Contents lists available at ScienceDirect
Digital Signal Processing
www.elsevier.com/locate/dsp

Enhancing the fundamental limits of sparsity pattern recovery

Zahra Shaeiri, Mohammad-Reza Karami *, Ali Aghagolzadeh

Department of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran

A R T I C L E I N F O

Article history:
Available online xxxx

Keywords:
Compressive sensing
Sparsity pattern recovery
Performance bounds of algorithms
Joint-Typicality decoder
Maximum Likelihood estimation

A B S T R A C T

Detecting the sparsity pattern or support set of a sparse vector from a small number of noisy linear measurements is a challenging problem in compressed sensing. This paper considers the problem of support recovery when statistical side information is available. From the standard linear and noisy measurement model with arbitrary sensing matrix and Gaussian additive noise and by exploiting the side information, a new linear model is derived which benefits from a larger sample size. The common potential benefits of the increase in the number of samples are revealed. The stability guarantees are then analyzed based on the new model. Two decoding schemes are taken for the support recovery task from the new framework, namely, Maximum Likelihood (ML) and Joint-Typicality (JT) decoding. Performance bounds of the support recovery from the new framework are developed and upper bounds are derived on the error probability of these decoders when they are fed with the prior knowledge which is the statistical properties of the new measurement noise. Finally, an extension is provided for when the noise is non-Gaussian. The results show that with the aid of the prior knowledge and using the new framework one can push the performance limits of the sparsity pattern recovery significantly. The approach is supported by extensive simulations including extension of LASSO to the new framework.

1. Introduction

Compressed sensing, one of the emerging research field in the last decade, has established that a small set of observations acquired via linear projections is adequate for estimation and recovery of signal, when the signal of interest has specific characteristics, *i.e.* is sparse in a certain basis. More concretely, it seeks an estimate of $\boldsymbol{w} \in \mathbb{R}^N$ via solving an underdetermined linear system of equations:

$$\boldsymbol{m} = \boldsymbol{S}\boldsymbol{w} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{S} \in \mathbb{R}^{n\times N}$ is a fat sensing or measurement matrix with $n \ll N$, $\boldsymbol{m} \in \mathbb{R}^n$ is the output noisy measurement of the k-sparse vector $\boldsymbol{w}$, *i.e.* it has $k < n \ll N$ nonzero entries, and $\boldsymbol{e}$ is the additive noise vector. Efforts to solve this problem have resulted in several breakthrough algorithms with affordable complexity, such as message passing [1], LASSO [2], Orthogonal Matching Pursuit (OMP) and variants thereof [3–6]. Concurrently, a great deal of researches derived performance and stability guarantees for these algorithms [7–10]. The most challenging aspect of sparse signal recovery is estimating its true support set or positions of its nonzero entries. Upon decoding the support set, the signal can be estimated simply by solving a least squares problem. Besides, there are many applications where finding the correct support set is more important than completely recovering the signal. Magnetoencephalography (MEG), electroencephalography (EEG), cognitive radio, subset selection in regression, and multi-user communication systems are examples of such applications [11–14]. In this paper, we are concerned with fundamental limits of the support recovery problem. Fundamental limits of any recovery problem can be achieved by analyzing the performance of an optimal decoder. These limits are highly valuable, since they reveal the gap between the performance of any tractable recovery algorithm and the ultimate performance limits. In recent years in several works, the authors studied information-theoretic limits of any estimator for exact and approximate support recovery and for single and multiple measurement vector models [15–38]. In this paper, we are interested in deriving the performance limits of sparsity pattern recovery when the decoder has some types of prior information. A new statistical model with a different measurement matrix and a different measurement noise vector is presented which is a reformulated version of the standard model (1). For the mentioned model, upper bounds of the probability of error events pertaining to the support recovery problem are derived for ML and JT decoders when the decoders have prior knowledge (for instance first and second moments of the new measurement noise vector). The results are further extended to the models with non-Gaussian ad-

* Corresponding author.
E-mail addresses: z.shaeiri@stu.nit.ac.ir (Z. Shaeiri), mkarami@nit.ac.ir (M.-R. Karami), aghagol@nit.ac.ir (A. Aghagolzadeh).

http://dx.doi.org/10.1016/j.dsp.2017.06.027
1051-2004/

ditive noise. Moreover, testing the new framework on the LASSO and getting considerable improvements proves the validity of the results numerically.

The rest of the paper is organized as follows. In section 2.1 the problem formulation and our contributions are stated and in section 2.2 relation to past works is discussed. The new framework is introduced and the stability guarantees are derived in section 3. Using the proposed model, performance bounds of the sparsity pattern recovery in terms of probability of error for the ML and the JT decoders and extension of the results to the non-Gaussian noise are provided in section 4. Simulations are presented and discussed in section 5. Conclusions are drawn in section 6. Finally, appendices are provided in section 7.

Notation

Bold letters are used for vectors and matrices and regular letters for scalar variables. $\mathbf{I}$ stands for the identity matrix and $\mathbf{0}$ denotes zero vector both with appropriate dimensions. Given Matrix $\boldsymbol{S}$ and vector $\boldsymbol{w}$, s_{ij} and w_i denote ij-th element of $\boldsymbol{S}$ and i-th element of $\boldsymbol{w}$, respectively. Given set $\mathcal{I}$, $\boldsymbol{w}_{\mathcal{I}}$ is a vector that consists of elements of $\boldsymbol{w}$ indexed by set $\mathcal{I}$, and $\boldsymbol{S}_{\mathcal{I}}$ is a sub-matrix of $\boldsymbol{S}$ containing only columns associated with the index set $\mathcal{I}$. $\mathbf{T}_{\boldsymbol{S}_{\mathcal{I}}} = \boldsymbol{S}_{\mathcal{I}}(\boldsymbol{S}_{\mathcal{I}}^t \boldsymbol{S}_{\mathcal{I}})^{-1}\boldsymbol{S}_{\mathcal{I}}^t$ is the orthogonal projection matrix onto the subspace spanned by the columns of $\boldsymbol{S}_{\mathcal{I}}$ and the orthogonal projection matrix onto the orthogonal complement of this subspace is denoted using $\mathbf{T}^{\perp}_{\boldsymbol{S}_{\mathcal{I}}} = \mathbf{I} - \mathbf{T}_{\boldsymbol{S}_{\mathcal{I}}}$. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices with identical number of columns. The Khatri–Rao product of $\mathbf{A}$ and $\mathbf{B}$ which is the column-wise Kronecker product of them is denoted by $\mathbf{A} \odot \mathbf{B}$. $|\mathcal{I}|$ denotes cardinality of the set $\mathcal{I}$. Finally, $Pr(.)$ refers to the probability measure.

2. Background

In the following subsections we first present the problem statement and outline of our contributions and then review the relevant works and their relation to our proposed approach.

2.1. Problem statement and contributions

Consider the measurement model (1) where $\boldsymbol{w}$ is k-sparse and k is known. In Theorems 4.1 and 4.2 we assume that the noise vector $\boldsymbol{e}$ has *i.i.d.* $N\left(0, \sigma^2\right)$ components. In Theorems 4.3 and 4.4 we assume that $\boldsymbol{e}$ is an arbitrary random vector. The majority of previous works have assumed that the measurement matrix has Gaussian *i.i.d.* entries. In contrast we take an arbitrary measurement matrix, *i.e.* it can have non-*i.i.d.* elements with any distribution (in its general domain). The unknown k-sparse vector $\boldsymbol{w}$ is assumed deterministic and its support set is denoted as i_{true} which is a k-element subset of the index set $\{1, 2, ..., N\}$. We assume symmetry with respect to i_{true}, *i.e.* it is uniformly distributed over all $\binom{N}{k}$ possible subsets of size k. Taking into account that we are only interested in the support set recovery, a strategy for this recovery has to be exploited. This is achieved by using a decoder which is a mapping from the given knowledge, namely $\boldsymbol{m}$, $\boldsymbol{S}$, and statistical characteristics of the additive noise, to an estimated support set with cardinality k. $\widehat{i}_D$ stands for such estimate with subscript indicating the type of decoder. The performance of the decoder is evaluated using exact and partial support recovery metrics and is derived in terms of the probability of error. Given the true support set i_{true}, the probability of sparsity pattern recovery error for the zero-one error metric or for the perfect recovery can be expressed as:

$$P_e = \sum_{i_{true}} Pr(\widehat{i}_D \neq i_{true} \mid i_{true}) Pr(i_{true}) \tag{2}$$

where the probability is taken over all measurement vectors. We also investigate approximate recovery considering the following less stringent form of the error probability:

$$P_{e,a} = \sum_{i_{true}} Pr\left(\frac{|\widehat{i}_D \cap i_{true}|}{|i_{true}|} < \theta | i_{true}\right) Pr\left(i_{true}\right) \tag{3}$$

in which $\theta \in (0, 1)$. $P_{e,a}$ considers recovery of most of the subspace information of the unknown vector $\boldsymbol{w}$ [21]. We desire to obtain tight upper bounds on the error probabilities of ML and JT estimators.

Maximum Likelihood decoder

We investigate performance of the ML decoder for support recovery which is optimal when there is no prior knowledge about $\boldsymbol{w}$ other than it being k-sparse. As was stated previously the true support set is denoted by i_{true} and the estimated support set is denoted as $\widehat{i}_{ML}$. The decoder searches exhaustively over all possible subsets $\widehat{i}_{ML} \subset \{1, 2, ..., N\}$ of size k and chooses the one which minimizes the quadratic norm $\|\boldsymbol{m} - \boldsymbol{S}_{\widehat{i}_{ML}} \widehat{\boldsymbol{w}}_{\widehat{i}_{ML}}\|_2^2$, where $\widehat{\boldsymbol{w}}_{\widehat{i}_{ML}} = \left(\boldsymbol{S}^T_{\widehat{i}_{ML}} \boldsymbol{S}_{\widehat{i}_{ML}}\right)^{-1} \boldsymbol{S}^T_{\widehat{i}_{ML}} \boldsymbol{m}$ is the least square estimator of $\boldsymbol{w}_{\widehat{i}_{ML}}$. The ML decoder makes an error if:

$$\|\boldsymbol{m} - \boldsymbol{S}_{\widehat{i}_{ML}} \widehat{\boldsymbol{w}}_{\widehat{i}_{ML}}\|_2^2 < \|\boldsymbol{m} - \boldsymbol{S}_{i_{true}} \widehat{\boldsymbol{w}}_{i_{true}}\|_2^2.$$

It is easy to see that:

$$\|\boldsymbol{m} - \boldsymbol{S}_{\widehat{i}_{ML}} \widehat{\boldsymbol{w}}_{\widehat{i}_{ML}}\|_2^2 = \|\left(\mathbf{I} - \mathbf{T}_{\boldsymbol{S}_{\widehat{i}_{ML}}}\right)\boldsymbol{m}\|_2^2 = \|\mathbf{T}^{\perp}_{\boldsymbol{S}_{\widehat{i}_{ML}}} \boldsymbol{m}\|_2^2,$$

and:

$$\|\boldsymbol{m} - \boldsymbol{S}_{i_{true}} \widehat{\boldsymbol{w}}_{i_{true}}\|_2^2 = \|\left(\mathbf{I} - \mathbf{T}_{\boldsymbol{S}_{i_{true}}}\right)\boldsymbol{m}\|_2^2 = \|\mathbf{T}^{\perp}_{\boldsymbol{S}_{i_{true}}} \boldsymbol{m}\|_2^2.$$

By assuming that $\text{rank}(\boldsymbol{S}_{i_{true}}) = k$ and for every $\widehat{i}_{ML}$, $\text{rank}(\boldsymbol{S}_{\widehat{i}_{ML}}) = k$, the error event associated with the ML decoder can be considered as [15]:

$$\xi_{ML} = \left\{\|\mathbf{T}^{\perp}_{\boldsymbol{S}_{\widehat{i}_{ML}}} \boldsymbol{m}\|^2 - \|\mathbf{T}^{\perp}_{\boldsymbol{S}_{i_{true}}} \boldsymbol{m}\|^2 < 0\right\}. \tag{4}$$

The ML decoder declares an error when at least one $\widehat{i}_{ML} \neq i_{true}$ is preferred to i_{true}. The overall error probability is then:

$$P_e = Pr\left(\bigcup_{\substack{\widehat{i}_{ML} \neq i_{true}, \\ |\widehat{i}_{ML}| = k}} \xi_{ML}\right). \tag{5}$$

Joint Typicality decoder

We also investigate the performance of a JT decoder. JT decoder is known to be asymptotically optimal. It characterizes the events based on their typicality. Thus, error events are expressed based on atypicality. Here, the following definition of the joint typicality property is exploited [21,39].

Joint Typicality property: The measurement vector $\boldsymbol{m}$ of the measurement model (1) and a set of indices $\mathcal{I} \subset \{1, 2, ..., N\}$ with cardinality k are υ-jointly typical, if $\text{rank}(\boldsymbol{S}_{\mathcal{I}}) = k$, and for any $\upsilon > 0$ the following event occurs:

$$\xi^{\mathcal{I}}_{JT} = \left\{\left|\frac{1}{n}\|\mathbf{T}^{\perp}_{\boldsymbol{S}_{\mathcal{I}}} \boldsymbol{m}\|^2 - \frac{1}{n}\mathbb{E}\left\{\|\mathbf{T}^{\perp}_{\boldsymbol{S}_{\mathcal{I}}} \boldsymbol{e}\|^2\right\}\right| < \upsilon\right\}. \tag{6}$$

The JT decoder outputs an estimate of the support set denoted by $\widehat{i}_{JT}$ which is a k-element subset of $\{1, 2, ..., N\}$. An error is declared when $\widehat{i}_{JT}$: