Sparse Estimation From Noisy Observations of an Overdetermined Linear System

Liang Dai\textsuperscript{a} and Kristiaan Pelckmans\textsuperscript{a}

\textsuperscript{a}Division of Systems and Control, Department of Information Technology, Uppsala University, Sweden
e-mail: liang.dai@it.uu.se, kristiaan.pelckmans@it.uu.se.

Abstract

This note studies a method for the efficient estimation of a finite number of unknown parameters from linear equations, which are perturbed by Gaussian noise. In case the unknown parameters have only few nonzero entries, the proposed estimator performs more efficiently than a traditional approach. The method consists of three steps: (1) a classical Least Squares Estimate (LSE), (2) the support is recovered through a Linear Programming (LP) optimization problem which can be computed using a soft-thresholding step, (3) a de-biasing step using a LSE on the estimated support set. The main contribution of this note is a formal derivation of an associated ORACLE property of the final estimate. That is, when the number of samples is large enough, the estimate is shown to equal the LSE based on the support of the true parameters.

Key words: System identification; Parameter estimation; Sparse estimation.

1 Problem settings

This note considers the estimation of a sparse parameter vector from noisy observations of a linear system. The formal definition and assumptions of the problem are given as follows. Let $n > 0$ be a fixed number, denoting the dimension of the underlying true parameter vector. Let $N > 0$ denote the number of equations (‘observations’). The observed signal $y \in \mathbb{R}^N$ obeys the following system:

$$y = Ax^0 + v,$$

where the elements of the vector $x^0 \in \mathbb{R}^n$ are considered to be the fixed but unknown parameters of the system. Moreover, it is assumed that $x_0$ is $s$-sparse (i.e. there are $s$ nonzero elements in the vector). Let $\mathcal{T} \subset \{1, \ldots, n\}$ denote the support set of $x^0$ (i.e. $x^0_i = 0 \iff i \notin \mathcal{T}$). Let $\mathcal{T}^c$ be the complement of $\mathcal{T}$, i.e. $\mathcal{T} \cup \mathcal{T}^c = \{1, 2, \ldots, n\}$ and $\mathcal{T} \cap \mathcal{T}^c = \emptyset$. The elements of the vector $v \in \mathbb{R}^N$ are assumed to follow the following distribution

$$v \sim \mathcal{N}(0, cI_N),$$

where $0 < c \in \mathbb{R}$.

Applications of such setup appear in many places, to name a few, see the applications discussed in Kump, Bai, Chan, Eichinger, and Li (2012) on the detection of nuclear material, and in Kukreja (2009) on model selection for aircraft test modeling (see also the Experiment 2 in Rojas and Hjalmarsson (2011) on the model selection for the AR model). In the experiment section, we will demonstrate an example which finds application in line spectral estimation, see Stoica and Moses (1997).

The matrix $A \in \mathbb{R}^{N \times n}$ with $N > n$ is a deterministic ‘sensing’ or ‘regressor’ matrix. Such a setting ($A$ is a ‘tall’ matrix) makes it different from the setting studied in compressive sensing, where the sensing matrix is ‘fat’, i.e. $N \ll n$. For an introduction to the compressive sensing theory, see e.g. Donoho (2006); Candès and Wakin (2008).

Denote the Singular Value Decomposition (SVD) of matrix $A \in \mathbb{R}^{N \times n}$ as

$$A = U \Sigma V^T,$$

in which $U \in \mathbb{R}^{N \times n}$ satisfies $U^T U = I_n$, $V \in \mathbb{R}^{n \times n}$ satisfies $V^T V = I_n$, and $\Sigma \in \mathbb{R}^{n \times n}$ is a diagonal matrix $\Sigma = \text{diag}(\sigma_1(A), \sigma_2(A), \ldots, \sigma_n(A))$. The results below make the following assumptions on $A$:

**Definition 1** We say that $\{A \in \mathbb{R}^{N \times n}\}_N$ are sufficiently rich if there exists a finite $N_0$ and $0 < c_1 \leq c_2$ such that for all $N > N_0$ the corresponding matrices
\( \mathbf{A} \in \mathbb{R}^{N \times n} \) obey

\[
    c_1 \sqrt{N} \leq \sigma_1(\mathbf{A}) \leq \sigma_2(\mathbf{A}) \leq \ldots \leq \sigma_n(\mathbf{A}) \leq c_2 \sqrt{N}, \tag{4}
\]

where \( \sigma_i(\mathbf{A}) \) denotes the \( i \)-th singular value of the matrix \( \mathbf{A} \), \( c_1, c_2 \in \mathbb{R}^+ \).

Note that the dependence of \( \mathbf{A} \) on \( N \) is not stated explicitly in order to avoid notational overload.

In Rojas and Hjalmarsson (2011) and Zou (2004), the authors make the assumption on \( \mathbf{A} \) that the sample covariance matrix \( \frac{1}{N} \mathbf{A}^T \mathbf{A} \) converges to a finite, positive-definite matrix:

\[
    \lim_{N \to \infty} \frac{1}{N} \mathbf{A}^T \mathbf{A} = \mathbf{D} > 0. \tag{5}
\]

This assumption is also known as Persistent Excitation (PE), see e.g. Söderström and Stoica (1989). Note that our assumption in Eq. (4) covers a wider range of cases. For example, Eq. (4) does not require the singular values of \( \sqrt{N} \mathbf{A} \) to converge, while only requires that they lie in \([c_1, c_2]\) when \( N \) increases.

Classically, properties of the Least Square Estimation (LSE) under the model given in Eq. (1) are given by the Gauss-Markov theorem. It says that the Best Linear Unbiased Estimation (BLUE) of \( \mathbf{A} \) that achieves the minimum variance among unbiased linear estimators. For details, please see the discussions in Leeb and Pötscher (2003); Kale (1983).

In this note, we will present a novel way to estimate the sparse vector \( \mathbf{x}^0 \), which also possesses the ORACLE property while with a lighter computational cost. The proposed estimator consists of three steps, in the first step, a normal LSE is conducted, the second step is to solve a LP (Linear Programming) problem, whose solution is given by a soft-thresholding step, finally, redo the LSE based on the support set of the estimated vector from the previous LP problem. Details will be given in Section 2.

In the following, the lower bold case will be used to denote a vector and capital bold characters are used to denote matrices. The subsequent sections are organized as follows. In section 2, we will describe the algorithm in detail and an analytical solution to the LP problem is given. In Section 3, we will analyze the algorithm in detail. In Section 4, we conduct several examples to illustrate the efficacy of the proposed algorithm and compare the proposed algorithm with other algorithms. Finally, we draw conclusions of the note.

### 2 Algorithm Description

The algorithm consists of the following three steps:

- **LSE**: Compute the SVD of matrix \( \mathbf{A} \) as \( \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \). The Least Square Error estimate (LSE) is then given as \( \mathbf{x}^0 = \mathbf{V} \Sigma^{-1} \mathbf{U}^T \mathbf{y} = \mathbf{A}^\dagger \mathbf{y} \).
- **LP**: Choose \( 0 < \epsilon < 1 \) and solve the following Linear Programming problem:

\[
    \mathbf{x}^{lp} = \arg \min_{\mathbf{x}} ||\mathbf{x}||_1 \text{ s.t. } ||\mathbf{x} - \mathbf{x}^0||_\infty \leq \lambda, \tag{6}
\]

where \( \lambda = \sqrt{\frac{c_0}{N}} \). Detect the support set \( \mathcal{T}^{lp} \) of \( \mathbf{x}^{lp} \).
- **RE-LSE**: Form the matrix \( \mathbf{A}_{\mathcal{T}^{lp}} \), which contains the columns of \( \mathbf{A} \) indexed by \( \mathcal{T}^{lp} \). Let \( \mathbf{A}^{\dagger}_{\mathcal{T}^{lp}} \) denote its pseudo-inverse. Then the final estimation \( \mathbf{x}^{rel} \) is given by \( \mathbf{x}^{rel} = \mathbf{A}^{\dagger}_{\mathcal{T}^{lp}} \mathbf{y} \), and \( \mathbf{x}^{rel} = \mathbf{x}^{lp} \mathbf{C} \), in which \( \mathcal{T}^{lp} \mathbf{C} \) denotes the complement set of \( \mathcal{T}^{lp} \).

Note that the LP problem has an analytical solution. Writing the \( \infty \) norm constraint explicitly as

\[
    \mathbf{x}^{lp} = \arg \min_{\mathbf{x}} \sum_{i=1}^{n} |x_i| \quad \text{s.t. } |x_i - x_i^{lp}| \leq \lambda, \text{ for } i = 1 \ldots n. \tag{7}
\]

We can see that there are no cross terms in both the objective function and the constraint inequalities, so each component can be optimized separately. From this ob-
servation, the solution of the LP problem is given as
\[ x_i^{lp} = \begin{cases} 
0, & \text{if } |x_i^{ls}| \leq \lambda \\
x_i^{ls} - \lambda, & \text{if } x_i^{ls} > \lambda \\
x_i^{ls} + \lambda, & \text{if } x_i^{ls} < -\lambda 
\end{cases} \]
for \( i = 1, 2, \ldots, n \). Such a solution \( x^{lp} \) is also referred to as an application of the soft-thresholding operation to \( x^{ls} \), see e.g. Donoho and Johnstone (1995). Several remarks related to the algorithm are given as follows.

**Remark 2** Note that the tuning parameter \( \lambda \) chosen as \( \lambda^2 = \frac{2n}{n^2} \) is very similar to the one (which is proportional to \( \frac{2}{n^2} \)) as given in Rosas and Hjalmarsson (2011) based on the Akaike’s Information Criterion (AIC).

**Remark 3** The order of \( \lambda \) chosen as \( -\frac{1}{2} + \frac{2}{n^2} \) is essential to make the asymptotic oracle property hold. Intuitively speaking, such a choice can make the following two facts hold.

1. Whenever \( \epsilon > 0 \), \( x^{lp} \) will lie in the feasible region of Eq. (6) with high probability.
2. The threshold decreases ‘slower’ (in the order of \( N \)) than the variance of the pseudo noise term \( \sqrt{\Sigma^{-1}} U^T v \). With such a choice, it is possible to get a good approximation of the support set of \( x^0 \) in the second step.

**Remark 4** Though the formulation of Eq. (6) is inspired by the Dantzig selector in Candès and Tao (2007), there are some differences between them.

1. As pointed out by one of the reviewer, both the proposed method and the Dantzig selector lie in the following class
   \[
   \min_x \|x\|_1 \text{ s.t. } \|W(x - x^{ls})\|_\infty \leq \lambda. \quad (8)
   \]
   If \( W \) is chosen as the identity matrix, we obtain the proposed method; If \( W \) is chosen as \( A^T A \), then we obtain the same formulation as given by the Dantzig selector.
2. As pointed out in Efron (2007), the solution path of the Dantzig selector behaves erratically with respect to the value of the regularizer parameter. However, the solution path of Eq. (6) with respect to the value of \( \lambda \) behaves regularly, which is due to the fact that, given \( \lambda \), the solution to Eq. (6) is given by the application of the soft-thresholding operation to the LSE estimation. When \( \lambda \) increases, the solution will decrease (or increase) linearly and when it hits zero, it will remain to be zero. This in turn implies computational advances when trying to find a \( s \)-sparse solution for given \( s \). A simple illustration of the solution path is given. Assume that \( n = 4 \) and \( x^{ls} = [2, 0.5, -1, -1.5]^T \), then the solution path to Eq. (6) w.r.t. \( \lambda \) is given as in Fig. (1).

Fig. 1. An illustration of the solution path to Eq. (6) w.r.t. \( \lambda \). When \( \lambda \) equals zero, the solution to Eq. (6) is \( x^{ls} \); when \( \lambda \) increases, the solution trajectory shrinks linearly to zero and then remains zero.

**Remark 5** From a computational point of view, the SCAD method needs to solve a non-convex optimization problem which will suffer from the multiple local minima, see the discussions in Trevor, Hastie, Tibshirani and Friedman (2005). So, the proposed scheme is mainly compared with techniques which can be solved as convex optimization problems. In Table 1, we list the computational steps needed for different methods. In the table, the term ST means the soft-thresholding operation, the term Re-LSE means ‘redo the LSE estimation after detecting the support set of the result obtained from the second step’. For a more precise description, see the Algorithm Description section. From this table, we can see that in the first step, all the methods need to do a LSE estimation; in the second step, except the proposed method (which is denoted by LP + Re-LSE), the other methods need to solve a LASSO optimization problem, which is more computationally involved than a simple soft-thresholding operation as needed by the proposed method; except the ADALASSO method, the other methods need to do a Re-LSE step, which is computationally easier if the sparsity level is low. From this table, we can also see that the main computational burden for the proposed method comes from the LSE (SVD) step.

| Method               | Step 1 | Step 2 | Step 3 |
|----------------------|--------|--------|--------|
| LP + Re-LSE         | LSE    | ST     | Re-LSE |
| ADALASSO            | LSE    | LASSO  | Re-LSE |
| A-SPARSEVA-AIC-RE   | LSE    | LASSO  | Re-LSE |
| A-SPARSEVA-BIC-RE   | LSE    | LASSO  | Re-LSE |

Table 1

Computational steps needed for different methods
Remark 6 Note that the proposed method does not need an "adaptive step" (i.e. to reweigh the cost function) in order to achieve the ORACLE property, which is different from the methods presented in Rojas and Hjalmarsson [2014] and Zou [2004].

3 Analysis of the algorithm

In this section, we will discuss the properties of the presented estimator. In the following, we will use $\sigma$ to denote the smallest singular value of $A$.

Remark 7 In the following sections, we assume that the noise variance equals one, i.e. $c = 1$, for the following reasons:

1. When the noise variance is given in advance, one can always re-scale the problem accordingly.
2. Even if the noise variance is not known explicitly (but is known to be finite), the support of $x^0$ will be recovered asymptotically. This is a direct consequence of the fact that finite, constant scalings do not affect the asymptotic statements, i.e. we can use the same $\lambda$ for any level of variance without influencing the asymptotic behavior.

The following facts (Lemma 1-3) will be needed for subsequent analysis. Since their proofs are standard, we state them without proofs here. Using the notations as introduced before, one has that

Lemma 1 $x^{ls} = x^0 + V \Sigma^{-1}U^T v$.

Lemma 2 $b = \Sigma V^T x^{ls} - \Sigma V^T x^0$ is a Gaussian random vector with distribution $N(0, I)$.

Lemma 3 Given $d > 0$, then

$$\int_{|t|>d} \frac{1}{\sqrt{2\pi}} e^{-t^2} dt \leq e^{-\frac{d^2}{2}}.$$

In the following, we will first analyze the probability that $x^0$ lies in the constraints set of the LP problem given by Eq. (6). Then we give an error estimation of the results given by Eq. (6). After this, we will discuss the capability of recovering the support set of $x^0$ by Eq. (6), which will lead to the asymptotic ORACLE property of the proposed estimator.

Lemma 4 For all $\lambda > 0$, one has that

$$\mathbb{P}\left(\|V^T x^{ls} - V^T x^0\|_\infty > \frac{\lambda}{\sqrt{n}}\right) \leq ne^{-\frac{\lambda^2}{2n}}.$$

Proof By Lemma 2, and noticing that $b = \Sigma V^T x^{ls} - \Sigma V^T x^0$ is a Gaussian random vector with distribution $N(0, I)$, we have that

$$\mathbb{P}\left(\|b\|_\infty > \frac{\lambda\sigma}{\sqrt{n}}\right)$$

$$\leq \mathbb{P}\left(\|\Sigma V^T x^{ls} - \Sigma V^T x^0\|_\infty > \frac{\lambda\sigma}{\sqrt{n}}\right)$$

$$= \mathbb{P}\left(\exists i, such that |b_i| > \frac{\lambda\sigma}{\sqrt{n}}\right)$$

$$\leq \sum_{i=1}^n \mathbb{P}\left(|b_i| > \frac{\lambda\sigma}{\sqrt{n}}\right).$$

Application of Lemma 3 gives the desired result.

Lemma 5 For all $\lambda > 0$, if $\|V^T x^{ls} - V^T x^0\|_\infty \leq \frac{\lambda}{\sqrt{n}}$, then $\|x^{ls} - x^0\|_\infty \leq \lambda$.

Proof Define $c = V^T x^{ls} - V^T x^0$, so we have $\|x^{ls} - x^0\|_\infty = \|Vc\|_\infty$. Analyze the $i$th element of $Vc$ that

$$|V_i c| \leq \|c\|_2 \leq \|c\|_\infty \sqrt{n} \leq \lambda.$$
Given $0 < \epsilon < 1$, and let $\lambda^2 = \frac{2n}{N}$, we have that
$$\mathbb{P} \left( \|x^{ls} - x^0\|_\infty \leq \lambda \right) \geq 1 - ne^{-c_1N\epsilon}.$$ 

Next, we will derive an error bound (in the $l_\infty$ norm) of the estimator given by the LP formulation. Define
$$h = x^{lp} - x^0,$$
as the error vector of LP formulation. We have that the error term $h$ is bounded as follows:

**Lemma 7** For any $\lambda > 0$, if $\|x^{ls} - x^0\|_\infty \leq \lambda$, then we have that $\|h\|_2^2 \leq 4\lambda^2$.

**Proof** We first consider the error vector on the $T^c$ which is given by $h_{T^c}$. From the property of the soft-shrithresholding operation, we have that $\|x^{ls}_{T^c}\|_\infty \leq \lambda$. It follows from the previous discussions that $x^{lp}$ is obtained by application of the soft-shrithresholding operator with the threshold $\lambda$ applied componentwise to $x^{ls}$, hence we obtain that $x^{lp}_{T^c} = 0$. This implies that $h_{T^c} = 0$.

Next we consider the error vector on the support $T$, denoted as $h_T$. From the property of the soft-shrithresholding operation, we have that $\|x^{ls}_T - x^{lp}_T\|_\infty \leq \lambda$. Then we have that $\|x^{ls}_T - x^{lp}_T\|_\infty \leq \|x^{ls}_T - x^{lp}_T\|_2 + \|x^{lp}_T - x^0\|_\infty \leq 2\lambda$

Combining both statements gives that $\|h\|_2^2 = \|h_T\|_2^2 + \|h_{T^c}\|_2^2 \leq |T|\|h_T\|_2^2 \leq 4\lambda^2$. 

Plugging in the $\lambda$ as chosen in previous section, we can get the error bound of the LP formulation. However, the estimate $x^{lp}$ is not the final estimation, instead it will be used to recover the support set of $x^0$. The following theorem states this result formally. For notational convenience, $T^{lp}(N)$ is used to denote the recovered support from the LP formulation, and $x^{rels}(N)$ then denotes the estimate after the second LSE step using $N$ observations. Finally, the vector $x^{ls-or}(N)$ denotes the LSE as if the support of $x^0$ were known (i.e. the ORACLE presents) using $N$ observations.

We will first get a weak support recovery result and based on this, we further prove that the support as recovered by the LP formulation will converge to the true support $T$ almost surely.

**Lemma 8** Given $0 < \epsilon < 1$, and assume that the matrix $A$ has singular values which satisfies Eq. (4), with constants $c_1, c_2$ as given there. Let $x_0 \triangleq \min \{|x_i|, i \in T\} \in \mathbb{R}^+$, and $\lambda^2 = \frac{2n}{N}$, then

(a) $\lim_{N \to \infty} \mathbb{P}(T = T^{lp}(N)) = 1,$

and

(b) $\lim_{N \to \infty} \mathbb{P}(x^{rels}(N) = x^{ls-or}(N)) = 1.$

**Proof** Let the vector $\breve{v}$ denote $\breve{v} = V\Sigma^{-1}U^Tv$. Since $x^{ls} = x^0 + V\Sigma^{-1}U^Tv$, one has that $x^{ls} = x^0 + \breve{v}$, in which $\breve{v}$ follows a normal distribution $N(0, V\Sigma^{-2}V^T)$. Without loss of generality, assume that $\breve{v}_1, \ldots, \breve{v}_N$ are the nonzero elements of $x^0$ and their values are positive. Since $\lambda$ decreases when $N$ increases, so there exist a number $N_1 \in \mathbb{N}$, such that $\lambda < \frac{\lambda}{\sqrt{N}}$ for all $N \geq N_1$. In the following derivations, we use $v_{i,j}$ to denote the element in the $i$th row, $j$th column of $V$ and $\breve{v}_i$ denotes the $i$th element of $\breve{v}$. When $N > N_1$, we have the following bound of $\mathbb{P}(T \neq T^{lp}(N))$:

$$\mathbb{P}(T \neq T^{lp}(N)) \leq \sum_{i=1}^N \mathbb{P}(\lambda - x_0 < \breve{v}_i < \lambda - x_0) + \sum_{i=1}^N \mathbb{P}(|\breve{v}_i| > \lambda) \leq n \mathbb{P}(\lambda - x_0 < \breve{v}_i < \lambda - x_0) + \sum_{i=1}^N \mathbb{P}(\lambda - x_0 < \breve{v}_i < \lambda - x_0) \leq \sum_{i=1}^N \mathbb{P}(\lambda - x_0 < \breve{v}_i < \lambda - x_0) + \sum_{i=1}^N \mathbb{P}(\lambda - x_0 < \breve{v}_i < \lambda - x_0) \leq \sum_{i=1}^N \mathbb{P}(\lambda - x_0 < \breve{v}_i < \lambda - x_0) + \sum_{i=1}^N \mathbb{P}(\lambda - x_0 < \breve{v}_i < \lambda - x_0) \leq 2c_2s\sqrt{n}N\epsilon e^{-\frac{1}{2}(c_1x_0)^2N + N\epsilon - c_1^2N\epsilon},$$

where $C = 2c_2s\sqrt{n}$. The second inequality in the chain holds due to the fact that the probability distribution function of $\breve{v}_i$ is monotonically increasing in the interval $[\lambda - x_0, \lambda - x_0]$, together with results in Lemma 3.

Then we can see that both terms in (9) will tend to 0 as $N \to \infty$ for any fixed $\epsilon > 0$, i.e. $\lim_{N \to \infty} \mathbb{P}(T^p(N) = T) = 1$. For the proof of (b), notice the following relation

$$\mathbb{P}(x^{rels}(N) = x^{ls-or}(N)) = \mathbb{P}(T^{rels}(N) = T).$$

From (a) we know that the right hand side will tend to 0 as $N$ tends to infinity, hence (b) is proven. 

Based on the previous lemma, we have

**Theorem 2** Given $0 < \epsilon < 1$, and assume that the matrix $A$ has singular values which satisfies Eq. (4), with
In the following, we will show that when \( N > N_1 \) that

\[
\mathbb{P}\left( \cap_{N=N_3}^{\infty} \left\{ \mathcal{T}^{lp}(N) = \mathcal{T} \right\} \right) = 1
\]

and

\[
\mathbb{P}\left( \cap_{N=N_3}^{\infty} \left\{ x^{\text{rel}}(N) = x^{\text{LASSO}}(N) \right\} \right) = 1.
\]

**Proof**

From the proof in the previous lemma, we have that when \( N > N_1 \)

\[
\mathbb{P}(\mathcal{T} \neq \mathcal{T}^{lp}(N))
\]

\[
\leq C N e^{-\frac{c_1 x_0^2}{2}} N + N e^{-c_2^2 n N^r}
\]

\[
= C e^{-\frac{c_1 x_0^2}{2} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + 1}
\]

Since \( 0 < \epsilon < 1 \) and \( x_0 > 0 \), one has that \( \frac{\epsilon \ln(N)}{2(c_1 x_0)^2} N \) and \( \frac{\epsilon \ln(N)}{c_1^2 n N^r} \) will tend to zero if \( N \to \infty \). Hence there exists a number \( N_2 \in \mathbb{N} \) such that for all \( N > N_3 \), \( \max(N_1, N_2) \) one has that \( \frac{\epsilon \ln(N)}{2(c_1 x_0)^2} N < \frac{1}{10} \) and \( \frac{\epsilon \ln(N)}{c_1^2 n N^r} < \frac{1}{5} \). Hence

\[
\sum_{N=N_3}^{\infty} \mathbb{P}(\mathcal{T}^{lp}(N) \neq \mathcal{T})
\]

\[
\leq \sum_{N=N_3}^{\infty} \sum_{N=N_3}^{\infty} C e^{-\frac{c_1 x_0^2}{2} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + 1}
\]

\[
\leq \int_{N=N_3-1}^{\infty} C e^{-\frac{c_1 x_0^2}{2} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + 1} dt
\]

\[
= A + B.
\]

Furthermore, it can be seen that

\[
A = \int_{N=N_3-1}^{\infty} C e^{-\frac{c_1 x_0^2}{2} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + 1} dt < \infty.
\]

In the following, we will show that \( B = \int_{N=N_3-1}^{\infty} e^{-\frac{c_1 x_0^2}{2} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + 1} \) By a change of variable using \( x = \frac{1}{c_1^2 n} e^{-\frac{c_1 x_0^2}{2} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + \frac{c_2^2 n N^r}{} + 1} \), we have that

\[
B = \int_{N=N_3-1}^{\infty} x^{\frac{1}{c_1^2 n}} e^{-x} dx < \frac{1}{c_1^2 n} \Gamma \left( \frac{1}{c_1^2 n} \right) < \infty
\]

with \( \Gamma \) the Gamma function. And hence

\[
\sum_{N=N_3}^{\infty} \mathbb{P}(\mathcal{T}^{lp}(N) \neq \mathcal{T}) < \infty.
\]

Application of the Borel-Cantelli lemma [4] implies that the events in \( \{ \mathcal{T} \neq \mathcal{T}^{lp}(N) \}_{N=N_3}^{\infty} \) will not happen infinitely often, i.e., there exists the number \( N' \in \mathbb{N} \) as defined in the assumptions of the theorem, such that \( \{ \mathcal{T} = \mathcal{T}^{lp}(N) \}_{N=N_3}^{\infty} \) will hold. Hence (a) and (b) are proven.

\[\square\]

### 4 Illustrative Experiments

This section supports the findings in previous sections with numerical examples and makes comparisons with other algorithms which possess the ORACLE property in the literature.

#### 4.1 Experiment 1

This example is taken from [Zou (2006)]. The setups are repeated as follows.

- \( x^0 \) is set to be \((3, 1.5, 0, 0, 2, 0, 0, 0)\)\(T\);
- Rows of matrix \( A \) are i.i.d. normal vectors;
- The correlation between the \( j_1 \)-th and the \( j_2 \)-th elements of each row are given as 0.5|\( j_1 - j_2 |\);
- The noise term \( \nu \in \mathbb{R}^N \) follows distribution \( \mathcal{N}(0, I_N) \).

Based on these setups, the proposed method and also the methods presented in [Rojas and Hjalmarsson (2011)] (the A-SPARSEVA-AIC-RE method and the A-SPARSEVA-BIC-RE methods) and [Zou (2006)], (the ADALASSO method) are applied to recover \( x^0 \). In this experiment, \( \epsilon \) for the proposed method is set to \( \frac{1}{4} \); \( \lambda_N \) for 'ADALASSO' is chosen as \( N^{1/2 - \gamma/4} \) (this choice satisfies all the assumptions in Theorem 2 in [Zou (2006)]), and \( \gamma \) is set to 1; the thresholding value (for detecting zero components from the solution of the Lasso problem) for the 'A-SPARSEVA-AIC-RE' and 'A-SPARSEVA-BIC-RE' are set to be \( 10^{-5} \) as suggested in [Rojas and Hjalmarsson (2011)]. For the comparison, we also include the experiment result obtained by using the LASSO method, in which we set the tuning parameter as \( \sqrt{N} \). In Fig. 2, for every \( N \), experiment is repeated 50 times to get the estimated MSE. The following abbreviations are used in Fig. 2: (1) the curve with tag 'LSE' gives the MSE of the estimates by the ordinary least square algorithm; (2) the curve with tag 'LP + RE-LSE' gives the MSE of the estimates given by the proposed algorithm; (3) the curve with tag 'ORACLE-LSE' gives the MSE of the estimates by the ORACLE least square estimation; (4) the curves with tags 'A-SPARSEVA-AIC-RE' and 'A-SPARSEVA-BIC-RE' give the MSE of the estimates by the methods presented in [Rojas and Hjalmarsson (2011)]; (5) the curve with tag 'ADALASSO' gives the MSE of the estimates by the ADALASSO method presented in [Zou (2006)]; (6) the curve with tag 'LASSO' gives the MSE of the estimates of the LASSO method.
Note that, when $N$ becomes large, the curves ‘LP + RE-LSE’ and ‘ORACLE-LSE’ exactly match each other.

Fig. 2. Performance of the different estimators from $N$ observations to estimate $x^0$. This picture indicates that the proposed estimator will give exactly the same performance as the ORACLE estimator for a large $N$ ($N \approx 75$).

Fig. 3 demonstrates the efficacy of support recovery of the LP formulation in Eq. (6) for different choices of $\epsilon$. In the plot, ‘portion’ is defined as the ratio of successful trials over the total number of trials. We conclude the empirical observations for this experiment in the caption of the figure.

In practice, cross validation technique could be exploited to choose the tuning parameters. In the following part, we will apply the 5-fold cross-validation method (see Trevor, Hastie, Tibshirani and Friedman (2003)) to choose the tuning parameters and then compare their performances based on the chosen tuning parameters. The procedure is as follows. At first, the tuning parameter is obtained by 5-fold cross validation, then it is applied to an independently generated test data which has the same dimension as the training data and the evaluation data. For different $N$, we run 100 i.i.d. realizations. In each realization, we record the value $\|\hat{x} - x^0\|_2^2$, where $\hat{x}$ denotes the estimate obtained by the estimator. $\epsilon$ are selected from $\{1/8, 1/4, 1/2\}$, $\gamma$ are selected from $\{1/2, 1, 2\}$, and $N$ are chosen from $\{20, 50, 100, 200, 300, 500\}$. The results are reported in Fig. 4.

4.2 Experiment 2

In this part, we perform an experiment for recovering the sinusoids from noisy measurements. The data is generated as follows:

$$y(t) = \sum_{k'=1}^{n'} c_{ik'} \sin(w_{ik'}t) + v(t).$$

Here both $\{w_{ik'}\}_{k'}$ and $\{c_{ik'}\}_{k'}$ are unknown, but we know that the frequencies do belong to a (larger, but of constant size) set $\{w_k\}_{k=1}^n$ of $n$ elements. By sampling the system with period $t_s$, we obtain the system

$$y = A x^0 + v,$$

where $y = [y(t_s), \ldots, y(Nt_s)]^T$. The matrix $A \in \mathbb{R}^{N \times n}$ is defined as follows. The $i$-th row of $A$ is given by

$$A_i = [\sin(iw_1 t_s), \sin(iw_2 t_s), \ldots, \sin(iw_n t_s)],$$

for $i = 1, \ldots, N$. The parameter term and noise term are defined as $c^0 = [c_1, c_2, \ldots, c_n]^T$, and $v = [v(t_s), v(2t_s), \ldots, v(Nt_s)]^T$.

In this experiment, $n = 10$ and $c^0 = (1, 1, 1, 0, \ldots, 0)^T$, $w_k = k$ for $k = 1, 2, \ldots, n$. We increase $N$ up to 500 and the noise vector $v$ satisfies $v \sim N(0, I_N)$. We also assume that only the first three entries in $\{w_k\}_{k=1}^n$ occur effectively in the system of Eq. (10) and the corresponding amplitudes are set to 1, i.e. $n' = 3$ and $i_1 = 1$, $i_2 = 2$, $i_3 = 3$. The sampling period $t_s$ is set to 0.1s.

The result using the proposed algorithm to recover $x^0$ is displayed in Fig. 5. It is again clear that the proposed
estimator is as efficient as the ORACLE estimator if one has enough samples. That is, from a finite $N$ onwards, the estimator couples tightly with the ORACLE estimator.

and for any $1 \leq i \leq n$ that:

\[
(A^T A)_{i,i} = \sum_{t=1}^{N} (\sin(t w_i t_s))^2 \geq \frac{N}{2} - C_{i,i}.
\]  

The proof is given in Appendix A. With this proposition, an application of Gersgorin circle theorem implies that the eigenvalues of $A^T A$ will increase with the order of $N$, which in turn implies Eq. (4).

5 Conclusion

This note presents an algorithm for solving an overdetermined linear system from noisy observations, specializing to the case where the true 'parameter' vector is sparse. The proposed method does not need one to solve explicitly an optimization problem: it rather requires one to compute twice the LSE step, as well to perform a computationally cheap soft-thresholding step. Also, it is shown formally that the proposed method achieves the ORACLE property. An open question is to quantify how many samples would be sufficient to guarantee exact recovery of $x_0$ for given sparsity level $s$. In this note, we resort to the asymptotic Borel-Cantelli Lemma (‘there exists such a number’), but it is often of interest to have an explicit characterization of this number. Another open question is that how to find a suitable weighting matrix $W$ and how to select the $\lambda$ in Eq. (8), in order to make Eq. (8) a practical improvement of the proposed method which also maintains the ORACLE property.
A Proof of Proposition 4

Proof The proof of (12) goes as follows. First

\[
\left| \sum_{t=1}^{N} \sin(tw_t t_s) \sin(tw_t t_s) \right| \\
= \frac{1}{2} \left| \sum_{t=1}^{N} (\cos(t(w_t - w_j) t_s) - \cos(t(w_i + w_j) t_s)) \right| \\
\leq \frac{1}{2} \left| \sum_{t=1}^{N} \cos(t(w_i - w_j) t_s) \right| + \frac{1}{2} \left| \sum_{t=1}^{N} \cos(t(w_i + w_j) t_s) \right| .
\]

We focus on bounding the term \( \left| \sum_{t=1}^{N} \cos(t(w_i - w_j) t_s) \right| \), the bound of the other term will follow along the same lines.

\[
\left| \sum_{t=1}^{N} \cos(t(w_i - w_j) t_s) \right| \\
= \left| \Re \left( \frac{1 - e^{j(N+1)(w_i - w_j) t_s}}{1 - e^{j(w_i - w_j) t_s}} \right) - 1 \right| \\
\leq \frac{1}{2} \left| \frac{1 - e^{j(N+1)(w_i - w_j) t_s}}{1 - e^{j(w_i - w_j) t_s}} \right| + 1 \\
\leq \frac{2}{1 - e^{j(w_i - w_j) t_s}} + 1,
\]

which is a constant which does not depend on \( N \), so inequality (12) is obtained.

In order to prove inequality (13), observe that

\[
\sum_{t=1}^{N} \left( \sin(tw_t t_s) \right)^2 = \frac{1}{2} \sum_{t=1}^{N} (1 - \cos(2tw_t t_s)) \geq \frac{N}{2} - \frac{1}{2} \left| \sum_{t=1}^{N} \cos(2tw_t t_s) \right| .
\]

Using previous bounding method, \( \frac{1}{2} \left| \sum_{t=1}^{N} \cos(2tw_t t_s) \right| \) is also bounded by a constant \( C_{1,i} \) which does not depend on \( N \). This concludes the proof. \( \Box \)