Blind Sparse Recovery From Superimposed Non-Linear Sensor Measurements

Martin Genzel
Department of Mathematics
Technische Universität Berlin
E-mail: genzel@math.tu-berlin.de

Peter Jung
Communications and Information Theory Group
Technische Universität Berlin
E-mail: peter.jung@tu-berlin.de

***Abstract*—In this work, we study the problem of sparse recovery from superimposed, non-linearly distorted measurements. This challenge is particularly relevant to wireless sensor networks that consist of autonomous and spatially distributed sensor units. Here, each of the M wireless sensors acquires m individual measurements of an s-sparse source vector $\boldsymbol{x}_0 \in \mathbb{R}^n$. All devices transmit simultaneously to a central receiver causing collisions. Since this process is imperfect, e.g., caused by low-quality sensors and the wireless channel, the receiver measures a superposition of corrupted signals. First, we will show that the source vector can be successfully recovered from $m = \mathcal{O}(s \log(2n/s))$ coherently communicated measurements via the vanilla Lasso. The more general situation of non-coherent communication can be modeled as a bilinear compressed sensing problem. In this setting, it will turn out that $m = \mathcal{O}(s \cdot \max\{\log(2n/s), M\})$ measurements are already sufficient for reconstruction using the (group) $\ell^{1,2}$-Lasso. In particular, as long as $M = \mathcal{O}(\log(2n/s))$ sensors are used, there is no substantial gain of performance when building a coherently communicating network. Finally, we shall discuss several practical implications and extensions of our approach.**

I. Motivation

Distributed sparse parameter estimation using wireless sensor networks is a promising approach to many environmental monitoring problems and forms a natural application of compressed sensing [1], [2]. Such networks have advantages over conventional sensing technologies in terms of costs, coverage, redundancy, and reliability. Typical applications are structural health monitoring, medical sensor solutions, traffic monitoring as well as warning systems for heat, fire, seismic activities, or meteorologic disturbances. While several communication standards, embedded platforms, and operating systems are available for such problem settings, some of the inherent limitations of these transceiver designs are low transmission and computing power due to battery saving. For example, radio-frequency (RF) components usually only provide low signal quality caused by phase noise and non-linear effects, such as ADC impairments or IQ imbalances (e.g., see "Dirty RF" [3]). It is therefore important to devise approaches to recovery under those non-ideal conditions.

In this work, we consider a model situation where multiple sensor nodes perform individual measurements on the same source. For example, each sensor reading could be a spatial sample of a temperature field in a building or measurements of the water flow and quality taken at different locations. The fluctuation of these quantities are typically specified by only a small number of active parameters which can be often modeled as a *sparse vector* $\boldsymbol{x}_0$ in a known transform domain (e.g., Fourier or wavelets). The task of the wireless sensor network is then to communicate $\boldsymbol{x}_0$ to a central fusion center in an autonomous manner. If all sensors use a common (synchronized) clock (e.g., see [4]) and have sufficient knowledge of the wireless channel, a *coherent cooperative transmission* becomes feasible. Due to the independent channel conditions, the probability of outage can be significantly reduced in that way, which is known as *cooperative* or *multiuser diversity* in communication engineering. In practice, such a network setup is however difficult to achieve. Therefore, *non-coherent cooperative transmission* has been investigated as well, although most works focus on achieving higher power gains at the receiver (e.g., see [5]).

The purpose of our approach is to reformulate the task of estimating the unknown vector $\boldsymbol{x}_0$ as a recovery problem with *superimposed, non-linearly distorted and (non-)coherently communicated measurements*. We will present two simple and robust methods to cope with this challenge and establish rigorous error bounds in the case of Gaussian sensing schemes.

II. Model Setup and Algorithmic Approaches

From a mathematical perspective, the above problem setup can be modeled as follows: Let $\boldsymbol{x}_0 \in \mathbb{R}^n$ be the *sparse vector* that we would like to recover via a network of M autonomous wireless sensors. In the i-th measurement step, all nodes $j = 1, \dots, M$ *directly* transmit their uncoded sensor readings $\langle \boldsymbol{a}_i^j, \boldsymbol{x}_0 \rangle$, where $\boldsymbol{a}_i^j \in \mathbb{R}^n$ denotes the i-th *measurement vector* of the j-th sensor. Note that this model is motivated by the desire to have an *autonomous* and *ad hoc* transmission procedure, bypassing additional resource and time overheads. Due to low-quality hardware components and the wireless channel, this leads to a superposition of non-linearly distorted signals at the receiver. The resulting overall *measurement process* therefore takes the form

$$y_i = \sum_{j=1}^{M} f_j(\langle \boldsymbol{a}_i^j, \boldsymbol{x}_0 \rangle) + e_i, \qquad i = 1, \dots, m. \tag{1}$$

Here, $e_i \sim \mathcal{N}(0, \nu^2)$ is independent noise and the scalar function $f_j \colon \mathbb{R} \to \mathbb{R}$ models the *non-linear distortion* of the j-th sensor device, which could be even *unknown* (see also Figure 1 for an illustration). The functions f_j particularly

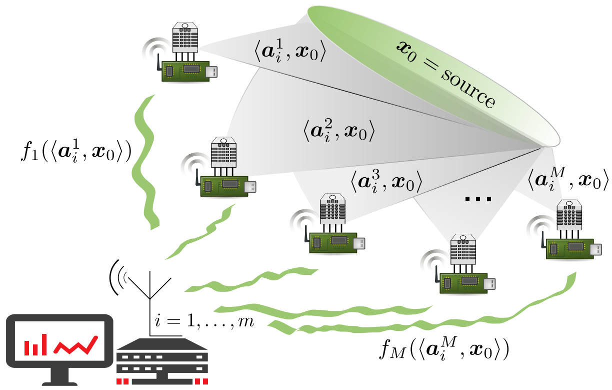

Fig. 1. A schematic sensor network: Each wireless sensor $j = 1, \dots, M$ acquires $i = 1, \dots, m$ individual measurements of a (sparse) source vector $\boldsymbol{x}_0 \in \mathbb{R}^n$ using different "viewpoints" $\boldsymbol{a}_i^j \in \mathbb{R}^n$. These measurements are simultaneously transmitted to a central receiver for recovery. Thereby, the sensor readings $\{\langle \boldsymbol{a}_i^j, \boldsymbol{x}_0 \rangle\}_{i=1}^m$ are affected by a possibly noisy and unknown non-linear distortion $f_j \colon \mathbb{R} \to \mathbb{R}$, which is typically caused by hardware imperfections and the wireless channel.

describe the common effects of the wireless channel and hardware imperfections. Let us briefly discuss two examples that often arise in applications:

Power amplifiers. A highly relevant disturbance is caused by the non-linear characteristics of low-cost amplifiers used at the nodes (e.g., see [6] for a widely used model). In the extreme case, this leads to a clipping at a certain *amplitude level* (threshold) $A > 0$:

$$f^{(A)}(v) := \begin{cases} v, & |v| \leq A, \\ A \cdot \operatorname{sign}(v), & \text{otherwise}. \end{cases} \tag{2}$$

While the sign (phase) of the signal is still preserved in this generic model, the amplitude undergoes some unknown (data-dependent) deformation.

Wireless channel. In a realistic setup, each node modulates its sensor readings on particular waveforms, propagating through the wireless channel after amplification. A filtering and sampling step is then performed at the central receiver. As a simple model, we may assume that the effective channel is approximately constant over the entire communication period, including all transceiver operations. Mathematically, this simply corresponds to a scalar multiplication:

$$f^{(h)}(v) := h \cdot v, \tag{3}$$

where $h \in \mathbb{R}$ is the *channel coefficient*, which may be unknown a priori. If the individual channel configuration of a sensor node is approximately known, one may at least determine the sign (phase) of the channel coefficient and consider $f^{(|h|)}(v) = |h| \cdot v$ instead. A common approach to achieve such a sign-compensation is the concept of *channel reciprocity*. Here, pilot signals are periodically broadcasted by the receiver to all sensors (simultaneously). Each sensor node can now estimate its individual coefficient in the reverse direction (downlink), and in that way, also infer on the parameter h. However, due to limited transmission power, this step usually only allows for compensating for the sign of h.

Put together, each contribution of the superposition (1) could be modeled by a function of the form $f_j = f^{(h_j)} \circ f^{(A)}$. But this is clearly just a simplified model of distortion, since the channel may be outdated in many applications and further disturbances could be involved, such as an oscillator mismatch or phase noise.

Our ultimate goal is to efficiently recover $\boldsymbol{x}_0$ from as few as possible measurement pairs $\{(\{\boldsymbol{a}_i^j\}_j, y_i)\}_{1 \leq i \leq m}$. The recent works of [7], [8] have shown that a single-index model, i.e., (1) with $M = 1$, can be estimated via the vanilla Lasso. As we will see in the following, such a strategy can be also adapted to the more general measurement scheme of (1).

A. Direct Method

In our first approach, we consider the case of *coherent communication* where the sensors are able to compensate for their individual phases to a certain degree. For that purpose, we try to mimic the additive structure of (1) directly by computing *superimposed measurement vectors* $\boldsymbol{a}_i := \sum_{j=1}^M \boldsymbol{a}_i^j$, $i = 1, \dots, m$, and solving the Lasso

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \sum_{i=1}^m (\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - y_i)^2 \quad \text{s.t. } \|\boldsymbol{x}\|_1 \leq R, \tag{P_R^{Dir}}$$

where the tuning parameter $R > 0$ controls the level of sparsity of the minimizer. Remarkably, this program does neither explicitly depend on the non-linearities f_j nor on the number of sensors M.[1] The measurement vectors $\boldsymbol{a}_i^j$ do not have to be known to the fusion center, so that the overall computational costs of (P_R^{Dir}) will not increase as M grows. In other words, the individual identities of the sensor nodes are not relevant for this method, which has important consequences for network planning and maintenance issues.

B. Lifting Method

Due to limited knowledge of the non-linearities f_j, there are some relevant situations where (P_R^{Dir}) fails to work. For example, if the channel changes rapidly or channel reciprocity is unfeasible, a reliable pre-compensation for the phases of the nodes may become difficult. This scenario is referred to as *non-coherent communication*. As a way out, one may rather try fit each sensor measurement of (1) individually. More precisely, we consider the following program:

$$\min_{\substack{\boldsymbol{x}^1, \dots, \boldsymbol{x}^M \\ \in \mathbb{R}^n}} \sum_{i=1}^m \Big(\sum_{j=1}^M \langle \boldsymbol{a}_i^j, \boldsymbol{x}^j \rangle - y_i \Big)^2 \quad \text{s.t. } \|[\boldsymbol{x}^1 \dots \boldsymbol{x}^M]\|_{1,2} \leq R, \tag{P_R^{Lift}}$$

where

$$\|[\boldsymbol{x}^1 \dots \boldsymbol{x}^M]\|_{1,2} := \sum_{k=1}^n \Big(\sum_{j=1}^M ([\boldsymbol{x}^j]_k)^2 \Big)^{1/2}$$

is the $\ell^{1,2}$-matrix norm with $[\boldsymbol{x}^j]_k$ denoting the k-th entry of $\boldsymbol{x}^j$. The purpose of the ℓ^2-group constraint in (P_R^{Lift}) is

[1] Even though the non-linearities f_j could be known, incorporating them directly into (P_R^{Dir}) would lead to a challenging *non-convex* problem.

to enforce a certain "coupling" between all vectors, since every sensor of (1) actually encodes the same source $\boldsymbol{x}_0$. It is sometimes also useful to regard (P_R^{Lift}) as a *lifting method* that concatenates the measurement vectors $\boldsymbol{A}_i := [\boldsymbol{a}_i^1 \dots \boldsymbol{a}_i^M]$ and source vectors $\boldsymbol{X} := [\boldsymbol{x}^1 \dots \boldsymbol{x}^M]$, contained in the "lifted" matrix space $\mathbb{R}^{n\times M}$ with Hilbert-Schmidt inner product $\langle\cdot,\cdot\rangle_{\text{HS}}$. Thus, ($P_R^{\text{Lift}}$) takes the form

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times M}} \sum_{i=1}^{m} (\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle_{\text{HS}} - y_i)^2 \quad \text{s.t. } \|\boldsymbol{X}\|_{1,2} \leq R.$$

Compared to the direct method (P_R^{Dir}), the approach of (P_R^{Lift}) has more degrees of freedom and therefore allows for more accurate estimations. However, this comes along with a (slightly) higher sample complexity and additional computational burdens, since (P_R^{Lift}) needs to operate in the higher dimensional space of $\mathbb{R}^{n\times M}$. Note that a similar lifting approach has been recently studied in [9], [10], considering problems from *self-calibration* and *sparse blind deconvolution*.

III. Main Results

In this paper, we shall present results for the Gaussian case, i.e., $\boldsymbol{a}_i^j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ are independent copies of a standard Gaussian vector (a more general setting will appear in [11]). In order to quantify the degree of distortion that is generated by the output functions f_j in (1), let us first define the following *scaling parameters*:

$$\mu_j := \mathbb{E}_g[f_j(g)\cdot g], \quad j = 1,\dots,M,$$
$$\bar{\mu} := \tfrac{1}{M}\textstyle\sum_{j=1}^M \mu_j,$$

where $g \sim \mathcal{N}(0,1)$. Intuitively, μ_j measures the expected rescaling and sign-change caused by f_j (compared to the identity), whereas $\bar{\mu}$ specifies the "average" rescaling of the entire measurement process (1).

We are now ready to state our main results, extending the recovery guarantees for single-index models from [7], [8]. Let us start with the direct method (P_R^{Dir}):

Theorem 1 (Direct Method)**.** *Let $\boldsymbol{x}_0$ be s-sparse*[2] *and $\|\boldsymbol{x}_0\|_2 = 1$. There exist numerical constants $C, C' > 0$ such that the following holds true with probability at least $1 - 5\exp(-C\delta^2 m)$ for every (fixed) $\delta \in (0,1]$: If*

$$m \geq C'\delta^{-2} s \log(\tfrac{2n}{s}),$$

then any minimizer $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ of (P_R^{Dir}) with $R = \|\bar{\mu}\boldsymbol{x}_0\|_1$ satisfies

$$\|\hat{\boldsymbol{x}} - \bar{\mu}\boldsymbol{x}_0\|_2 \leq (\sigma_{Dir}^2 + \tfrac{\nu^2}{M})^{\frac{1}{2}}\cdot\delta, \tag{6}$$

where[3] *$\sigma_{Dir}^2 := \frac{1}{M}\sum_{j=1}^M \|f_j(g) - \bar{\mu} g\|_{\psi_2}^2$ with $g \sim \mathcal{N}(0,1)$.*

Roughly speaking, successful reconstruction from superimposed sensor measurements (1) becomes feasible if m exceeds $s\log(2n/s)$, which resembles the typical flavor of results from compressed sensing theory. We would like to emphasize that the non-linearities f_j as well as the sensor count M affect the error bound (6) only in terms of the (constant) rescaling factor $\bar{\mu}$ and the *model variance* σ_{Dir}^2. Moreover, the impact of the additive noise $e_i \sim \mathcal{N}(0,\nu^2)$ becomes even smaller in (6) as M grows. This behavior is well-known in the linear case (i.e., all f_j are linear functions) and the above result shows that this "rule-of-thumb" is even valid in the non-linear situation. It particularly implies that enlarging a coherently transmitting wireless sensor network may lead to more accurate and stable recovery.

Operationally, if $\bar{\mu} \neq 0$, the recovered vector $\hat{\boldsymbol{x}}$ only needs to be rescaled by a factor $1/\bar{\mu}$. The significance of Theorem 1 is however lost if $\bar{\mu} \approx 0$, which may arise in the non-coherent setting. The next result shows that this drawback can be handled by the lifting method (P_R^{Lift}):

Theorem 2 (Lifting Method)**.** *Let $\boldsymbol{x}_0$ be s-sparse and $\|\boldsymbol{x}_0\|_2 = 1$. There exist numerical constants $C, C' > 0$ such that the following holds true with probability at least $1 - 5\exp(-C\delta^2 m)$ for every (fixed) $\delta \in (0,1]$: If*

$$m \geq C'\delta^{-2} s \cdot \max\{M, \log(\tfrac{2n}{s})\}, \tag{7}$$

then any minimizer $[\hat{\boldsymbol{x}}^1 \dots \hat{\boldsymbol{x}}^M] \in \mathbb{R}^{n\times M}$ of (P_R^{Lift}) with $R = \|[\mu_1\boldsymbol{x}_0 \dots \mu_M\boldsymbol{x}_0]\|_{1,2}$ satisfies

$$\Big(\tfrac{1}{M}\sum_{j=1}^M \|\hat{\boldsymbol{x}}^j - \mu_j\boldsymbol{x}_0\|_2^2\Big)^{1/2} \leq (\sigma_{Lift}^2 + \tfrac{\nu^2}{M})^{\frac{1}{2}}\cdot\delta, \tag{8}$$

where $\sigma_{Lift}^2 := \frac{1}{M}\sum_{j=1}^M \|f_j(g) - \mu_j g\|_{\psi_2}^2$ with $g \sim \mathcal{N}(0,1)$.

The error bound of (8) states that each column $\hat{\boldsymbol{x}}^j$ of the minimizer approximates a scaled version of $\boldsymbol{x}_0$. In that way, it is even possible to recover the unknown scaling factors μ_j. Interestingly, this essentially solves the *bilinear problem* of factorizing the matrix

$$\boldsymbol{x}_0\boldsymbol{\mu}^\top = [\mu_1\boldsymbol{x}_0 \dots \mu_M\boldsymbol{x}_0] \in \mathbb{R}^{n\times M},$$

where $\boldsymbol{\mu} := (\mu_1,\dots,\mu_M) \in \mathbb{R}^M$. Theorem 2 shows that ($P_R^{\text{Lift}}$) is indeed capable of handling more complicated scenarios than (P_R^{Dir}). In terms of required measurements (7), this improvement comes at no costs as long as the size of the network is $M = \mathcal{O}(\log(2n/s))$. But for larger networks, one has to take significantly more measurements because $m = \mathcal{O}(s\cdot M)$ then grows linearly with the sensor count. But note that the number of unknown non-zero parameters is actually $s + M$. This gap between multiplicative and additive scaling is due to the fact that (P_R^{Lift}) does not account for low-rankness, while $\boldsymbol{x}_0\boldsymbol{\mu}^\top$ is in fact of rank one. While the focus of this work is clearly on *non-linear* distortions, there has been recent progress in matrix factorization from *linear* measurements, breaking the above (multiplicative) complexity barrier (e.g., see [12], [13]).

Remark. There are several important extensions that were omitted here due to length restrictions, but will be contained in the full version of this work [11]:

1) The optimal tuning parameter R of our Lasso approaches is usually unknown in practice. But it can be

[2] That means, at most s entries of $\boldsymbol{x}_0$ are non-zero.

[3] Here, $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm (with respect to g).

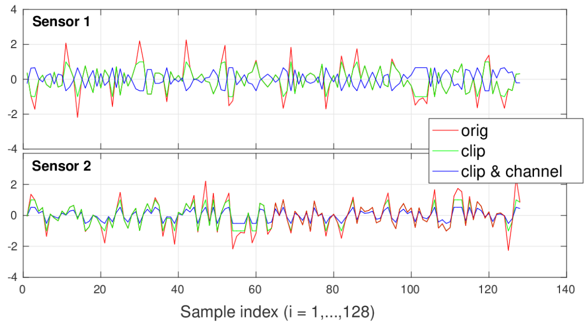

Fig. 2. Transmitted signals of the first two sensors $j = 1, 2$: The original signal (red), clipped channel input (green) and clipped channel output (blue). Note that the first sensor is affected by a sign change, i.e., $\mathrm{sign}(h_j) = -1$.

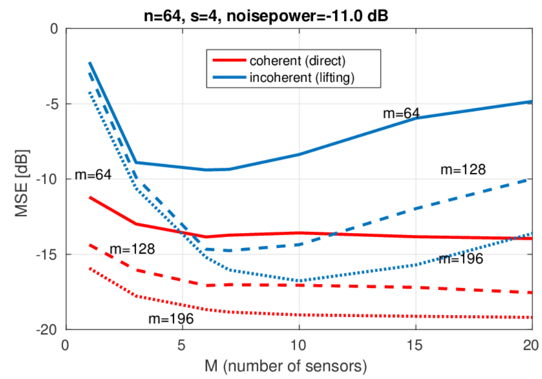

Fig. 3. Mean squared error (MSE) for signal reconstruction with coherent (red) and non-coherent transmission (blue) via (P_R^{Dir}) and (P_R^{Lift}), respectively.

shown that similar error bounds (with an additional error term) still hold true if R is not exactly chosen as in Theorem 1 and Theorem 2.

2) One may go beyond the assumption of Gaussianity and allow for sub-Gaussian random vectors $\boldsymbol{a}_i^j$ with invertible covariance matrix.
3) Different assumptions on the noise are also permitted in (1) and $\boldsymbol{x}_0$ does not have to be exactly sparse.

IV. Practical Scope and Numerics

Our main results indicate that, for coherent transmission and recovery, the direct method (P_R^{Dir}) should be preferred over (P_R^{Lift}), since it is more efficient and requires less measurements. However, Theorem 1 also states that a successful source estimation relies on the assumption that $|\bar{\mu}| \gg 0$, which in turn asks for a certain prior knowledge of the channel environment. Consequently, the lifting method (P_R^{Lift}) becomes relevant in the non-coherent situation, where the (signs of the) sensor parameters μ_j are mostly unknown. Although taking more measurements for recovery of $\boldsymbol{x}_0$, it eventually even allows us to estimate the $\mu_1, \dots, \mu_M$. Thus, to some extent, (P_R^{Lift}) enables to "learn" the underlying system configuration. Once such a "calibration step" has been done, we may continue using the direct approach, tuned by our additional knowledge of the sensors (see also Section VI).

Numerical Experiments

In the following, we shall validate our two recovery approaches numerically. For this purpose, we have generated normalized s-sparse random vectors $\boldsymbol{x}_0 \in \mathbb{R}^n$ with $n = 64$ and $s = 4$. Each of the $j = 1, \dots, M$ sensors performs $i = 1, \dots, m$ measurements of $\boldsymbol{x}_0$ with $\boldsymbol{a}_i^j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Figure 2 illustrates the signal vectors $[f_j(\langle \boldsymbol{a}_i^j, \boldsymbol{x}_0\rangle)]_{1\leq i\leq m}$ for the first two sensors $j = 1, 2$. Each of the $m = 128$ measurements taken by a sensor node (red) undergoes a clipping (2) with threshold $A = 1$ (green) and is then transmitted into the channel, which corresponds to a scalar multiplication by coefficients $h_j \sim \mathcal{N}(0, 1)$ according to (3) (blue).

For testing coherent transmission, we assume that the phases of the individual channels have been already resolved, as discussed in the course of equation (3). Hence, we consider $f_j(v) = |h_j| \cdot f^{(A)}(v)$ as non-linearities here. The recovery is then performed by solving (P_R^{Dir}) with $R = \bar{\mu} \cdot \sqrt{s}$ and rescaling by $1/\bar{\mu}$. For the non-coherent setting, we just have $f_j(v) = h_j \cdot f^{(A)}(v)$ and apply (P_R^{Lift}) with $R = \sqrt{M} \cdot \sqrt{s}$ for recovery. The reconstruction results are shown in Figure 3 for different numbers of sensors M and measurements m. The MSE of the coherent case (red) decreases as m and M grow. In particular, it becomes almost constant for large M, which coincides with the observation that the model variance σ^2_{Dir} dominates the error bound (6) in Theorem 1. On the contrary, there is obviously a "turning point" in the non-coherent setup (blue). For sufficiently small sensor counts, the recovery error indeed drops with M up to a certain level. Above this threshold, more measurements are required to achieve the same accuracy. This behavior is precisely reflected by the statement of Theorem 2, which indicates that $M = \mathcal{O}(\log(2n/s))$ is the "ideal" size of a network.

V. Proof Ideas and Sketch

The key idea of our approaches (P_R^{Dir}) and (P_R^{Lift}) is to approximate (or to fit) *non-linear* measurements $y_1, \dots, y_m$ by an appropriate *linear* counterpart. These strategies are particular examples of a more general concept, namely the *K-Lasso* (cf. [7]):

$$\min_{\tilde{\boldsymbol{x}} \in \mathbb{R}^d} \sum_{i=1}^{m} (\langle \tilde{\boldsymbol{a}}_i, \tilde{\boldsymbol{x}}\rangle - y_i)^2 \quad \text{s.t. } \tilde{\boldsymbol{x}} \in K. \qquad (P_K)$$

Here, $\tilde{\boldsymbol{a}}_1, \dots, \tilde{\boldsymbol{a}}_m \in \mathbb{R}^d$ are certain measurement vectors and the convex *signal set* $K \subset \mathbb{R}^d$ imposes structural assumptions on the solution. For the direct method, we have chosen $\tilde{\boldsymbol{a}}_i := \boldsymbol{a}_i = \sum_{j=1}^M \boldsymbol{a}_i^j \in \mathbb{R}^n$ and K equals a scaled ℓ^1-unit ball ($d = n$), whereas for the lifting method, $\tilde{\boldsymbol{a}}_i := \boldsymbol{A}_i = [\boldsymbol{a}_i^1 \dots \boldsymbol{a}_i^M] \in \mathbb{R}^{n\times M}$ and K is a scaled $\ell^{1,2}$-unit ball ($d = n \cdot M$).

A major challenge is now to establish a relationship between a minimizer of (P_K) and the underlying observation model of $y_1, \dots, y_m$. The following recovery result gives a general answer to this question and forms the basis of Theorem 1 and Theorem 2. Its proof is based on recent results on multiplier processes [14] and matrix concentration [15].

Theorem 3 ([11]). *Let $\tilde{\boldsymbol{a}}_1, \dots, \tilde{\boldsymbol{a}}_m$ be i.i.d. standard Gaussian vectors in $\mathbb{R}^d$. Moreover, assume that $y_1, \dots, y_m$ are drawn independently from a certain (unknown) random distribution ("measurement model"). Fix any vector $\tilde{\boldsymbol{x}}_0 \in K$, where $K \subset \mathbb{R}^d$ is known and convex. Then, there exist numerical constants $C, C', C'' > 0$ such that the following holds with probability at least $1 - 5\exp(-C\delta^2 m)$ for every (fixed) $\delta \in (0, 1]$: If* [4]

$$m \geq C'\delta^{-2}[w_1(\mathcal{C}(K, \tilde{\boldsymbol{x}}_0))]^2, \tag{9}$$

then any minimizer $\hat{\tilde{\boldsymbol{x}}} \in \mathbb{R}^d$ of (P_K) satisfies

$$\|\hat{\tilde{\boldsymbol{x}}} - \tilde{\boldsymbol{x}}_0\|_2 \leq \sigma_{\tilde{\boldsymbol{x}}_0} \cdot \delta + C''\tau_{\tilde{\boldsymbol{x}}_0}, \tag{10}$$

where

$$\tau_{\tilde{\boldsymbol{x}}_0} := \tau(\tilde{\boldsymbol{a}}_i, y_i, \tilde{\boldsymbol{x}}_0) := \sup_{\tilde{\boldsymbol{x}} \in S^{d-1}} \mathbb{E}_{\tilde{\boldsymbol{a}}_i, y_i}[(\langle \tilde{\boldsymbol{a}}_i, \tilde{\boldsymbol{x}}_0\rangle - y_i)\langle \tilde{\boldsymbol{a}}_i, \tilde{\boldsymbol{x}}\rangle],$$

$$\sigma_{\tilde{\boldsymbol{x}}_0} := \sigma(\tilde{\boldsymbol{a}}_i, y_i, \tilde{\boldsymbol{x}}_0) := \|\langle \tilde{\boldsymbol{a}}_i, \tilde{\boldsymbol{x}}_0\rangle - y_i\|_{\psi_2}.$$

It is somewhat surprising that Theorem 3 holds for *every* choice of $\tilde{\boldsymbol{x}}_0$. However, in order to turn the error bound of (10) into a meaningful statement, one needs to show that the *expected noise correlation* $\tau_{\tilde{\boldsymbol{x}}_0}$ is sufficiently small. We shall see below that for each of our methods there exists a choice of $\tilde{\boldsymbol{x}}_0$ for which $\tau_{\tilde{\boldsymbol{x}}_0} = 0$. This basically means that the *model mismatch* $z_i := \langle \tilde{\boldsymbol{a}}_i, \tilde{\boldsymbol{x}}_0\rangle - y_i$ (sometimes referred to as "noise") is uncorrelated with the measurement vectors $\tilde{\boldsymbol{a}}_i$. In general, $\tau_{\tilde{\boldsymbol{x}}_0}$ forms the key parameter that specifies the capability of the K-Lasso of fitting non-linear models. For our superimposed measurement model (1), we have the following:

$$\text{Direct method:} \quad \tau(\tfrac{1}{\sqrt{M}}\boldsymbol{a}_i, \tfrac{1}{\sqrt{M}}y_i, \bar{\mu}\boldsymbol{x}_0) = 0, \tag{11}$$

$$\text{Lifting method:} \quad \tau(\boldsymbol{A}_i, y_i, \boldsymbol{x}_0\boldsymbol{\mu}^\top) = 0. \tag{12}$$

The second important quantity of Theorem 3 is clearly the conic mean width $w_1(\mathcal{C}(K, \tilde{\boldsymbol{x}}_0))$. It is formally defined as

$$w_1(\mathcal{C}(K, \tilde{\boldsymbol{x}}_0)) := \mathbb{E}_{\boldsymbol{g}}[\sup_{\tilde{\boldsymbol{x}} \in \mathcal{C}(K, \tilde{\boldsymbol{x}}_0) \cap B_2^d} \langle \boldsymbol{g}, \tilde{\boldsymbol{x}}\rangle], \tag{13}$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\mathcal{C}(K, \tilde{\boldsymbol{x}}_0) := \{\lambda(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}_0) \in \mathbb{R}^d \mid \tilde{\boldsymbol{x}} \in K, \lambda > 0\}$ denotes the *descent cone* of K at $\tilde{\boldsymbol{x}}_0$. Intuitively, this geometric parameter measures the complexity of K in a neighborhood of the source vector $\tilde{\boldsymbol{x}}_0$, and by (9), it gets related to the sample complexity of the actual recovery problem (see also [16]). In our specific setup, one can establish the following upper bounds on $w_1(\mathcal{C}(K, \tilde{\boldsymbol{x}}_0))$:

Direct method: For $K := \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_1 \leq \|\bar{\mu}\boldsymbol{x}_0\|_1\}$,

$$w_1(\mathcal{C}(K, \bar{\mu}\boldsymbol{x}_0)) \leq \sqrt{Cs\log(\tfrac{2n}{s})}. \tag{14}$$

Lifting method: For $K := \{\boldsymbol{X} \in \mathbb{R}^{n\times M} \mid \|\boldsymbol{X}\|_{1,2} \leq \|\boldsymbol{x}_0\boldsymbol{\mu}^\top\|_{1,2}\}$,

$$w_1(\mathcal{C}(K, \boldsymbol{x}_0\boldsymbol{\mu}^\top)) \leq \sqrt{Cs \cdot \max\{M, \log(\tfrac{2n}{s})\}}. \tag{15}$$

Finally, observing that $\sigma_{\bar{\mu}\boldsymbol{x}_0} \leq \sigma_{\text{Dir}} + \nu/\sqrt{M}$ and $\sigma_{\boldsymbol{x}_0\boldsymbol{\mu}^\top} \leq \sigma_{\text{Lift}} + \nu/\sqrt{M}$, the claims of Theorem 1 and Theorem 2 immediately follow from Theorem 3, using (11), (12) and (14), (15), respectively.

[4]Here, $w_1(\mathcal{C}(K, \tilde{\boldsymbol{x}}_0))$ denotes the so-called *conic mean width* of K at $\tilde{\boldsymbol{x}}_0$. A formal definition is given below in (13).

VI. Future Directions

While both approaches (P_R^{Dir}) and (P_R^{Lift}) come along with several up- and downsides, it is quite natural to consider a *hybrid method* that combines the "raw" measurement vectors $\boldsymbol{a}_i^1, \dots, \boldsymbol{a}_i^M$ in a more sophisticated manner. In that way, one may easily improve the recovery performance if prior information about the network is given, e.g., if sensor configurations are partially known. The abstract framework of Theorem 3 would again enable a rigorous treatment of this setting.

Another extension of great practical relevance is to go beyond (sub-)Gaussian designs and to allow for structured measurements, such as random samples from an orthonormal system.

Acknowledgment

M.G. is supported by the Einstein Center for Mathematics Berlin (ECMath) under project grant CH2. P.J. is partially supported by DFG grant JU 2795/3.

References

[1] C. Luo, F. Wu, J. Sun, and C. Chen, "Compressive Data Gathering for Large-Scale Wireless Sensor Networks," in *Proceedings of the 15th Annual International Conference on Mobile Computing and Networking*, 2009, pp. 145–156.

[2] G. Cao, P. Jung, S. Stanczak, and F. Yu, "Data Aggregation and Recovery in Wireless Sensor Networks Using Compressed Sensing," *Int. J. Sens. Netw.*, vol. 22, no. 4, pp. 209–219, 2016.

[3] G. Fettweis, M. Löhning, D. Petrovic, M. Windisch, P. Zillmann, and W. Rave, "Dirty RF: A new paradigm," *Int. J. Wireless Inform. Network.*, vol. 14, no. 2, pp. 133–148, 2007.

[4] D. R. Brown, G. B. Prince, and J. A. McNeill, "A method for carrier frequency and phase synchronization of two autonomous cooperative transmitters," in *IEEE 6th Workshop on Signal Processing Advances in Wireless Communications, 2005.*, 2005, pp. 260–264.

[5] A. Scaglione and Y.-W. Hong, "Opportunistic large arrays: cooperative transmission in wireless multihop ad hoc networks to reach far distances," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2082–2092, 2003.

[6] C. Rapp, "Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digital Sound Broadcasting Signal," in *Second European Conference on Satellite Communications*, 1991, pp. 179–184.

[7] Y. Plan and R. Vershynin, "The generalized Lasso with non-linear observations," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1528–1537, 2016.

[8] M. Genzel, "High-Dimensional Estimation of Structured Signals From Non-Linear Observations With General Convex Loss Functions," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1–19, 2017.

[9] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Probl.*, vol. 31, no. 11, p. 115002, 2015.

[10] A. Flinth, "Sparse blind deconvolution and demixing through $\ell_{1,2}$-minimization," 2016, arXiv preprint: 1609.06357.

[11] M. Genzel and P. Jung, "Recovering Structured Data From Superimposed Non-Linear Measurements," *in preparation*, 2017.

[12] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[13] K. Lee, Y. Wu, and Y. Bresler, "Near optimal compressed sensing of a class of sparse low-rank matrices via sparse power factorization," 2016, arXiv preprint: 1312.0525.

[14] S. Mendelson, "Upper bounds on product and multiplier empirical processes," *Stoch. Proc. Appl.*, vol. 126, no. 12, pp. 3652–3680, 2016.

[15] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin, "A simple tool for bounding the deviation of random matrices on geometric sets," in *Geometric aspects of functional analysis*, ser. Lecture Notes in Math. Springer, Berlin, 2016, to apprear (arXiv:1603.00897).

[16] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.