho}} = [\tilde{\rho}_1, \ldots, \tilde{\rho}_K]^\intercal \in \mathbb{C}^K \,, \tag{7}$$

whose components $\tilde{\rho}_k$ correspond to the amplitudes of the M sources at the grid points $\vec{\boldsymbol{y}}_k$, $k = 1, \ldots, K$, with $K \gg M$. This vector has components $\tilde{\rho}_k = \alpha_j$ if $\vec{\boldsymbol{y}}_k = \vec{\boldsymbol{z}}_j$ for some $j = 1, \ldots, M$, while the others are zero.

Denoting by $G(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}; \omega)$ the Green's function for the propagation of a wave of angular frequency ω from point $\vec{\boldsymbol{y}}$ to point $\vec{\boldsymbol{x}}$, we define the single-frequency Green's function vector that connects a point $\vec{\boldsymbol{y}}$ in the IW with all the sensors on the array located at points $\vec{\boldsymbol{x}}_r$, $r = 1, \ldots, N_r$, so

$$\boldsymbol{g}(\vec{\boldsymbol{y}}; \omega) = [G(\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{y}}; \omega), G(\vec{\boldsymbol{x}}_2, \vec{\boldsymbol{y}}; \omega), \ldots, G(\vec{\boldsymbol{x}}_N, \vec{\boldsymbol{y}}; \omega)]^\intercal \in \mathbb{C}^{N_r} \,.$$

In three dimensions, $G(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}; \omega) = \dfrac{\exp\{\mathrm{i}\omega|\vec{\boldsymbol{x}} - \vec{\boldsymbol{y}}|/c_0\}}{4\pi|\vec{\boldsymbol{x}} - \vec{\boldsymbol{y}}|}$ if the medium is homogeneous. Hence, the signals of frequencies ω_l recorded at the sensors locations $\vec{\boldsymbol{x}}_r$ are

$$b(\vec{\boldsymbol{x}}_r, \omega_l) = \sum_{j=1}^{M} \alpha_j G(\vec{\boldsymbol{x}}_r, \vec{\boldsymbol{z}}_j; \omega_l) \,, \quad r = 1, \ldots, N_r \,.$$

They form the single-frequency data vector $\boldsymbol{b}(\omega_l) = [b(\vec{\boldsymbol{x}}_1, \omega_l), b(\vec{\boldsymbol{x}}_2, \omega_l), \ldots, b(\vec{\boldsymbol{x}}_N, \omega_l)]^\intercal \in \mathbb{C}^{N_r}$. As several frequencies ω_l, $l = 1, \ldots, N_f$, are used to recover (7), all the recorded data are stacked in the multi-frequency column data vector

$$\boldsymbol{b} = [\boldsymbol{b}(\omega_1)^\intercal, \boldsymbol{b}(\omega_2)^\intercal, \ldots, \boldsymbol{b}(\omega_{N_f})^\intercal]^\intercal \in \mathbb{C}^N \,, \text{with } N = N_r N_f \,. \tag{8}$$

A. The inverse problem with linear data

When the data (8) are available and reliable, one can form the linear system

$$\mathcal{A}\,\boldsymbol{\rho} = \boldsymbol{b} \tag{9}$$

to recover (7). Here, $\mathcal{A}$ is the $N \times K$ measurement matrix whose columns $\boldsymbol{a}_k$ are the multi-frequency Green's function vectors

$$\boldsymbol{a}_k = \frac{1}{c_k} [\boldsymbol{g}(\vec{\boldsymbol{y}}_k; \omega_1)^\intercal, \boldsymbol{g}(\vec{\boldsymbol{y}}_k; \omega_2)^\intercal, \ldots, \boldsymbol{g}(\vec{\boldsymbol{y}}_k; \omega_{N_f})^\intercal]^\intercal \in \mathbb{C}^N \,, \tag{10}$$

where c_k are scalars that normalize these vectors to have ℓ_2-norm one, and

$$\boldsymbol{\rho} = \text{diag}(c_1, c_2, \ldots, c_K)\, \tilde{\boldsymbol{\rho}}, \tag{11}$$

where $\tilde{\boldsymbol{\rho}}$ is given by (7). Then, one can solve (9) for the unknown vector $\boldsymbol{\rho}$ using a number of ℓ_2 and ℓ_1 inversion methods to find the sought image. In general, ℓ_2 methods are robust but the resulting resolution is low. On the other hand, ℓ_1 methods provide higher resolution but they are much more sensitive to noise in the data. Hence, they cannot be used with poor quality data unless one carefully takes care of the noise.

B. The inverse problem with quadratic cross correlation data

In many instances, imaging with cross correlations helps to form better and more robust images. This is the case, for example, when one uses high frequency signals and has a low-budget measurement system with inexpensive sensors that are not able to resolve the signals well. Another situation is when the raw data (8) can be measured but it is more convenient to image with cross correlations because they help to mitigate the effects of the inhomogeneities of the medium between the sources and the sensors [2], [10]

Assume that all the cross-correlated data contained in the matrix

$$B = \boldsymbol{b}\, \boldsymbol{b}^* \in \mathcal{C}^{N\times N} \tag{12}$$

are available for imaging. Then, one can consider the linear system

$$\mathcal{A} X \mathcal{A}^* = B\,, \tag{13}$$

and seek the correlated image $X = \boldsymbol{\rho}\, \boldsymbol{\rho}^* \in \mathcal{C}^{K\times K}$ that solves it. The unknown matrix X is rank 1 and, hence, one possibility is to look for a low-rank matrix by using nuclear norm minimization as it was suggested for imaging with intensities-only in [4], [3]. This is possible in theory, but it is unfeasible when the problem is large because the number of unknowns grows quadratically and, therefore, the computational cost rapidly becomes prohibitive. For example, to form an image with 1000×1000 pixels one would have to solve a system with 10^{12} unknowns.

Instead, we suggest the following strategy. We propose to vectorize both sides of (13) so

$$\text{vec}(\mathcal{A} X \mathcal{A}^*) = \text{vec}(B)\,, \tag{14}$$

where $\text{vec}(\cdot)$ denotes the vectorization of a matrix formed by stacking its columns into a single column vector. Then, we use the Kronecker product $\otimes$, and its property $\text{vec}(PQR) = (R^T \otimes P)\text{vec}(Q)$, to express the matrix multiplications as the linear transformation

$$(\bar{\mathcal{A}} \otimes \mathcal{A})\, \text{vec}(X) = \text{vec}(B)\,. \tag{15}$$

With this formulation of the problem we can use an ℓ_1 minimization algorithm to form the images, which is much faster than a nuclear norm minimization algorithm that needs to compute the SVD of the iterate matrices. However, with just this approach the main obstacle is not overcome, as the dimensionality still grows quadratically with the number of unknowns K. Hence, we propose a dimension reduction strategy that uses a *Noise Collector* [17] to absorb a component of the data vector that does not provide extra information about the signal support. We point out that this component is not a gaussian random vector as in [17], but a deterministic vector resulting from the off-diagonal terms of X that are neglected.

III. The noise collector and dimension reduction

A. The Noise Collector

The *Noise Collector* [17] is a method to find the vector $\boldsymbol{\chi} \in \mathbb{C}^{\mathcal{K}}$ in

$$T\, \boldsymbol{\chi} = \boldsymbol{d}_0 + \boldsymbol{e}\,, \tag{16}$$

from highly incomplete measurement data $\boldsymbol{d} = \boldsymbol{d}_0 + \boldsymbol{e} \in \mathbb{C}^{\mathcal{N}}$ possibly corrupted by noise $\boldsymbol{e} \in \mathbb{C}^{\mathcal{N}}$, where $1 \ll \mathcal{N} < \mathcal{K}$. Here, T is a general measurement matrix of size $\mathcal{N}\times\mathcal{K}$, whose columns have unit length. The main results in [17] ensure that we can still recover the support of $\boldsymbol{\chi}$ when the data is noisy by looking at the support of $\boldsymbol{\chi}_\tau$ found as

$$\begin{array}{c}(\boldsymbol{\chi}_\tau, \boldsymbol{\eta}_\tau) = \arg\min_{\boldsymbol{\chi},\boldsymbol{\eta}} \left(\tau \|\boldsymbol{\chi}\|_{\ell_1} + \|\boldsymbol{\eta}\|_{\ell_1}\right), \\ \text{subject to } T\boldsymbol{\chi} + \mathcal{C}\boldsymbol{\eta} = \boldsymbol{d},\end{array} \tag{17}$$

with an $O(1)$ no-phantom weight τ, and a *Noise Collector* matrix $\mathcal{C} \in \mathbb{C}^{\mathcal{N}\times\Sigma}$ with $\Sigma = \mathcal{N}^\beta$, for $\beta > 1$. If the noise $\boldsymbol{e}$ is Gaussian, then the columns of $\mathcal{C}$ can be chosen independently and at random on the unit sphere $\mathbb{S}^{\mathcal{N}-1} = \left\{x \in \mathbb{R}^{\mathcal{N}}, \|x\|_{\ell_2} = 1\right\}$. The weight $\tau > 1$ is chosen so it is expensive to approximate $\boldsymbol{e}$ with the columns of T, but it cannot be taken too large because then we lose the signal $\boldsymbol{\chi}$ that gets absorbed by the *Noise Collector* as well. Intuitively, τ is a measure of the rate at which the signal is lost as the noise increases. For practical purposes, τ is chosen as the minimal value for which $\boldsymbol{\chi} = 0$ when the data is pure noise, i.e., when $\boldsymbol{d}_0 = 0$. The key property is that the optimal value of τ does not depend on the level of noise and, therefore, it is chosen in advance, before the *Noise Collector* is used for a specific task. We have the following result.

Theorem 1: [17] Fix $\beta > 1$, and draw $\Sigma = \mathcal{N}^\beta$ columns to form the Noise Collector $\mathcal{C}$, independently, from the uniform distribution on $\mathbb{S}^{\mathcal{N}-1}$. Let $\boldsymbol{\chi}$ be an M-sparse solution of the noiseless system $T\boldsymbol{\chi} = \boldsymbol{d}_0$, and $\boldsymbol{\chi}_\tau$ the solution of (17) with $\boldsymbol{d} = \boldsymbol{d}_0 + \boldsymbol{e}$. Denote the ratio of minimum to maximum significant values of $\boldsymbol{\chi}$ as

$$\gamma = \min_{i\in \text{supp}(\boldsymbol{\chi})} \frac{|\chi_i|}{\|\boldsymbol{\chi}\|_{\ell_\infty}}. \tag{18}$$

Assume that the columns of T are incoherent, so that

$$|\langle \boldsymbol{t}_i, \boldsymbol{t}_j \rangle| \leqslant \frac{1}{3M} \text{ for all } i \text{ and } j. \tag{19}$$

Then, for any $\kappa > 0$, there are constants $\tau = \tau(\kappa, \beta)$, $c_1 = c_1(\kappa, \beta, \gamma)$, and $\mathcal{N}_0 = \mathcal{N}_0(\kappa, \beta)$ such that, if the noise level satisfies

$$\max\left(1, \|\boldsymbol{e}\|_{\ell_2}\right) \leqslant c_1 \frac{\|\boldsymbol{d}_0\|^2_{\ell_2}}{\|\boldsymbol{\chi}\|_{\ell_1}} \sqrt{\frac{\mathcal{N}}{\ln \mathcal{N}}}, \tag{20}$$

then $\text{supp}(\boldsymbol{\chi}_\tau) = \text{supp}(\boldsymbol{\chi})$ for all $\mathcal{N} > \mathcal{N}_0$ with probability $1 - 1/\mathcal{N}^\kappa$.

To gain a better understanding of this theorem, let us consider the case where T is the identity matrix (the classical

denoising problem) and all coefficients of $\boldsymbol{d}_0 = \boldsymbol{\chi}$ are either 1 or 0. Then $\|\boldsymbol{d}_0\|_{\ell_2}^2 = \|\boldsymbol{\chi}\|_{\ell_1} = M$. In this case, an acceptable level of noise is

$$\|\boldsymbol{e}\|_{\ell_2} \lesssim \|\boldsymbol{d}_0\|_{\ell_2} \sqrt{\frac{\mathcal{N}}{M \ln \mathcal{N}}} \sim \sqrt{\frac{\mathcal{N}}{\ln \mathcal{N}}}. \tag{21}$$

The estimate (21) implies that we can handle more noise as we increase the number of measurements. This holds for two reasons. Firstly, a typical noise vector $\boldsymbol{e}$ is almost orthogonal to the columns of T, so

$$|\langle \boldsymbol{t}_i, \boldsymbol{e} \rangle| \leqslant c_0 \sqrt{\frac{\ln \mathcal{N}}{\mathcal{N}}} \|\boldsymbol{e}\|_{\ell_2} \tag{22}$$

for some $c_0 = c_0(\kappa)$ with probability $1 - 1/\mathcal{N}^\kappa$. In particular, a typical noise vector $\boldsymbol{e}$ is almost orthogonal to the signal subspace V. More formally, suppose V is the M-dimensional subspace spanned by the column vectors $\boldsymbol{t}_j$ with j in the support of $\boldsymbol{\chi}$, and let $W = V^\perp$ be the orthogonal complement to V. Consider the orthogonal decomposition $\boldsymbol{e} = \boldsymbol{e}^v + \boldsymbol{e}^w$, such that $\boldsymbol{e}^v$ is in V and $\boldsymbol{e}^w$ is in W. Then,

$$\|\boldsymbol{e}^v\|_{\ell_2} \lesssim \sqrt{\frac{M}{\mathcal{N}}} \|\boldsymbol{e}\|_{\ell_2}$$

with high probability that tends to 1, as $\mathcal{N} \to \infty$. In Theorem 1, a quantitative estimate of this convergence is $1 - 1/\mathcal{N}^\kappa$. It means that if a signal is sparse so $M \ll \mathcal{N}$, then we can recover it for very low signal-to-noise ratios. Secondly, and more importantly, if the columns of the noise collector $\mathcal{C}$ are also almost orthogonal to the signal subspace, then it is too expensive to approximate the signal $\boldsymbol{d}_0$ with the columns of $\mathcal{C}$ and, hence, we have to use the columns of the measurement matrix T. If we draw the columns of $\mathcal{C}$, independently, from the uniform distribution on $\mathbb{S}^{\mathcal{N}-1}$, then they will be almost orthogonal to the signal subspace with high probability. It is again estimated as $1 - 1/\mathcal{N}^\kappa$ in Theorem 1. Finally, the incoherence condition (19) implies that it is too expensive to approximate the signal $\boldsymbol{d}_0$ with columns T that are not in the support of $\boldsymbol{\chi}$ and, hence, there are no false positives.

In Theorem 1 we used randomness twice: the noise vector $\boldsymbol{e}$ was random and the columns of the noise collector were drawn at random. Note that in both cases randomness could be replaced by deterministic conditions requiring that $\boldsymbol{e}$ and the columns of $\mathcal{C}$ are almost orthogonal to the signal subspace. It is natural to assume that the noise vector $\boldsymbol{e}$ is a random variable and, as we explain in [17], the columns of $\mathcal{C}$ are random because it is hard to construct a deterministic $\mathcal{C}$ that satisfies the almost orthogonality conditions. In the present work we still construct the matrix $\mathcal{C}$ randomly, but we sometimes treat the vector $\boldsymbol{e}$ as deterministic, as for example, in our Theorem 2. Inspection of the proofs in [17] shows that the only condition on $\boldsymbol{e}$ we need to verify from Theorem 1 is (22). Thus, the next Theorem is a deterministic reformulation of Theorem 1. The proof is given in A.

Theorem 2: Assume conditions on $\boldsymbol{\chi}$, T, and $\mathcal{C}$ are as in Theorem 1 and define γ as in (18). Then, for any $\kappa > 0$, there are constants $\tau_0 = \tau_0(\kappa, \beta)$, $c_0 = c_0(\kappa, \beta)$, and $\mathcal{N}_0 = \mathcal{N}_0(\kappa, \beta, \gamma)$, $\alpha = \alpha(c_0, \kappa, \beta)$ such that the following two claims hold.

(i) If $\boldsymbol{e}$ satisfies (22) for all $\boldsymbol{t}_i$, $i \notin \text{ supp } (\boldsymbol{\chi})$; all columns of T satisfy

$$|\langle \boldsymbol{t}_i, \boldsymbol{t}_j \rangle| \leqslant c_0 \frac{\sqrt{\ln \mathcal{N}}}{\sqrt{\mathcal{N}}} \tag{23}$$

for all i and j; the sparsity M is such that

$$M \leqslant \alpha \frac{\sqrt{\mathcal{N}}}{\sqrt{\ln \mathcal{N}}}; \tag{24}$$

and $\tau \geqslant \tau_0$, then $\text{supp}(\boldsymbol{\chi}_\tau) \subset \text{supp}(\boldsymbol{\chi})$ with probability $1 - 1/\mathcal{N}^\kappa$.

(ii) If, in addition, the noise is not large, so

$$|\langle \boldsymbol{t}_m, \boldsymbol{e} \rangle| \leqslant \min_{i \in \text{supp}(\boldsymbol{\chi})} |\chi_i|/2 \tag{25}$$

for all $\boldsymbol{t}_m$, $m \in \text{supp}(\boldsymbol{\chi})$, and

$$\|\boldsymbol{e}\|_{\ell_2} \leqslant c_1 \|\boldsymbol{\chi}\|_{\ell_1} \tag{26}$$

for some c_1, then $\text{supp}(\boldsymbol{\chi}) = \text{supp}(\boldsymbol{\chi}_\tau)$ for all $\mathcal{N} > \mathcal{N}_0$ with probability $1 - 1/\mathcal{N}^\kappa$.

In contrast to Theorem 1, we require in Theorem 2 condition (22) to hold only for for $\boldsymbol{t}_i$, $i \notin \text{ supp } (\boldsymbol{\chi})$, that is for the columns of T outside the support of $\boldsymbol{\chi}$. For the columns inside the support, $i \in \text{ supp } (\boldsymbol{\chi})$, we relax condition (22) to condition (25). Thus Theorem 2 has slightly weaker assumptions than Theorem 1. For a random $\boldsymbol{e}$ this weakening in not essential, because one needs to know the support of $\boldsymbol{\chi}$ in advance. It turns out that for our $\boldsymbol{e}$ this weakening will become important (see Remark 1 in the end of B) .

B. Dimension reduction for quadratic cross correlation data

The $N^2 \times K^2$ linear problem (15) that uses quadratic cross correlation data is notoriously hard to solve due to its high dimensionality. Therefore, we propose the following strategy for robust dimensionality reduction. The idea is to treat the contribution of the off-diagonal elements of $X = \boldsymbol{\rho}\,\boldsymbol{\rho}^* \in \mathbb{C}^{K \times K}$ as *noise* and, thus, use the *Noise Collector* to absorb it. Namely, we define

$$\boldsymbol{\chi} = \text{diag}(X) = [|\rho_1|^2, |\rho_2|^2, \ldots, |\rho_K|^2]^T, \tag{27}$$

and re-write (15) as

$$T\,\boldsymbol{\chi} + \mathcal{C}\,\boldsymbol{\eta} = \boldsymbol{d}, \tag{28}$$

where we replace the off-diagonal elements by the Noise Collector term $\mathcal{C}\,\boldsymbol{\eta}$ and

$$T = (\bar{\mathcal{A}} \otimes \mathcal{A})_{\boldsymbol{\chi}} \tag{29}$$

contains only the K columns of $\bar{\mathcal{A}} \otimes \mathcal{A}$ corresponding to $\boldsymbol{\chi}$. Thus, the size of $\boldsymbol{\chi}$ is $\mathcal{K}$ and the size of T is $\mathcal{N} \times \mathcal{K}$, with $\mathcal{K} = K$ and $\mathcal{N} = N^2$. In practice, the measurements may be subsampled as well, so the size of the system can be further reduced to $\mathcal{N} \times \mathcal{K}$, with $\mathcal{N} = O(N)$ and $\mathcal{K} = K$.

Problem (28) can be understood as an exact linearization of the classical phase retrieval problem, where all the interference terms $\rho_i \rho_j^*$ for $i \neq j$ are absorbed in $\mathcal{C}\,\boldsymbol{\eta}$, with $\boldsymbol{\eta}$ being an unwanted vector considered to be noise in this formulation. In other words, the phase retrieval problem with K unknowns has been transformed to the linear problem (28) that also

has K unknowns. Note, though, that in phase retrieval only autocorrelation measurements are considered, while in (28) we also use cross-correlation measurements.

In the next theorem we use all the measurements $\boldsymbol{d} \in \mathbb{C}^{\mathcal{N}}$, so $\mathcal{N} = N^2$ in (28). This is done for simplicity of presentation, but in practice $\mathcal{N} = O(N)$ measurements are enough. We will choose a solution of (28) using (17). As in Theorems 1 and 2, the vector $\boldsymbol{\eta}$ in (28) has $\mathcal{N}^\beta$ entries that do not have physical meaning. Its only purpose is to absorb the off-diagonal contributions in $\boldsymbol{e} = \boldsymbol{d} - T\boldsymbol{\chi}$. We point out that the magnitude of $\boldsymbol{e}$ is not small if $M \geq 2$. Indeed, the contribution of $\boldsymbol{\chi} = \mathrm{diag}(X)$ to the data $\boldsymbol{d}$ is of order M, while the contribution of the off-diagonal terms of X is of order M^2. Furthermore, the vector $\boldsymbol{e}$ is not independent of $\boldsymbol{\chi}$ anymore.

Theorem 3: Fix $|\rho_i|$. Suppose the phases $\rho_i/|\rho_i|$ are independent and uniformly distributed on the (complex) unit circle. Suppose X is a solution of (15), $\boldsymbol{\chi} = \mathrm{diag}(X)$ is M-sparse, and $T = (\bar{\mathcal{A}} \otimes \mathcal{A})_{\boldsymbol{\chi}} : \mathbb{C}^{\mathcal{K}} \to \mathbb{C}^{\mathcal{N}}$, $\mathcal{K} = K$ and $\mathcal{N} = N^2$. Fix $\beta > 1$, and draw $\Sigma = \mathcal{N}^\beta$ columns for $\mathcal{C}$, independently, from the uniform distribution on $\mathbb{S}^{\mathcal{N}-1}$. Denote

$$\Delta = \sqrt{N} \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|, \tag{30}$$

and define γ as in (18). Then, for any $\kappa > 0$, there are constants $\alpha = \alpha(\kappa, \gamma, \Delta)$, $\tau = \tau(\kappa, \beta)$, and $\mathcal{N}_0 = \mathcal{N}_0(\kappa, \beta, \gamma, \Delta)$ such that the following holds. If $M \leqslant \alpha N/\sqrt{\ln N}$ and $\boldsymbol{\chi}_\tau$ is the solution of (17), then $\mathrm{supp}(\boldsymbol{\chi}) = \mathrm{supp}(\boldsymbol{\chi}_\tau)$ for all $\mathcal{N} > \mathcal{N}_0$ with probability $1 - 1/\mathcal{N}^\kappa$.

The proof of Theorem 3 is given in B. In Theorem 3 the scaling for sparse recovery is $M \leqslant \alpha N/\sqrt{\ln N}$. This result is in good agreement with our numerical experiments, see Figure 7. In order to obtain this scaling we introduced our probabilistic framework - in Theorem 3 assuming that the phases of the signals are random. The idea is that a vector with random phases better describes a typical signal in many applications. The dimension reduction, however, could be done without introducing the probabilistic framework. We state and prove a deterministic version of Theorem 3 in C for completeness. In this case the scaling for sparse recovery is more conservative: $M \leqslant \alpha\sqrt{N}/\sqrt{\ln N}$, and it does not agree with our numerical experiments.

IV. Algorithmic implementation

A key point of the proposed strategy is that the M-sparsest solution of (28) can be effectively found by solving the minimization problem

$$(\boldsymbol{\chi}_\tau, \boldsymbol{\eta}_\tau) = \arg\min_{\boldsymbol{\chi},\boldsymbol{\eta}} \left(\tau\|\boldsymbol{\chi}\|_{\ell_1} + \|\boldsymbol{\eta}\|_{\ell_1}\right), \quad \text{subject to } T\,\boldsymbol{\chi} + \mathcal{C}\boldsymbol{\eta} = \boldsymbol{d}, \tag{31}$$

with an $O(1)$ no-phatom weight τ. Here, T is an $\mathcal{N} \times \mathcal{K}$ matrix, $\mathcal{C}$ an $\mathcal{N} \times \mathcal{N}^\beta$ matrix, with β close to one, $\boldsymbol{\chi}$ is a $\mathcal{K} \times 1$ vector and $\boldsymbol{\eta}$ is a $\mathcal{N}^\beta \times 1$ vector. The main property of this approach is that if the matrix T is incoherent enough, so its columns satisfy assumption (19) of Theorem 1, the ℓ_1-norm minimal solution of (31) has a zero false discovery rate for any level of noise, with probability that tends to one as the dimension of the data $\mathcal{N}$ increases to infinity.

To find the minimizer in (31), we define the function

$$\begin{aligned} F(\boldsymbol{\chi}, \boldsymbol{\eta}, \boldsymbol{z}) &= \lambda\left(\tau\|\boldsymbol{\chi}\|_{\ell_1} + \|\boldsymbol{\eta}\|_{\ell_1}\right) \\ &+ \frac{1}{2}\|T\boldsymbol{\chi} + \mathcal{C}\boldsymbol{\eta} - \boldsymbol{d}\|_{\ell_2}^2 + \langle \boldsymbol{z}, \boldsymbol{d} - T\boldsymbol{\chi} - \mathcal{C}\boldsymbol{\eta}\rangle \end{aligned} \tag{32}$$

for a no-phantom weight τ, and determine the solution as

$$\max_{\boldsymbol{z}} \min_{\boldsymbol{\chi},\boldsymbol{\eta}} F(\boldsymbol{\chi}, \boldsymbol{\eta}, \boldsymbol{z}). \tag{33}$$

This strategy finds the minimum in (31) exactly for all values of the regularization parameter λ. Thus, the method is fully automated, meaning that it has no tuning parameters. To determine the exact extremum in (33), we use the iterative soft thresholding algorithm GeLMA [16] that works as follows.

Pick a value for the no-phantom weight τ; for optimal results calibrate τ to be the smallest value for which $\boldsymbol{\chi} = 0$ when the algorithm is fed with pure noise. In our numerical experiments we use $\tau = 2$. Next, pick a value for the regularization parameter, for example $\lambda = 1$, and choose step sizes $\Delta t_1 < 2/\|[T\,|\,\mathcal{C}]\|^2$ and $\Delta t_2 < \lambda/\|T\|^1$. Set $\boldsymbol{\chi}_0 = \mathbf{0}$, $\boldsymbol{\eta}_0 = \mathbf{0}$, $\boldsymbol{z}_0 = \mathbf{0}$, and iterate for $k \geqslant 0$:

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{d} - T\,\boldsymbol{\chi}_k - \mathcal{C}\,\boldsymbol{\eta}_k\,, \\ \boldsymbol{\chi}_{k+1} &= \mathcal{S}_{\tau\,\lambda\Delta t_1}(\boldsymbol{\chi}_k + \Delta t_1\, T^*(\boldsymbol{z}_k + \boldsymbol{r}))\,, \\ \boldsymbol{\eta}_{k+1} &= \mathcal{S}_{\lambda\Delta t_1}(\boldsymbol{\eta}_k + \Delta t_1\, \mathcal{C}^*(\boldsymbol{z}_k + \boldsymbol{r}))\,, \\ \boldsymbol{z}_{k+1} &= \boldsymbol{z}_k + \Delta t_2\, \boldsymbol{r}\,, \end{aligned} \tag{34}$$

where $\mathcal{S}_r(y_i) = \mathrm{sign}(y_i)\max\{0, |y_i| - r\}$.

A. The Noise Collector: construction and properties

To construct the *Noise Collector* matrix $\mathcal{C} \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}^\beta}$ that satisfies the assumptions of Theorem 1 one could draw $\mathcal{N}^\beta$ normally distributed $\mathcal{N}$-dimensional vectors, normalized to unit length. Thus, the additional computational cost incurred for implementing the *Noise Collector* in (34), due to the terms $\mathcal{C}\boldsymbol{\eta}_k$ and $\mathcal{C}^*(\boldsymbol{z}_k + \boldsymbol{r})$, would be $O(\mathcal{N}^{\beta+1})$, which is not very large as we use $\beta \approx 1.5$ in practice. The computational cost of (34) without the *Noise Collector* mainly comes from the matrix vector multiplications $T\,\boldsymbol{\chi}_k$ which can be done in $O(\mathcal{N}\mathcal{K})$ operations and, typically, $\mathcal{K} \gg \mathcal{N}$.

To further reduce the additional computational time and memory requirements we use a different construction procedure that exploits the properties of circulant matrices. The idea is to draw instead a few normally distributed $\mathcal{N}$-dimensional vectors of length one, and construct from each one of them a circulant matrix of dimension $\mathcal{N} \times \mathcal{N}$. The columns of these matrices are still independent and uniformly distributed on $\mathbb{S}^{\mathcal{N}-1}$, so they satisfy the assumptions of Theorem 1. The full *Noise Collector* matrix is then formed by concatenating these circulant matrices together.

More precisely, the *Noise Collector* construction is done in the following way. We draw $\mathcal{N}^{\beta-1}$ normally distributed $\mathcal{N}$-dimensional vectors, normalized to unit length. These are the generating vectors of the *Noise Collector*. To these vectors are associated $\mathcal{N}^{\beta-1}$ circulant matrices $\mathcal{C}_i \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$, $i =$

[1]Choosing two step sizes instead of the smaller one Δt_1 improves the convergence speed.

$1, \ldots, \mathcal{N}^{\beta-1}$, and the *Noise Collector* matrix is constructed by concatenation of these $\mathcal{N}^{\beta-1}$ matrices, so

$$\mathcal{C} = [\mathcal{C}_1 \,|\mathcal{C}_2 \,|\mathcal{C}_3 \,|\ldots \,|\mathcal{C}_{\mathcal{N}^{\beta-1}}] \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}^\beta}.$$

We point out that the *Noise Collector* matrix $\mathcal{C}$ is not stored, only the $\mathcal{N}^{\beta-1}$ generating vectors are saved in memory. On the other hand, the matrix vector multiplications $\mathcal{C}\boldsymbol{\eta}_k$ and $\mathcal{C}^*(\boldsymbol{z}_k + \boldsymbol{r})$ in (34) can be computed using these generating vectors and FFTs [11]. This makes the complexity associated to the *Noise Collector* $O(\mathcal{N}^\beta \log(\mathcal{N}))$.

To explain this further, we recall briefly below how a matrix vector multiplication can be performed using the FFT for a circulant matrix. For a generating vector $\boldsymbol{c} = [c_0, c_1, \ldots, c_{\mathcal{N}-1}]$, the $\mathcal{C}_i$ circulant matrix takes the form

$$\mathcal{C}_i = \begin{bmatrix} c_0 & c_{\mathcal{N}-1} & \ldots & c_1 \\ c_1 & c_0 & \ldots & c_2 \\ \vdots & & \ddots & \vdots \\ c_{\mathcal{N}-1} & c_{\mathcal{N}-2} & \ldots & c_0 \end{bmatrix}.$$

This matrix can be diagonalized by the Discrete Fourier Transform (DFT) matrix, i.e.,

$$\mathcal{C}_i = \mathcal{F}\Lambda\mathcal{F}^{-1}$$

where $\mathcal{F}$ is the DFT matrix, $\mathcal{F}^{-1}$ is its inverse, and Λ is a diagonal matrix such that $\Lambda = \text{diag}(\mathcal{F}\boldsymbol{c})$, where $\boldsymbol{c}$ is the generating vector. Thus, a matrix vector multiplication $\mathcal{C}_i\boldsymbol{\eta}$ is performed as follows: (i) compute $\hat{\boldsymbol{\eta}} = \mathcal{F}^{-1}\boldsymbol{\eta}$, the inverse DFT of $\boldsymbol{\eta}$ in $\mathcal{N}\log(\mathcal{N})$ operations, (ii) compute the eigenvalues of $\mathcal{C}_i$ as the DFT of $\boldsymbol{c}$, and component wise multiply the result with $\hat{\boldsymbol{\eta}}$ (this step can also be done in $\mathcal{N}\log(\mathcal{N})$ operations), and (iii) compute the FFT of the vector resulting from step (ii) in, again, $\mathcal{N}\log(\mathcal{N})$ operations.

Consequently, the cost of performing the multiplication $\mathcal{C}\boldsymbol{\eta}_k$ is $\mathcal{N}^{\beta-1}\mathcal{N}\log(\mathcal{N}) = \mathcal{N}^\beta\log(\mathcal{N})$. As the cost of finding the solution without the *Noise Collector* is $O(\mathcal{N}\mathcal{K})$ due to the terms $T\boldsymbol{\chi}_k$, the additional cost due to the *Noise Collector* is negligible since $\mathcal{K} \gg \mathcal{N}^{\beta-1}\log(\mathcal{N})$ because, typically, $\mathcal{K} \gg \mathcal{N}$ and $\beta \approx 1.5$.

We emphasise that in the dimension reduced signal recovery problem (28) that uses cross-correlated data, $\mathcal{K} = K$ and $\mathcal{N} = O(N)$, so the computational cost of finding its sparsest solution is $O(NK)$, as the original problem that uses linear data.

V. Numerical results

We consider processing of passive array signals. We seek to determine the positions $\vec{\boldsymbol{z}}_j$ and the complex amplitudes α_j of M point sources, $j = 1, \ldots, M$, from measurements of polychromatic signals on an array of receivers; see Figure 1. The source imaging problem is considered here for simplicity. The active array imaging problem can be cast under the same linear algebra framework even when multiple scattering is important [5].

The array consists of $N_r = 21$ receivers located at $x_r = -\frac{a}{2} + \frac{r-1}{N_r-1}a$, $r = 1, \ldots, N_r$, where $a = 100\lambda$ is the array aperture. The imaging window (IW) is at range $L = 100\lambda$ from the array and the bandwidth $B = f_0/3$ of the emitted pulse is $1/3$ of the central frequency f_0, so the resolution in range is $c/B = 3\lambda$ while in cross-range it is $\lambda L/a = \lambda$. We consider a high frequency microwave imaging regime with central frequency $f_0 = 60$GHz corresponding to $\lambda_0 = 5$mm. We make measurements for $N_f = 21$ equally spaced frequencies spanning a bandwidth $B = 20$GHz. The array aperture is $a = 50$cm, and the distance from the array to the center of the IW is $L = 50$cm. Then, the resolution is $\lambda_0 L/a = 5$mm in the cross-range (direction parallel to the array) and $c_0/B = 15$mm in range (direction of propagation). These parameters are typical in microwave scanning technology [14].

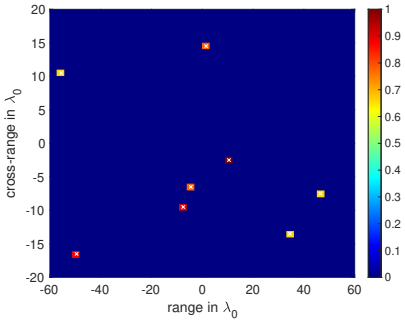

Fig. 2. The true $\boldsymbol{\chi} = \text{diag}(X) = \text{diag}(\boldsymbol{\rho}\,\boldsymbol{\rho}^*)$, i.e., the absolute values squared of the point sources amplitudes.. The dimension of the image is $K = 1681$.

We consider an IW with $K = 1681$ pixels which makes the dimension of $X = \boldsymbol{\rho}\,\boldsymbol{\rho}^*$ equal to $K^2 = 2825761$. The pixel dimensions, i.e., the resolution of the imaging system, is $5\text{mm} \times 15\text{mm}$. The total number of measurements is $N = N_r N_f = 441$. Thus, we can form $N^2 = 194481$ cross-correlations over frequencies and locations.

Let us first note that with these values for N and K, which in fact are not big, we cannot form the full $K^2 \times K^2$ matrix $(\bar{\mathcal{A}} \otimes \mathcal{A})$ so as to solve (15) for the $K^2 \times 1$ vector $\text{vec}(X)$ because of its huge dimensions. Instead, we propose to reduce the dimensionality of the problem to K unknowns corresponding to $\text{diag}(X)$, neglecting all the off-diagonal terms that correspond to the interference terms $\rho_k\rho^*_{k'}$ for $k \neq k'$. Their contributions to the cross-correlated data are treated as *noise*, which is absorbed in a fictitious vector $\boldsymbol{\eta}$ using a *Noise Collector*. We stress that this *noise* is never small if $M \geq 2$, as its contribution to the the cross-correlated data is of order $O(M^2)$, while the contribution of $\text{diag}(X)$ is only of order $O(M)$.

In the following examples, we consider imaging of $M = 8$ point sources; see Fig. 2. Instead of the N^2 cross-correlated data, which are in principle available, we only use $\mathcal{N} = 21N$ cross-correlated data picked at random. This reduces even more the dimensionality of the problem that we are solving.

In Fig. 3, we present the results when the used data is noise-free. The left column shows the results when we use the ℓ_1 algorithm (34); the top plot is the recovered image and the bottom plot the recovered $\boldsymbol{\chi} = \text{diag}(X) = \text{diag}(\boldsymbol{\rho}\,\boldsymbol{\rho}^*)$ vector. The support of the sources is exact but the amplitudes are not. If it is important for an application to recover the amplitudes with precision, one can consider in a second step

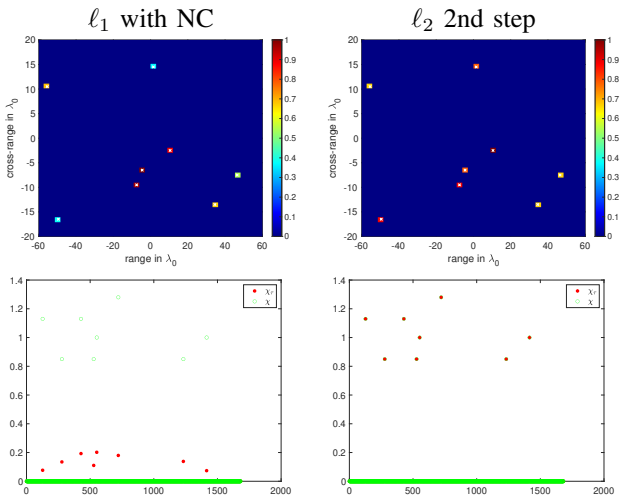

Fig. 3. Imaging $M = 8$ sources using correlations and the NC. The dimension of the image is $K = 1681$. The dimension of the linear data is $N = 441$. The ℓ_1 images are obtained using $21N$ of the N^2 correlation data. Noise free data.

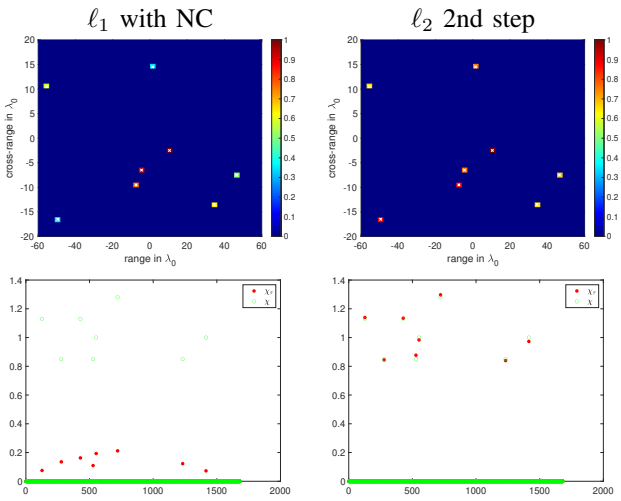

Fig. 4. Imaging $M = 8$ sources using correlations and the NC. The dimension of the image is $K = 1681$. The dimension of the linear data is $N = 441$. The ℓ_1 images are obtained using $21N$ of the N^2 correlation data. Data with 10dB SNR.

the full problem (15) for vec(X) with all the interference terms $\rho_k \rho_{k'}^*$ for $k \neq k'$, but restricted to the exact support found in the first step. If there is no noise in the data, this second step finds the exact values of the amplitudes efficiently using an ℓ_2 minimization method; see the right column of Fig. 3.

In Figs. 4 and 5 we consider the same configuration of sources but we add white Gaussian noise to the data. The resulting SNR values are 10dB and 0dB, respectively. In both cases, the solutions obtained in the first step look very similar to the one obtained in Fig. 3 for noise free data. This is so, because the noise in the data is dominated by the neglected interference terms. The actual effect of the additive noise is only seen in the 2nd step when we solve for vec(X), restricted to the support, using an ℓ_2 minimization method. Indeed, when the data are noisy we cannot recover the exact values of the amplitudes. Still, since an ℓ_2 method is used on the correct support, the reconstructions are extremely robust and give very good results, even when the SNR is 0dB.

To illustrate the robustness of the reconstructions of the entire matrix $X = \boldsymbol{\rho}\,\boldsymbol{\rho}^*$ we also plot in Fig. 6 the angle of X_τ compared to the angle of X restricted on the support recovered during the first step. We get an exact reconstruction for noise-free data. The error in the reconstruction increases as the SNR decreases but the results are very satisfactory even for the 0dB SNR case.

Again, the big advantage of the proposed ℓ_1 minimization approach that seeks only for the components of diag(X), and uses a *Noise Collector* to absorb the interference terms that are treated as noise, is that it is linear in the number of pixels K instead of quadratic. This allows us to consider large scale problems. Moreover, as we observed in the results of Figs. 3 to 6, the number of data $\mathcal{N}$ used to recover the images do not need to be N^2, but only a multiple of N.

In Fig. 7 we illustrate the performance of the proposed ℓ_1 approach for different sparsity levels M and data sizes $\mathcal{N}$. There is no additive noise added to the data in this figure.

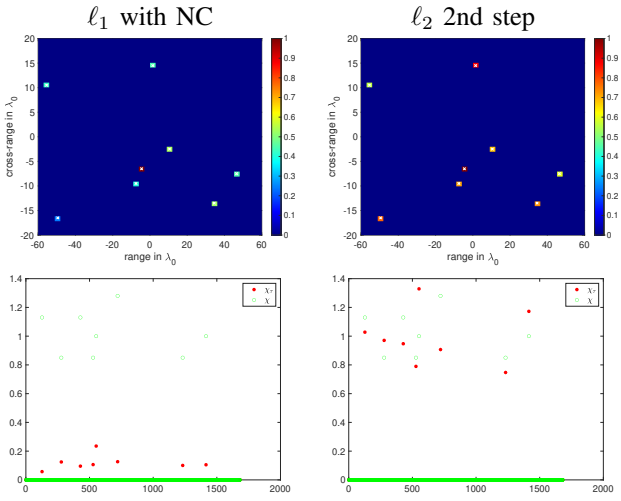

Fig. 5. Imaging $M = 8$ sources using correlations and the NC. The dimension of the image is $K = 1681$. The dimension of the linear data is $N = 441$. The ℓ_1 images are obtained using $21N$ of the N^2 correlation data. Data with 0dB SNR.

Success in recovering the true support of the unknown $\boldsymbol{\chi}$ corresponds to the value 1 (yellow) and failure to 0 (blue). The small phase transition zone (green) contains intermediate values. The red line is the the estimate $\sqrt{\mathcal{N}}/(2\sqrt{\ln \mathcal{N}})$. These results are obtained by averaging over 10 realizations.

VI. Conclussions

In this paper, we consider the problem of sparse signal recovery from cross correlation measurements. The unknown in this case is the correlated matrix signal $X = \boldsymbol{\rho}\boldsymbol{\rho}^*$ whose dimension grows quadratically with the size K of $\boldsymbol{\rho}$ and, hence, inversion becomes computationally unfeasible as K becomes large. To overcome this issue, we propose a novel dimension reduction strategy. Specifically, we vectorize the problem and consider only the diagonal terms $|\rho_i|^2$ of X as unknown. The contribution of the off-diagonal interference terms $\rho_i \rho_j^*$ for

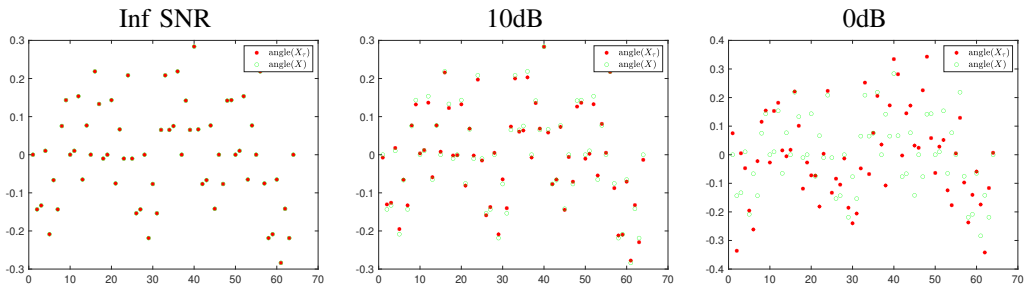

Fig. 6. Imaging $M = 8$ sources using correlations and the NC. The dimension of the image is $K = 1681$. The dimension of the linear data is $N = 441$. The angle of the components of X_τ compared to angle of the components of the true X restricted on the support.

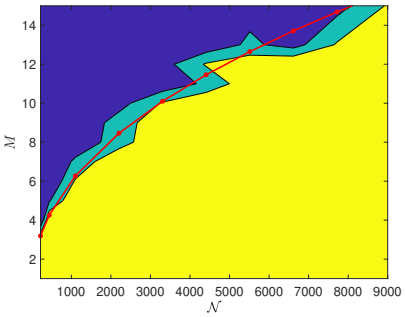

Fig. 7. Algorithm performance for exact support recovery during the first step using ℓ_1 and the Noise Collector. Success corresponds to the value 1 (yellow) and failure to 0 (blue). The small phase transition zone (green) contains intermediate values. The red line is the estimate $\sqrt{\mathcal{N}}/(2\sqrt{\ln \mathcal{N}})$. Ordinate and abscissa are the data used $\mathcal{N}$ and the sparsity M.

$i \neq j$ is treated as noise, which is absorbed using the *Noise Collector* approach introduced in [17]. In this way, we are able to relate the (noisy) data to the diagonal terms $|\rho_i|^2$ of X through a linear transformation. This allows us to recover the signal exactly using efficient ℓ_1-minimization algorithms. The cost of solving this dimension reduced problem is similar to the one using linear data. Furthermore, our numerical experiments show that the suggested approach is robust with respect to additive noise in the data. Finally, we point out that when using cross-correlated data the maximum level of sparsity that can be recovered increases to $O(N/\sqrt{\ln N})$ instead of $O(\sqrt{N}/\sqrt{\ln N})$ for the linear data.

ACKNOWLEDGMENTS

The work of M. Moscoso was partially supported by Spanish MICINN grant FIS2016-77892-R. The work of A.Novikov was partially supported by NSF DMS-1813943 and AFOSR FA9550-20-1-0026. The work of G. Papanicolaou was partially supported by AFOSR FA9550-18-1-0519. The work of C. Tsogka was partially supported by AFOSR FA9550-17-1-0238 and FA9550-18-1-0519.

APPENDIX

A. Proof of Theorem 2

Proof: To prove the first claim, we repeat the proof of Theorem 2 from [17]. Define H_1 as the convex hull of the columns of $\mathcal{C}$, and H_2 as the convex hull of the columns $\boldsymbol{t}_i$ of T in the support of $\boldsymbol{\chi}$, as follows.

$$H_1 = H_1(\tau) = \left\{ x \in \mathbb{R}^N \,\middle|\, x = \tau \sum_{i=1}^{\Sigma} \xi_i \boldsymbol{c}_i,\ \sum_{i=1}^{\Sigma} |\xi_i| \leqslant 1 \right\},$$

$$H_2 = \left\{ x \in \mathbb{R}^N \,\middle|\, x = \sum_{i \in \text{supp}(\boldsymbol{\chi})} \xi_i \boldsymbol{t}_i,\ \sum_{i=1}^{K} |\xi_i| \leqslant 1 \right\},$$

and

$$H(\tau) = \{\xi h_1 + (1-\xi) h_2, 0 \leqslant \xi \leqslant 1, h_i \in H_i\}.$$

Suppose the $(M+1)$-dimensional space V is spanned by $\boldsymbol{e}$ and the column vectors $\boldsymbol{t}_j$, with j in the support of $\boldsymbol{\chi}$. Denote by $\boldsymbol{t}_i^v$ the orthogonal projections of $\boldsymbol{t}_i$ on V. We will prove that $\text{supp}(\boldsymbol{\chi}_\tau) \subset \text{supp}(\boldsymbol{\chi})$ if for any $\boldsymbol{t}_j$, $j \notin \text{supp}(\boldsymbol{\chi})$, we have $\boldsymbol{t}_j^v \subset H(\tau)$ strictly (i.e. $\boldsymbol{t}_j^v \cap \partial H(\tau) = \emptyset$) Fix $j \notin \text{supp}(\boldsymbol{\chi})$, and suppose

$$\boldsymbol{t}_j^v = \xi_0 \boldsymbol{t}_0 + \sum_{k=1}^{M} \xi_k \boldsymbol{t}_{i_k}, \text{ where all } i_k \in \text{supp}(\boldsymbol{\chi}), \boldsymbol{t}_0 = \frac{\boldsymbol{e}}{\|\boldsymbol{e}\|}. \tag{35}$$

Suppose $|\xi_k| = \max_{n \leqslant M} |\xi_n|$. Multiply (35) by $|\xi_k| \boldsymbol{t}_k^v / \xi_k$. Using (22) and (23) we obtain

$$c_0 \frac{\sqrt{\ln \mathcal{N}}}{\sqrt{\mathcal{N}}} \geqslant |\xi_k| \left(1 - M c_0 \frac{\sqrt{\ln \mathcal{N}}}{\sqrt{\mathcal{N}}} \right)$$

Choose α in (24) so that

$$M c_0 \frac{\sqrt{\ln \mathcal{N}}}{\sqrt{\mathcal{N}}} \leqslant \frac{1}{4}. \tag{36}$$

Then,

$$\left(1 - M c_0 \frac{\sqrt{\ln \mathcal{N}}}{\sqrt{\mathcal{N}}} \right) \geqslant \frac{3}{4},$$

and therefore,

$$|\xi_k| \leqslant \frac{4 c_0}{3} \frac{\sqrt{\ln \mathcal{N}}}{\sqrt{\mathcal{N}}}$$

for all $k = 0, 1, 2, \ldots, M$. Hence, $\sum_{k=1}^{M} |\xi_k| \leqslant 1/3$. By the Milman's extension of the Dvoretzky's theorem [15] we can find $\tau_0 = O(1)$ so that

$$4 c_0 \frac{\sqrt{\ln \mathcal{N}}}{\sqrt{\mathcal{N}}} \boldsymbol{t}_0 := \tilde{\boldsymbol{t}}_0 \in H_1(\tau_0)$$

with probability $1-1/\mathcal{N}^{\kappa}$. Therefore,

$$\begin{gathered}\boldsymbol{t}_j^v=\tilde{\xi}_0\tilde{\boldsymbol{t}}_0+\sum_{k=1}^M\xi_k\boldsymbol{t}_{i_k},\ \text{where all } i_k\in\operatorname{supp}(\boldsymbol{\chi})\\ \text{and } |\tilde{\xi}_0|+\sum_{k=1}^M|\xi_k|\leqslant 1/3+1/3\leqslant 2/3\end{gathered}$$

and $\tilde{\boldsymbol{t}}_0\in H_1(\tau)$ for all $\tau\geqslant\tau_0$. Therefore, $\boldsymbol{t}_j^v\subset H(\tau)$ strictly.

To prove the second claim, we repeat the proof of Theorem 3 from [17]. Estimate (36) implies we can assume $\langle\boldsymbol{t}_i,\boldsymbol{t}_j\rangle=0$ for $i\neq j$, $i,j\in\operatorname{supp}(\boldsymbol{\chi})$ - this will only replace the constant c_1 in (26) to $2c_1/\sqrt{3}$ at most. Suppose V^i are the 2-dimensional spaces spanned by $\boldsymbol{e}$ and $\boldsymbol{t}_i$ for $i\in\operatorname{supp}(\boldsymbol{\chi})$. By the Milman's extension of the Dvoretzky's theorem [15] all $\lambda H(\tau)\cap V^i$ look like rounded rhombi depicted on Fig. 8, and $\lambda H_1(\tau)\cap V^i\subset B^i_{\lambda\tau}$ with probability $1-N^{-\kappa}$, where $B^i_{\lambda\tau}$ is a 2-dimensional ℓ_2-ball of radius $\lambda\tau c_0\sqrt{\ln\mathcal{N}}/\sqrt{\mathcal{N}}$. Thus $\lambda H(\tau)\cap V^i\subset H^i_{\lambda\tau}$ with probability $1-N^{-\kappa}$, where $H^i_{\lambda\tau}$ is the convex hull of $B^i_{\lambda\tau}$ and a vector $\lambda\boldsymbol{f}_i$, $\boldsymbol{f}_i=\chi_i\|\boldsymbol{\chi}\|_{\ell_1}^{-1}\boldsymbol{t}_i$. Then $\operatorname{supp}(\boldsymbol{\chi}_\tau)=\operatorname{supp}(\boldsymbol{\chi})$, if there exists λ_0 so that $\chi_i\boldsymbol{t}_i+\boldsymbol{e}$ lies on the flat boundary of $H^i_{\lambda_0}$ for all $i\in\operatorname{supp}(\boldsymbol{\chi})$.

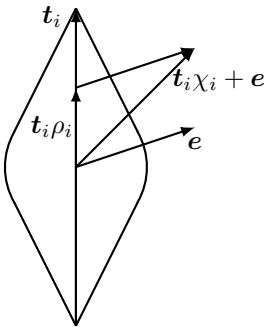

Fig. 8. An intersection of $H(\tau)$ with the $\operatorname{span}(\boldsymbol{t}_i,\boldsymbol{e})$ is a rounded rhombus.

If $\min_{i\in\operatorname{supp}(\boldsymbol{\chi})}|\chi_i|\geqslant\gamma\|\boldsymbol{\chi}\|_\infty$, then there exists a constant $c_2=c_2(\gamma)$ such that if $\chi_i\boldsymbol{t}_i+\boldsymbol{e}$ lies on the flat boundary of H^i_λ for some i and some λ, then there exists λ_0 so that $\chi_i\boldsymbol{t}_i+c_2\boldsymbol{e}$ lies on the flat boundary of $H^i_{\lambda_0}$ for all $i\in\operatorname{supp}(\boldsymbol{\chi})$. If

$$\frac{|\langle\boldsymbol{t}_i,\chi_i\boldsymbol{t}_i+\boldsymbol{e}\rangle|}{\|\chi_i\boldsymbol{t}_i+\boldsymbol{e}\|_{\ell_2}}\geqslant\frac{\tau c_0\sqrt{\ln\mathcal{N}}}{\sqrt{\mathcal{N}}\|\boldsymbol{f}_i\|_{\ell_2}}=\frac{\tau c_0\|\boldsymbol{\chi}\|_{\ell_1}\sqrt{\ln\mathcal{N}}}{\sqrt{\mathcal{N}}|\chi_i|},\quad(37)$$

then $\chi_i\boldsymbol{t}_i+c_2\boldsymbol{e}$ lies on the flat boundary of H^i_λ.

Since $|\langle\boldsymbol{t}_i,\chi_i\boldsymbol{t}_i+\boldsymbol{e}\rangle|\geqslant|\chi_i|/2$ by (25), inequality (37) holds if

$$\frac{|\chi_i|}{\|\chi_i\boldsymbol{t}_i+\boldsymbol{e}\|_{\ell_2}}\geqslant\frac{2\tau c_0\|\boldsymbol{\chi}\|_{\ell_1}\sqrt{\ln\mathcal{N}}}{\sqrt{\mathcal{N}}|\chi_i|}.$$

By (26) and using $\|\boldsymbol{\chi}\|_{\ell_1}\leqslant M$ the last inequality is true if

$$M\leqslant\frac{\sqrt{\mathcal{N}}}{\tau c_1c_0\sqrt{\ln\mathcal{N}}}.$$

The last inequality is true if α in (24) is small enough. Thus, $\operatorname{supp}(\boldsymbol{\chi}_\tau)=\operatorname{supp}(\boldsymbol{\chi})$. ∎

B. Proof of Theorem 3

Proof: We need to verify that all conditions of Theorem 2 are satisfied. Choose c_0, τ_0, $\mathcal{N}_0$ and α so that Theorem 2 is satisfied with probability $1-\frac{1}{3\mathcal{N}^{\kappa}}$. Note that we can increase c_0, τ_0, $\mathcal{N}_0$ and decrease α in this proof if necessary. We denote by $(\bar{\mathcal{A}}\otimes\mathcal{A})_{k,l}$ the column of $\bar{\mathcal{A}}\otimes\mathcal{A}$ that arises from a tensor product $\bar{\boldsymbol{a}}_k\otimes\boldsymbol{a}_l$. If we use all N^2 of the data the columns $\bar{\mathcal{A}}\otimes\mathcal{A}$,then

$$\begin{aligned}\langle(\bar{\mathcal{A}}\otimes\mathcal{A})_{k,l},(\bar{\mathcal{A}}\otimes\mathcal{A})_{m,n}\rangle&=\sum_{i=1}^N\sum_{j=1}^N\bar{a}_{k,i}a_{l,j}a_{m,i}\bar{a}_{n,j}\\&=\langle\boldsymbol{a}_m,\boldsymbol{a}_k\rangle\langle\boldsymbol{a}_l,\boldsymbol{a}_n\rangle.\end{aligned}$$

In particular, all columns of $\bar{\mathcal{A}}\otimes\mathcal{A}$ have length 1. Therefore,

$$|\langle\boldsymbol{t}_i,\boldsymbol{t}_j\rangle|=\left|(\bar{\mathcal{A}}\otimes\mathcal{A})_{i,i},(\bar{\mathcal{A}}\otimes\mathcal{A})_{j,j}\right|=|\langle\boldsymbol{a}_i,\boldsymbol{a}_j\rangle|^2\leqslant\frac{\Delta^2}{N},$$

and condition (23) is verified if we choose $\mathcal{N}_0$ large enough.

Now we obtain

$$\lambda_1M\leqslant\|\boldsymbol{e}\|_{\ell_2}\leqslant\lambda_2M\quad(38)$$

with high probability. Note that (38) implies (26) because $\gamma M\leqslant\|\boldsymbol{\chi}\|_{\ell_1}\leqslant M$. We write

$$\|\boldsymbol{e}\|_{\ell_2}^2=\|\boldsymbol{\chi}\|_{\ell_1}^2+2\|\boldsymbol{\chi}\|_{\ell_1}\Xi_1+\Xi_2,$$

where

$$\Xi_1=\sum_{k,l,k\neq l}\bar{\rho}_k\rho_l\langle\boldsymbol{a}_k,\boldsymbol{a}_l\rangle,\quad(39)$$

and

$$\Xi_2=\sum_{\text{all indices different}}\rho_k\bar{\rho}_l\bar{\rho}_m\rho_n\langle\boldsymbol{a}_m,\boldsymbol{a}_k\rangle\langle\boldsymbol{a}_l,\boldsymbol{a}_n\rangle.\quad(40)$$

By Hanson-Wright inequality (48)

$$\mathbb{P}\left(|\Xi_1|>t\right)\leqslant2\exp\left(-\frac{t^2/32}{\|\mathbf{M}\|_F^2}\right)$$

where $\mathbf{M}$ is a matrix with components $|\rho_k\rho_l|\langle\boldsymbol{a}_k,\boldsymbol{a}_l\rangle$, $\|\mathbf{M}\|_F$ is its Frobenius (Hilbert-Schmidt) norm. Since $|\langle\boldsymbol{a}_k,\boldsymbol{a}_l\rangle|\leqslant\Delta/\sqrt{N}$, we obtain $\|\mathbf{M}\|_F\leqslant\Delta M/\sqrt{N}$ (in our set-up $\|\boldsymbol{\rho}\|_{\ell_\infty}=1$). Take $t=\gamma M/8\leqslant\|\boldsymbol{\chi}\|_{\ell_1}/8$ and obtain

$$\mathbb{P}\left(|\Xi_1|>\gamma M/8\right)\leqslant2\exp\left(-cN\right),c=c(\gamma),$$

which is negligible for large N. Thus

$$|\Xi_1|\leqslant\frac{\|\boldsymbol{\chi}\|_{\ell_1}}{8}\quad(41)$$

with probability $1-2\exp\left(-cN\right)$. Observe that $\Xi_2=(\Xi_1)^2-\Xi_3$, where

$$\Xi_3=\sum_{m=l\text{ or }k=n\text{ or both}}\rho_k\bar{\rho}_l\bar{\rho}_m\rho_n\langle\boldsymbol{a}_m,\boldsymbol{a}_k\rangle\langle\boldsymbol{a}_l,\boldsymbol{a}_n\rangle.$$

For Ξ_3 we can use a deterministic estimate:

$$|\Xi_3|\leqslant2\frac{c_0^2M^3}{N}\leqslant c\alpha\frac{\|\boldsymbol{\chi}\|_{\ell_1}^2}{\sqrt{\ln N}}\leqslant\frac{\|\boldsymbol{\chi}\|_{\ell_1}^2}{16}.$$

For $(\Xi_1)^2$ we use (41). Using the union bound, we obtain

$$\frac{1}{2}\|\boldsymbol{\chi}\|_{\ell_1}^2\leqslant\|\boldsymbol{e}\|_{\ell_2}^2\leqslant\frac{3}{2}\|\boldsymbol{\chi}\|_{\ell_1}^2\quad(42)$$

with probability $1-2\exp\left(-cN\right)$. Thus, (38) holds with probability $1-2\exp\left(-cN\right)$.

We will now prove (22). For $m \notin \text{supp}(\boldsymbol{\chi})$, consider a random variable

$$\begin{aligned}\Theta_m &= \langle \boldsymbol{t}_m, \boldsymbol{e}\rangle \\ &= \sum_{k,l,k\neq l} \bar{\rho}_k \rho_l \langle \boldsymbol{t}_m, (\bar{\mathcal{A}}\otimes\mathcal{A})_{k,l}\rangle \\ &= \sum_{k,l,k\neq l} \bar{\rho}_k \rho_l \langle \boldsymbol{a}_m, \boldsymbol{a}_k\rangle\langle \boldsymbol{a}_m, \boldsymbol{a}_l\rangle.\end{aligned} \tag{43}$$

We have

$$\left|\langle \boldsymbol{t}_m, (\bar{\mathcal{A}}\otimes\mathcal{A})_{k,l}\rangle\right| = |\langle \boldsymbol{a}_m, \boldsymbol{a}_k\rangle\langle \boldsymbol{a}_m, \boldsymbol{a}_l\rangle| \leqslant \frac{\Delta^2}{N} \tag{44}$$

if $m \neq k$, and $m \neq l$. If $\mathbf{M}$ is a matrix with components $|\rho_k\rho_l|\langle \boldsymbol{a}_m, \boldsymbol{a}_k\rangle\langle \boldsymbol{a}_m, \boldsymbol{a}_l\rangle$, then $\|\mathbf{M}\|_F \leqslant \Delta^2 M/N$. Using (38) choose $t = c_0\frac{\sqrt{\ln N^2}}{N}\|\boldsymbol{e}\|_{\ell_2} > c_0\frac{\gamma}{2}\frac{M\sqrt{\ln N}}{N}$ in Hanson-Wright inequality (48) to obtain:

$$\begin{aligned}\mathbb{P}\left(|\Theta_m| > c_0\tfrac{\sqrt{\ln N^2}}{N}\|\boldsymbol{e}\|_{\ell_2}\right) &\leqslant \mathbb{P}\left(|\Theta_m| > c_0\frac{\gamma}{2}\frac{M\sqrt{\ln N}}{N}\right) \\ &\leqslant 2\exp\left(-\frac{\gamma^2 c_0^2 \ln N}{128\Delta^4}\right).\end{aligned}$$

Then (22) holds with probability $1-\frac{1}{3\mathcal{N}^\kappa}$ if c_0 is large enough.

We will now prove (25). For $m \in \text{supp}(\boldsymbol{\chi})$ decompose

$$\Theta_m = \langle \boldsymbol{t}_m, \boldsymbol{e}\rangle = \Theta_m^1 + \Theta_m^2$$

where

$$\Theta_m^1 = \sum_{k,l,k\neq l,k\neq m,l\neq m} \bar{\rho}_k\rho_l\langle \boldsymbol{a}_m, \boldsymbol{a}_k\rangle\langle \boldsymbol{a}_m, \boldsymbol{a}_l\rangle$$

and

$$\Theta_m^2 = \sum_{k,k\neq m} \left(\bar{\rho}_m\rho_k + \bar{\rho}_k\rho_m\right)\langle \boldsymbol{a}_m, \boldsymbol{a}_k\rangle \tag{45}$$

The distribution of the random variable Θ_m^1 has exactly the same behavior as Θ_m for $m \notin \text{supp}(\boldsymbol{\chi})$. We therefore have

$$\begin{aligned}\mathbb{P}\left(|\Theta_m^1| > \tfrac{\gamma}{4}\|\chi\|_{\ell_\infty}\right) &\leqslant 2\exp\left(-c\frac{N^2}{\Delta^4 M^2}\right) \\ &\leqslant 2\exp(-\tilde{c}\ln N/\alpha^2) \\ &\leqslant \frac{1}{6}\frac{1}{\mathcal{N}^\kappa},\end{aligned}$$

by Hanson-Wright inequality 48 if α is small enough. . If $m = l$ (or $m = k$) then

$$\left|\langle \boldsymbol{t}_m, (\bar{\mathcal{A}}\otimes\mathcal{A})_{k,m}\rangle\right| = |\langle \boldsymbol{a}_m, \boldsymbol{a}_k\rangle| \leqslant \frac{\Delta}{\sqrt{N}}. \tag{46}$$

If we condition on $\boldsymbol{\rho}_m$, then Θ_m^2 is a sum of independent random variables. Therefore by Hoeffding's inequality

$$\mathbb{P}\left(|\Theta_m^2| > t\right) \leqslant 2\exp\left(-c\frac{t^2}{b^2}\right), \text{ where } b^2 \leqslant \frac{c_0^2 M}{N} \leqslant \frac{\Delta^2\alpha}{\ln N}.$$

Choosing t appropriately we obtain

$$\mathbb{P}\left(|\Theta_m^2| > \frac{\gamma}{4}\|\chi\|_{\ell_\infty}\right) \leqslant \frac{1}{6}\frac{1}{\mathcal{N}^\kappa}.$$

by choosing α small enough. Using the union bound we conclude that

$$\mathbb{P}\left(|\Theta_m| > \frac{\gamma}{2}\|\chi\|_{\ell_\infty}\right) \leqslant \frac{1}{3}\frac{1}{\mathcal{N}^\kappa}$$

for $m \in \text{supp}(\boldsymbol{\chi})$. Applying the union bound we conclude that conditions (25), (22) and estimates in Theorem 2 hold with probability $1-\frac{1}{\mathcal{N}^\kappa}$. This completes the proof. ■

Remark 1: The proof of Theorem 3 reveals why we had to assume (25) for $m \in \text{supp}(\boldsymbol{\chi})$. When $m \notin \text{supp}(\boldsymbol{\chi})$ then $\langle \boldsymbol{t}_m, \boldsymbol{e}\rangle$ is estimated in (43) using (44). When $m \in \text{supp}(\boldsymbol{\chi})$ then $\langle \boldsymbol{t}_m, \boldsymbol{e}\rangle$ contains Θ_m^2 given by (45). For Θ_m^2 we cannot use (44), and we have to use a weaker estimate (46).

C. A deterministic version of Theorem 3

Theorem 4: Suppose X is a solution of (15), $\boldsymbol{\chi} = \text{diag}(X)$ is M-sparse, $\boldsymbol{d} \in \mathbb{C}^{\mathcal{N}}$, $\mathcal{N} = N^2$, and $T = (\bar{\mathcal{A}}\otimes\mathcal{A})_{\boldsymbol{\chi}} : \mathbb{C}^{\mathcal{K}} \to \mathbb{C}^{\mathcal{N}}$. Fix $\beta > 1$, and draw $\Sigma = \mathcal{N}^\beta$ columns for $\mathcal{C}$, independently, from the uniform distribution on $\mathbb{S}^{\mathcal{N}-1}$ and define γ as in (18) and Δ as in (30). Then, for any $\kappa > 0$, there are constants $\alpha = \alpha(\kappa,\gamma,\Delta)$, $\tau = \tau(\kappa,\beta)$, and $\mathcal{N}_0 = \mathcal{N}_0(\kappa,\beta,\gamma,\Delta)$ such that the following holds. If $M \leqslant \alpha\sqrt{N}$ and $\boldsymbol{\chi}_\tau$ is the solution (17), then $\text{supp}(\boldsymbol{\chi}) = \text{supp}(\boldsymbol{\chi}_\tau)$ for all $\mathcal{N} > \mathcal{N}_0$ with probability $1 - 1/\mathcal{N}^\kappa$.

Proof: We need to verify that all conditions of Theorem 2 are satisfied non-probabilistically. Conditions (23) is already verified in the proof of Theorem 3 under even weaker assumptions than in Theorem 4. Therefore we only need to verify estimates (26), (25) and (22).

Since

$$\boldsymbol{e} = \sum_{k\neq l}\bar{\rho}_k\rho_l(\bar{\mathcal{A}}\otimes\mathcal{A})_{k,l},$$

we have

$$\begin{aligned}\|\boldsymbol{e}\|_{\ell_2}^2 &\leqslant \frac{2\Delta}{\sqrt{N}}\sum_{all\ indices}\chi_k\,|\rho_{m_1}|\,|\rho_{m_2}| \\ &+\frac{\Delta^2}{N}\sum_{all\ indices}|\rho_{m_1}|\,|\rho_{m_2}|\,|\rho_{k_1}|\,|\rho_{k_2}| \\ &+\sum_{k,m}\chi_k\chi_m\end{aligned}$$

and thus

$$\begin{aligned}\|\boldsymbol{e}\|_{\ell_2}^2 &\leqslant 2\Delta\alpha\|\boldsymbol{\chi}\|_{\ell_1}\|\boldsymbol{\rho}\|_{\ell_2}^2 + \Delta^2\alpha^2\|\boldsymbol{\rho}\|_{\ell_2}^4 + \|\boldsymbol{\chi}\|_{\ell_1}^2 \\ &= (1+\Delta\alpha)^2\|\boldsymbol{\chi}\|_{\ell_1}^2.\end{aligned}$$

Thefore estimate (26) holds.

A non-probabilistic version of estimate (25) is as follows. For $m \in \text{supp}(\boldsymbol{\chi})$ we have

$$\begin{aligned}|\langle \boldsymbol{t}_m, \boldsymbol{e}\rangle| &= \left|\sum_{k,l,k\neq l}\bar{\rho}_k\rho_l\langle \boldsymbol{t}_m, (\bar{\mathcal{A}}\otimes\mathcal{A})_{k,l}\rangle\right| \\ &\leqslant 2\sum_{k,k\neq m}|\rho_k|\,|\rho_m|\left|\langle \boldsymbol{t}_m, (\bar{\mathcal{A}}\otimes\mathcal{A})_{k,m}\rangle\right| \\ &+\sum_{k,l,k\neq l\neq m}|\rho_k|\,|\rho_l|\left|\langle \boldsymbol{t}_m, (\bar{\mathcal{A}}\otimes\mathcal{A})_{k,l}\rangle\right| \\ &\leqslant \frac{2\Delta}{\sqrt{N}}\sum_k|\rho_k|\,|\rho_m| + \frac{\Delta^2}{N}\sum_{k,l}|\rho_k|\,|\rho_l| \\ &\leqslant \left(\frac{2\Delta M}{\sqrt{N}} + \frac{\Delta^2 M^2}{N}\right)\|\boldsymbol{\rho}\|_{\ell_\infty}^2 \\ &= \left(\frac{2\Delta M}{\sqrt{N}} + \frac{\Delta^2 M^2}{N}\right)\|\boldsymbol{\chi}\|_{\ell_\infty} \\ &\leqslant \frac{\gamma}{2}\|\boldsymbol{\chi}\|_{\ell_\infty}.\end{aligned}$$

if α is small enough.

We now obtain a lower bound on $\|\boldsymbol{e}\|_{\ell_2}$. For Ξ_1 and Ξ_2 in (39) and (40), respectively, we have

$$|\Xi_1| \leqslant \frac{\Delta M^2}{\sqrt{N}} \leqslant \Delta\alpha M, \quad |\Xi_2| \leqslant \frac{\Delta^2 M^4}{N} \leqslant \Delta^2\alpha^2 M^2.$$

Since

$$\|\boldsymbol{e}\|_{\ell_2}^2 = \|\boldsymbol{\chi}\|_{\ell_1}^2 + 2\|\boldsymbol{\chi}\|_{\ell_1}\Xi_1 + \Xi_2, \text{ and } \|\boldsymbol{\chi}\|_{\ell_1} = M$$

we can choose α so that

$$M/2 = \|\boldsymbol{\chi}\|_{\ell_1}/2 \leqslant \|\boldsymbol{e}\|_{\ell_2}^2.$$

To show (22) observe that

$$\left|\langle \boldsymbol{t}_m, (\bar{\mathcal{A}} \otimes \mathcal{A})_{k,l}\rangle\right| = |\langle \boldsymbol{a}_m, \boldsymbol{a}_k\rangle\langle \boldsymbol{a}_m, \boldsymbol{a}_l\rangle| \leqslant \Delta^2/N,$$

because $m \neq k$, and $m \neq l$. Therefore

$$\begin{aligned} |\langle \boldsymbol{t}_m, \boldsymbol{e}\rangle| &= \left|\sum_{k,l,k\neq l} \bar{\rho}_k\rho_l\langle \boldsymbol{t}_m, (\bar{\mathcal{A}} \otimes \mathcal{A})_{k,l}\rangle\right| \\ &\leqslant \frac{\Delta^2 M^2}{N} \leqslant \frac{\Delta^2\alpha\|\boldsymbol{e}\|_{\ell_2}}{\sqrt{N}}, \end{aligned}$$

and (22) follows either for choosing α small or $\ln N$ large. ■

D. Hansen-Wright's Inequality for bounded symmetric random variables

For simplicity of presentation all random variables here are real. Suppose X_i are independent sub-gaussian random variables, $\mathbb{E}(X_i) = 0$, and the sub-gaussian norms $\|X_i\|_{\psi_2} \leqslant K$. Consider

$$\Xi = \sum_{i,j} X_j X_i m_{ij},$$

where m_{ij} are entries of a deterministic $M \times M$ diagonal-free (i.e. $m_{ii} = 0$) matrix $\mathbf{M}$. The Hanson-Wright inequality (see e.g. [18]) is

$$\mathbb{P}\left(|\Xi| > t\right) \leqslant 2\exp\left(-c\min\left(\frac{t^2}{K^4\|\mathbf{M}\|_F^2}, \frac{t}{K^2\|\mathbf{M}\|}\right)\right) \tag{47}$$

where $\|\mathbf{M}\|_F$ is the Frobenius (Hilbert-Schmidt) norm of $\mathbf{M}$, and $\|\mathbf{M}\|$ is its operator norm. If we use this inequality in the proof of our Theorem 3, then the result becomes weaker than Theorem 1 by a factor of $\sqrt{\ln N}$ because in our setting

$$\min\left(\frac{t^2}{K^4\|\mathbf{M}\|_F^2}, \frac{t}{K^2\|\mathbf{M}\|}\right) = \frac{t}{K^2\|\mathbf{M}\|}.$$

In order to obtain Theorem 3 in its present form, we need a slight strengthening of (47). Our proof is a modification of two proofs from [20] and [18]. It may already exist in the literature, but we were not able to find it. Therefore we provide it here for the reader's convenience. We assume that our random variables are symmetric and bounded. This holds if a random variable is uniformly distributed on the (complex) unit circle as in Theorem 3.

Theorem 5: (Hansen-Wright inequality for bounded symmetric random variables) Suppose X_i are independent symmetric random variables, with $\|X_i\|_{\ell_\infty} \leqslant K$. Let $\Xi = \sum_{i\neq j} X_j X_i m_{ij}$. Then

$$\mathbb{P}\left(|\Xi| > t\right) \leqslant 2\exp\left(-\frac{t^2/32}{K^4\|\mathbf{M}\|_F^2}\right). \tag{48}$$

Proof: By replacing X_i with X_i/K we can assume $K = 1$. By Chebyshev's inequality

$$\mathbb{P}\left(\Xi > t\right) = \mathbb{P}\left(e^{\lambda\Xi} > e^{\lambda t}\right) \leqslant e^{-\lambda t}\mathbb{E}\left(e^{\lambda\Xi}\right) \tag{49}$$

for any $\lambda > 0$. We now use decoupling. Consider independent Bernoulli random variables $\mu_i = 0$ or 1 with probability $1/2$. Since $\mathbb{E}(\mu_i)(1-\mu_j) = 1/4$ for $i \neq j$ we conclude $\Xi = 4\mathbb{E}_\mu\Xi_\mu$, where

$$\Xi_\mu = \sum_{i\neq j} \mu_j(1-\mu_i) X_j X_i m_{ij},$$

and $\mathbb{E}_\mu$ is conditional expectation with respect to $\mu = (\mu_1, \ldots, \mu_M)$. Using independence of $X = (X_1, \ldots, X_M)$ and μ, and applying Jensen's inequality we obtain

$$\mathbb{E}\left(e^{\lambda\Xi}\right) = \mathbb{E}_X\left(e^{\lambda\Xi}\right) \leqslant \mathbb{E}_X\mathbb{E}_\mu\left(e^{4\lambda\Xi_\mu}\right) = \mathbb{E}_\mu\mathbb{E}_X\left(e^{4\lambda\Xi_\mu}\right)$$

where $\mathbb{E}_X$ is conditional expectation with respect to X. This implies there exist a realization of μ such that $\mathbb{E}_X\left(e^{\lambda\Xi}\right) \leqslant \mathbb{E}_X\left(e^{4\lambda\Xi_\mu}\right)$ for this μ. Fix this μ and the corresponding set of indices $\Lambda_\mu = \{i|\delta_i = 1\}$. Then we can write $\Xi_\mu = \sum_{i\neq j, i\in\Lambda_\mu, j\in\Lambda_\mu^c} X_j X_i m_{ij}$. Since the random variables X_i, $i \in \Lambda_\mu$ and X_i, $i \in \Lambda_\mu^c$ are independent, their distribution will not change if we replace X_i, $i \in \Lambda_\mu^c$ by $X_i^{'}$, $i \in \Lambda_\mu^c$, where $X_i^{'}$ is an independent copy of X_i. In other words we have

$$\begin{gathered} \mathbb{E}_{X,X'}\left(e^{\lambda\Xi}\right) \leqslant \mathbb{E}_X\left(e^{4\lambda\tilde{\Xi}_\mu}\right), \\ \text{where } \tilde{\Xi}_\mu = \textstyle\sum_{i\neq j, i\in\Lambda_\mu, j\in\Lambda_\mu^c} X_j^{'} X_i m_{ij}. \end{gathered}$$

We now claim that

$$\mathbb{E}_{X,X'}\left(e^{4\lambda\tilde{\Xi}_\mu}\right) \leqslant \mathbb{E}_{X,X'}\left(e^{4\lambda\tilde{\Xi}}\right), \text{ where } \tilde{\Xi} = \sum_{i\neq j} X_j^{'} X_i m_{ij}.$$

Indeed, Lemma 6.1.2 in [20] states that $\mathbb{E}\left(F(Y)\right) \leqslant \mathbb{E}\left(F(Y+Z)\right)$ for any convex function F, if Y and Z are independent and $\mathbb{E}(Z) = 0$. In our case we take $F(x) = e^{4\lambda x}$, $Y = \tilde{\Xi}_\mu$ and $Z = \tilde{\Xi} - \tilde{\Xi}_\mu$. If we condition on X_i, $i \in \Lambda_\mu$ and X_i, $i \in \Lambda_\mu^c$, then Y is fixed, Z is independent Y and its conditional expectation is zero. Hence the following decoupling estimate is obtained.

$$\mathbb{E}\left(e^{\lambda\Xi}\right) \leqslant \mathbb{E}\left(e^{4\lambda\tilde{\Xi}}\right), \text{ where } \tilde{\Xi} = \sum_{i\neq j} X_j^{'} X_i m_{ij}. \tag{50}$$

By independence

$$\mathbb{E}\left(e^{4\lambda\tilde{\Xi}}\right) = \prod_{i\neq j}\mathbb{E}\left(e^{4\lambda m_{ij}X_i X_j^{'}}\right).$$

Since random variables are symmetric

$$\begin{aligned} \mathbb{E}\left(e^{4\lambda m_{ij}X_i X_j^{'}}\right) &= \frac{1}{2}\mathbb{E}\left(e^{4\lambda m_{ij}\left|X_i X_j^{'}\right|} + e^{-4\lambda m_{ij}\left|X_i X_j^{'}\right|}\right) \\ &\leqslant \mathbb{E}\left(e^{8\lambda^2 m_{ij}^2\left|X_i X_j^{'}\right|^2}\right) \leqslant e^{8\lambda^2 m_{ij}^2}. \end{aligned}$$

Using the last two estimates in (50) we obtain

$$\mathbb{E}\left(e^{\lambda\Xi}\right) \leqslant e^{8\lambda^2\|\mathbf{M}\|_F^2}.$$

Plugging the lsat inequality in 49 and optimizing over λ we obtain

$$\mathbb{P}\left(\Xi > t\right) \leqslant \inf_{\lambda>0} e^{-\lambda t + 8\lambda^2\|\mathbf{M}\|_F^2} < e^{-\frac{t^2}{32\|\mathbf{M}\|_F^2}}.$$

■

REFERENCES

[1] BECK, AMIR, and TEBOULLE, MARC, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM J. Img. Sci. 2 (2009), pp.183–202.

[2] A. BAKULIN AND R. CALVERT, *The virtual source method: Theory and case study,* Geophysics, 71 (2006), pp. SI139–SI150.

[3] E. J. CANDÈS, Y. C. ELDAR, T. STROHMER, AND V. VORONINSKI, *Phase Retrieval via Matrix Completion,* SIAM J. Imaging Sci. 6 (2013), pp. 199–225.

[4] A. CHAI, M. MOSCOSO AND G. PAPANICOLAOU, *Array imaging using intensity-only measurements*, Inverse Problems 27 (2011), 015005.

[5] A. CHAI, M. MOSCOSO AND G. PAPANICOLAOU, *Imaging strong localized scatterers with sparsity promoting optimization*, SIAM J. Imaging Sci. 10 (2014), pp. 1358–1387.

[6] E. DASKALAKIS, C. EVANGELIDIS, J. GARNIER, N. MELIS, G. PAPANICOLAOU AND C. TSOGKA, *Robust seismic velocity change estimation using ambient noise recordings*, Geophys. J. Int. (2016).

[7] J. GARNIER AND G. PAPANICOLAOU, *Passive Sensor Imaging Using Cross Correlations of Noisy Signals in a Scattering Medium*, SIAM J. Imaging Sci. 2 (2009), pp. 396-437.

[8] J. GARNIER AND G. PAPANICOLAOU, *Role of scattering in virtual source imaging*, SIAM Journal of Imaging Science 7 (2014), pp. 1210–1236.

[9] J. GARNIER AND G. PAPANICOLAOU, *Passive imaging with ambient noise*, Cambridge University Press, 2016.

[10] J. GARNIER, G. PAPANICOLAOU, A. SEMIN, C. TSOGKA, Signal to Noise Ratio Analysis in Virtual Source Array Imaging, SIAM Journal of Imaging Science 8 (2015), pp. 248–279.

[11] R. M. GRAY, *Toeplitz and Circulant Matrices: A Review, Foundations and Trends in Communications and Information Theory* **2** (2006), pp. 155–239.

[12] P. R. GRIFFITHS AND J. A. DE HASETH, *Fourier Transform Infrared Spectrometry*, John Wiley & Sons Inc., Hoboken, 2007.

[13] T. HELIN, M. LASSAS, L. OKSANEN, AND T. SAKSAL, *Correlation based passive imaging with a white noise source,* Journal de Mathématiques Pures et Appliquées 116 (2018), pp. 132–160.

[14] J. LAVIADA, A. ARBOLEYA-ARBOLEYA, Y. ALVAREZ-LOPEZ, C. GARCIA-GONZALEZ AND F. LAS-HERAS, *Phaseless synthetic aperture radar with efficient sampling for broadband near-field imaging: Theory and validation,* IEEE Trans. Antennas Propag. 63 (2015), pp. 573–584.

[15] V.D MILMAN, *A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies,* Funkcional. Anal. i Priložen. 5 (1971), pp. 28–37.

[16] M. MOSCOSO, A. NOVIKOV, G. PAPANICOLAOU AND L. RYZHIK, *A differential equations approach to l1-minimization with applications to array imaging*, Inverse Problems 28 (2012), 105001.

[17] M. MOSCOSO, A. NOVIKOV, G. PAPANICOLAOU, C. TSOGKA, *The noise collector for sparse recovery in high dimensions*, Proceedings of the National Academy of Science 117 (2020), pp. 11226–11232, doi: 10.1073/pnas.1913995117.

[18] M. RUDELSON, R. VERSHYNIN, *Hanson-Wright inequality and sub-gaussian concentration*. Electron. Commun. Probab. 18 (2013), paper no. 82, 9 pp. doi:10.1214/ECP.v18-2865.

[19] G. T. SCHUSTER, *Seismic Interferometry*, Cambridge University Press, Cambridge, 2009.

[20] R. VERSHYNIN, *High-dimensional probability*. An introduction with applications in data science, Cambridge University Press, 2018.

[21] K. WAPENAAR, E. SLOB, R. SNIEDER, AND A. CURTIS, *Tutorial on seismic interferometry: Part 2 - Underlying theory and new advances*, Geophysics, 75 (2010), pp. 75A211–75A227.