UPPSALA
UNIVERSITET

IT Licentiate theses
2014-003

On some sparsity related problems and the randomized Kaczmarz algorithm

LIANG DAI

UPPSALA UNIVERSITY
Department of Information Technology

On some sparsity related problems and the randomized Kaczmarz algorithm

Liang Dai
liang.dai@it.uu.se

April 2014

Division of Systems and Control
Department of Information Technology
Uppsala University
Box 337
SE-751 05 Uppsala
Sweden

http://www.it.uu.se/

Dissertation for the degree of Licentiate of Philosophy in Electrical Engineering with Specialization in Signal Processing and System Identification

ISSN 1404-5117
Printed by the Department of Information Technology, Uppsala University, Sweden

Abstract

This thesis studies several problems related to recovery and estimation. Specifically, these problems are about sparsity and low-rankness, and the randomized Kaczmarz algorithm. This thesis includes four papers referred to as Paper I, Paper II, Paper III, and Paper IV.

Paper I considers how to make use of the fact that the solution to an overdetermined system is sparse. This paper presents a three-stage approach to accomplish the task. We show that this strategy, under the assumptions as made in the paper, achieves the oracle property.

In Paper II, a Hankel-matrix completion problem arising in system theory is studied. Specifically, the use of the nuclear norm heuristic for this task is considered. Theoretical justification for the case of a single real pole is given. Results show that for the case of a single real pole, the nuclear norm heuristic succeeds in the matrix completion task. Numerical simulations indicate that this result does not always carry over to the case of two real poles.

Paper III discusses a screening approach for improving the computational performance of the Basis Pursuit De-Noising problem. The key ingredient for this work is to make use of an efficient ellipsoid update algorithm. The results of the experiments show that the proposed scheme can improve the overall time complexity for solving the problem.

Paper IV studies the choice of the probability distribution for implementing the row-projections in the randomized Kaczmarz algorithm. The result proves that a probability distribution resulting in a faster convergence of the algorithm can be found by solving a related Semi-Definite Programming optimization problem.

List of papers

[I] L. Dai and K. Pelckmans, Sparse estimation from noisy observations of an overdetermined linear system, accepted by *Automatica*.

[II] L. Dai and K. Pelckmans, On the nuclear norm heuristic for a Hankel matrix recovery problem, submitted for possible publication.

[III] L. Dai and K. Pelckmans, An ellipsoid based, two-stage screening test for BPDN, *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012.

[IV] L. Dai, M. Soltanalian and K. Pelckmans, On the randomized Kaczmarz algorithm, *IEEE Signal Processing Letters*, 21(3):330-333, 2014.

Contents

1 Introduction **2**

2 Paper summaries **8**

2.1 Paper I 8

2.2 Paper II 9

2.3 Paper III 11

2.4 Paper IV 11

1 Introduction

In many practical applications, one needs to recover some hidden quantities from measurements. For example, the celebrated Shannon's sampling theory [27] is about how to recover a signal from its sampled measurements. If the measurement process is linear, i.e. the problem can be formulated as a system of linear equations, then the signal recovery task is to infer the original signal from the set of linear equations which describes the the measurement process. And usually, the signal of interest has some structural properties. For instance, it is sparse, or it is sparse under the representation with certain basis. The word *sparse* here means that only a small number of the elements in a vector are nonzero. Such recovery task usually is termed the sparse estimation problem. This is an active and wide range research topic. Two recent books dedicated to this topic are [12, 13]. This section will give a very brief introduction to this topic, with a focus on the case when the set of linear equations are underdetermined.

The bold lower case will be used to denote a vector, bold upper case will be used to denote a matrix. $\|\cdot\|_2$ denotes the spectral norm of a matrix as well as the l_2 norm for a given vector. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian probability distribution function with mean μ and variance σ^2. For the other notions and conventions, we will explain them accordingly when encountered. We start with studying the following problem:

Question 1 *Is it possible to solve $\mathbf{x}_0$ from the following set of linear equations*

$$\mathbf{A}\mathbf{x}_0 = \mathbf{b}, \tag{1}$$

in which $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x}_0 \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$, given that $m < n$?

The answer in general is no if no other assumptions are made. The reason is as follows. Let matrix $\mathbf{B} \in \mathbb{R}^{n\times(n-m)}$ be a matrix satisfying $\mathbf{AB} = 0$. Then we have that any vector of the form $\mathbf{x}_0 + \mathbf{Bk}$, $\mathbf{k} \in \mathbb{R}^{n-m}$ will also be a feasible solution to (1), which follows from

$$\mathbf{A}(\mathbf{x}_0 + \mathbf{Bk}) = \mathbf{A}\mathbf{x}_0 + \mathbf{ABk} = \mathbf{b}.$$

Based on this, one further question can be posed as follows:

Question 2 *If additional assumptions on the matrix $\mathbf{A}$ and $\mathbf{x}_0$ are made, is it possible to recover $\mathbf{x}_0$ from the set of linear equations (1)?*

There are many answers to this question. Here we present one possible solution based on the assumptions of the sparsity of the hidden vector $\mathbf{x}_0$ and the spark property of matrix $\mathbf{A}$.

We first introduce the concept of spark [11] for a matrix. Given matrix $\mathbf{A}$, the spark of $\mathbf{A}$ is defined as the minimum number of the linearly dependent columns of $\mathbf{A}$. I.e., it is given by:

$$\text{spark}(\mathbf{A}) \triangleq \min_{\mathbf{x} \neq 0} \|\mathbf{x}\|_0, \;\; \text{s.t.} \quad \mathbf{A}\mathbf{x} = 0,$$

in which $\|\mathbf{x}\|_0$ denotes the number of the nonzero elements in vector $\mathbf{x}$.

In other words, if the spark of $\mathbf{A}$ is greater or equal to $2k+1$ where $k \in \mathbb{N}$, then any $2k$ columns of $\mathbf{A}$ will be linearly independent. Inspired by this intuition, and assuming that we have $\text{spark}(\mathbf{A}) > 2k$ and $\|\mathbf{x}_0\|_0 \leq k$, then we can recover $\mathbf{x}_0$ using the following approach, i.e. finding the sparsest representation of $\mathbf{y}$ under the basis $\mathbf{A}$. Correctness of this approach will be analyzed shortly, see also the discussions in [11].

Theorem 1 *Given $k \in \mathbb{N}$, if spark$(\mathbf{A}) > 2k$ and $\|\mathbf{x}_0\|_0 \leq k$ hold, when solving the following optimization problem:*

$$\begin{aligned} \hat{\mathbf{x}}_0 = & \arg\min_{\mathbf{x} \in \mathbb{R}^n} \; \|\mathbf{x}\|_0 \\ & \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}_0 = \mathbf{b}, \end{aligned} \tag{2}$$

one has that $\hat{\mathbf{x}}_0 = \mathbf{x}_0$.

The reasoning behind Theorem 1 goes as follows. If $\hat{\mathbf{x}}_0$ is different from $\mathbf{x}_0$ and $\|\hat{\mathbf{x}}_0\|_0 \leq k$, $\mathbf{A}\hat{\mathbf{x}}_0 = \mathbf{b}$, it will imply that $\mathbf{A}(\mathbf{x}_0 - \hat{\mathbf{x}}_0) = 0$ holds, in which $\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|_0 \leq 2k$. This leads to the fact that there exist $2k$ (or less than $2k$) columns of matrix $\mathbf{A}$ which are linearly dependent. This contradicts the fact that $\text{spark}(\mathbf{A}) > 2k$.

In [11], a lower bound to the spark of a matrix is given as

$$\text{spark}(\mathbf{A}) \geq 1 + \frac{1}{\mu(\mathbf{A})}, \tag{3}$$

in which the mutual coherence of matrix $\mathbf{A}$ is defined as

$$\mu(\mathbf{A}) = \max_{1 \leq i \neq j \leq n} \frac{|\mathbf{v}_i^T \mathbf{v}_j|}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}, \tag{4}$$

where $\mathbf{v}_i$ denotes the i-th column of matrix $\mathbf{A}$. From (3) and (4), we can see that a lower dependency correlation between different columns of $\mathbf{A}$ will result in a higher spark of $\mathbf{A}$.

Remark 1 *Here, we discuss several examples of matrices regarding their spark properties.*

- *Random matrix*

  When the elements of the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ *are generated identically independently distributed (i.i.d.) according to the standard Normal distribution, then with high probability,* $\mathrm{spark}(\mathbf{A}) = m + 1$ *holds.*

- *Vandermonde matrix*

  The Vandermonde matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ *is defined as* $\mathbf{A}_{i,j} = \lambda_j^{i-1}$*, and* $\lambda_i \neq \lambda_j$ *if* $i \neq j$*. Since any submatrix of* $\mathbf{A}$ *with size* $m \times m$ *is also a Vandermonde matrix, the determinant of the submatrix will be nonzero. I.e. any* m *columns of the matrix are linearly independent. This implies that* $\mathrm{spark}(\mathbf{A}) = m + 1$.

- *Grassmannian Frame*

  When μ *achieves the so-called Welch bound [20], i.e.* $\mu(\mathbf{A}) = \sqrt{\frac{n-m}{m(n-1)}}$*, then the matrix* $\mathbf{A}$ *is called a Grassmannian Frame. For such class of matrices, one has that* $\mathrm{spark}(\mathbf{A}) = m + 1$*. In [25], an iterative projection method is proposed to design such Grassmannian Frames.*

However, it turns out that the problem given in (2) is NP-hard to solve [16], which leads to the following question:

Question 3 *Is it possible to find a computationally tractable way to find the solution of (2) if more assumptions are made?*

To answer this question, the Restricted Isometry Property (RIP) and the l_1 relaxation approach will be introduced. Sufficient conditions for the l_1 relaxation approach based on the RIP property to recover the solution to (2) will also be discussed.

For a given integer k, the Restricted Isometry Constant of matrix $\mathbf{A}$ is defined as

$$\delta_k = \max_{|S| \leq k} \|\mathbf{A}_S^T \mathbf{A}_S - \mathbf{I}\|_2, \tag{5}$$

in which $\mathbf{A}_S$ denotes the submatrix of $\mathbf{A}$ with the columns indexed by the set S, and $\mathbf{I}$ indicates the identity matrix with corresponding size.

Based on δ_k, the following relation for $\mathbf{A}$ is implied by (5)

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_k)\|\mathbf{x}\|_2^2, \text{ for} \forall \mathbf{x} \in \mathbb{R}^n \text{ with } \|\mathbf{x}\|_0 \leq k.$$

This property is termed the Restricted Isometry Property (RIP) of the matrix $\mathbf{A}$ in [4].

Remark 2 *A connection between the δ_{2k} and the spark property of $\mathbf{A}$ is that if $\delta_{2k} < 1$, it holds that $\mathrm{spark}(\mathbf{A}) \geq 2k+1$. From this angle, the RIP property can be regarded as a generalization of the spark property of a matrix.*

As discussed before, the problem in (2) is NP-hard to solve, one way to get around that is to *relax* the non convex $\|\mathbf{x}\|_0$ with its '*convex envelope*' $\|\mathbf{x}\|_1 \triangleq \sum_{i=1}^{n} |x_i|$. Formally, the convex envelope for a function $f(x)$ is defined as the largest function $g(x)$, which is convex and satisfies $g(x) \leq f(x)$. For more discussions about the convex envelope for a non convex function, see chapter five in [14].

By making use of the convex envelope of $\|\mathbf{x}\|_0$, the optimization problem for the l_1 relaxation approach is given as:

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\min_{\mathbf{x}\in\mathbb{R}^n} \ \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{aligned} \tag{6}$$

Different from the approach in (2), the l_1 relaxation approach can be formulated as a Linear Programming (LP) problem, which is computationally tractable. One possible way to write the formulation in (6) as an equivalent LP problem is given as follows.

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\min_{\mathbf{x},\mathbf{t}\in\mathbb{R}^n} \sum_{i=1}^{n} t_i \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & -\mathbf{t} \leq \mathbf{x} \leq \mathbf{t}. \end{aligned} \tag{7}$$

To solve the LP problem, many techniques are applicable. For instances, the interior point method and the simplex method. Note that the l_1 relaxation approach is also termed the Basis Pursuit in the literature, see [8].

Remark 3 *Except the l_1 relaxation approach, the greedy methods, such as the Orthogonal Matching Pursuit (OMP) [21] algorithm, the Iterative Hard Thresholding (IHT) [3] algorithm and the Subspace Pursuit (SP) [9] algorithm are alternatives for solving the problem (2). In each step, these methods refine the estimation iteratively with cheap computations. The greedy methods work well especially when the parameter vector is ultrasparse [26].*

Based on the RIP property, the following result for (6) can be established [4].

Theorem 2 *Suppose that $\|\mathbf{x}_0\|_0 \leq k$, and the matrix $\mathbf{A}$ has the property that $\delta_{2k} < \sqrt{2} - 1$, then by solving (6), it gives that $\hat{\mathbf{x}} = \mathbf{x}_0$.*

Remarkably, when the entries of $\mathbf{A}$ are i.i.d. $\mathcal{N}(0, \frac{1}{m})$ random variables, and

$$m \geq \frac{C}{\delta^2} log(n/k)k + ln(\frac{1}{\epsilon}) \tag{8}$$

holds where C is a universal constant, then $\mathbf{A}$ will have that $\delta_k \leq \delta$ with probability $1 - \epsilon$. For a derivation of the result, see [1].

Note that, for such random matrix, if $m = 2k$, then the spark of $\mathbf{A}$ will be $2k + 1$, which implies that (2) is sufficient to find the k-sparse vector $\mathbf{x}_0$. In order to apply (6) to recover $\mathbf{x}_0$, one needs to have more observations, as indicated by the $log(n/k)$ factor before k in (8).

Remark 4 *By inspecting the reasonings behind Theorem 1, we can find that if for any vector $\mathbf{h}$ satisfying $\mathbf{A}\mathbf{h} = 0$, it holds that $\|\mathbf{h}\|_0 \geq 2k + 1$, then the conclusion of Theorem 1 also holds. This suggests that a suitable characterization of the nullspace of $\mathbf{A}$ can also leads to similar results which are based on the RIP property of $\mathbf{A}$. See for example the results in [29].*

Next, we will discuss how to adapt the formulation (6) to deal with the noisy case, i.e. when (1) is replaced by the following equation

$$\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{e}, \tag{9}$$

in which $\mathbf{e} \in \mathbb{R}^m$ represents the noise term.

Question 4 *How to get a reasonable estimate of $\mathbf{x}_0$ in the noisy case?*

In the following, two formulations which can deal with the noisy case will be discussed, i.e. the Quadratically Constrained Linear Program (QCLP) and the Dantzig selector. The QCLP is given as

$$\begin{aligned} \hat{\mathbf{x}}_q(\epsilon) &= \underset{\mathbf{x} \in \mathbb{R}^n}{\arg\min} \ \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \epsilon, \end{aligned} \tag{10}$$

in which $\epsilon > 0$ is the tuning parameter depending on the size of the noise.

Remark 5 *There exist similar formulations of the QCLP, which is named the Basis Pursuit De-Noising (BPDN), and the Least Absolute Shrinkage and Selection Operator (LASSO) . The BPDN is given as follows:*

$$\hat{\mathbf{x}}_b(\lambda) = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \ \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{11}$$

where $\lambda > 0$ is a noise level dependent tuning parameter.

The LASSO is given as follows:

$$\hat{\mathbf{x}}_l(\lambda) = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq t, \tag{12}$$

where $t > 0$ is a noise level dependent tuning parameter.

Note that, for a given $\epsilon > 0$ in (10), there exist a corresponding $\lambda \geq 0$ for (11) and a corresponding $t \geq 0$ for (12), such that these formulations return the same optimizer, i.e. $\hat{\mathbf{x}}_q = \hat{\mathbf{x}}_l = \hat{\mathbf{x}}_b$, see [24].

The performance result of (10) is stated in the following theorem, see e.g. [6].

Theorem 3 *If $\mathbf{A}$ satisfies the RIP property with $\delta_{2k} \leq \sqrt{2} - 1$, $\|\mathbf{x}_0\|_0 = k$, and the entries of $\mathbf{e}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, then it holds that*

$$\|\hat{\mathbf{x}}_q(2\sqrt{m}\sigma) - \mathbf{x}_0\|_2 \leq \frac{8\sqrt{1+\delta_{2k}}\sqrt{m}\sigma}{1-(2+\sqrt{2})\delta_{2k}}, \tag{13}$$

with probability larger than $1 - \exp(-c_0 m)$, in which $c_0 > 0$ is a constant.

The formulation known as the Dantzig selector was suggested in [7] for recovering a sparse signal in the noisy case. The Dantzig selector is given as follows

$$\begin{aligned} \hat{\mathbf{x}}_d(\lambda) &= \arg\min_{\mathbf{x}\in\mathbb{R}^n} \ \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \|\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})\|_\infty \leq \lambda, \end{aligned} \tag{14}$$

where $\lambda \geq 0$ is a noise level dependent tuning parameter. Different from the QCLP, the Dantzig selector can be reformulated as a Linear Programming problem using the same idea as transforming (6) into (7).

A similar result on the performance of (14) is summarized as follows, see e.g. [7]:

Theorem 4 *If $\mathbf{A}$ satisfies the RIP property with $\delta_{2k} \leq \sqrt{2} - 1$, $\|\mathbf{x}_0\|_0 = k$, and the entries of $\mathbf{e}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, then one has that*

$$\|\hat{\mathbf{x}}_d(2\sqrt{\log(n)}\sigma) - \mathbf{x}_0\|_2 \leq \frac{4\sqrt{2}\sqrt{1+\delta_{2k}}}{1-(2+\sqrt{2})\delta_{2k}}\sqrt{k\log(n)}\sigma, \tag{15}$$

with probability larger than $1 - \frac{1}{n}$.

Remark 6 *We make the following two remarks:*

1. *Comparing the performance bounds given by QCLP and Dantzig selector, when both m and n are fixed, (15) will give a tighter bound than (13) when k is small.*
2. *Notice that, if the true support set of $\mathbf{x}_0$ is known, the least square estimator will give the estimation of $\mathbf{x}_0$ with error $\|\mathbf{x} - \mathbf{x}_0\|_2^2$ of the order $k\sigma^2$. In [2], it is proven that the Cramér-Rao bound for estimating $\mathbf{x}_0$ is of order $k\sigma^2$. In the case when the support of $\mathbf{x}_0$ is unknown, the recovery error in (15) given by the Dantzig selector is amplified by an extra $\log(n)$ term. Such a property is termed near oracle property of the Dantzig selector in [7].*

For more discussions and comparisons between these two methods, please see [7].

2 Paper summaries

In the following descriptions, we will follow the notations used in the previous section.

2.1 Paper I

Paper 1 presents an approach for the estimation of a sparse vector $\mathbf{x}_0 \in \mathbb{R}^n$ from linear observations which are perturbed by Gaussian noise. The basic idea of the approach is to make use of the Least Squares estimator, while exploiting the sparsity information of $\mathbf{x}_0$. The observed signal $\mathbf{b} \in \mathbb{R}^m$ obeys the following system:

$$\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{e}, \tag{16}$$

and also $m > n$ is assumed. The method consists of the following steps:

1. A classical Least Squares Estimation (LSE);

2. The support is recovered through a Linear Programming optimization problem;

3. A de-biasing step using a LSE on the estimated support set.

It turns out that the LP problem in the second step can be implemented by a soft thresholding operation. For a given value $x \in \mathbb{R}$, when the soft thresholding operation with threshold λ is applied, the obtained vector x_{th} is given as

$$x_{th} = \begin{cases} 0, & \text{if } |x| \leq \lambda \\ x - \lambda, & \text{if } x > \lambda \\ x + \lambda, & \text{if } x < -\lambda \end{cases} .$$

Note that x_{th} can also be obtained by solving the following two optimization problems:

$$x_{th} = \arg\min_{y \in \mathbb{R}} \frac{1}{2}(y - x)^2 + \lambda |y|, \tag{17}$$

and

$$x_{th} = \arg\min_{y \in \mathbb{R}} |y| \quad \text{s.t.} \quad |y - x| \leq \lambda. \tag{18}$$

In this work, the soft thresholding operation is generated by (18). Remark that the formulation (17) is a key ingredient for the derivation of the iterative soft thresholding method [10, 18] and the coordinate descent method [28, 15] for solving the BPDN problem.

The main result of this work is summarized in Theorem 2, which says that, when the number of the observed signal increases, the estimator is able to detect the support of the *true* parameters almost surely, i.e.

$$\mathbb{P}\left(\cap_{m=n'}^{\infty} \{\mathcal{T}^{lp}(m) = \mathcal{T}\} \right) = 1,$$

in which n' is a fixed finite number, $\mathcal{T}^{lp}(m)$ is the estimated support in the second step when the number of observations is m, and $\mathcal{T}$ denotes the true support set of $\mathbf{x}_0$.

2.2 Paper II

This paper studies the completion of a Hankel matrix related with the system theory, by making use of the nuclear norm heuristic. Similar to (2), a rank minimization problem is formulated:

$$\begin{aligned} &\min_{\mathbf{A}\in\mathbb{R}^{m\times n}} \operatorname{rank}(\mathbf{A}) \\ &\text{s.t. } \mathcal{R}\left(\mathbf{A}\right) = \mathbf{b}, \end{aligned} \tag{19}$$

where the linear map $\mathcal{R} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ and the vector $\mathbf{b} \in \mathbb{R}^p$ are given.

In the same spirit of (6), $\operatorname{rank}(\mathbf{A})$ is replaced by its convex envelope function (termed the nuclear norm) $\|\mathbf{A}\|_* \triangleq \sum_{i=1}^{k} \sigma_i$, with $\{\sigma_i\}_{i=1}^{k}$ denoting the singular values of $\mathbf{A}$. Then, the nuclear norm minimization problem, i.e. the relaxation of (19) is given as follows:

$$\begin{aligned} &\min_{\mathbf{A}\in\mathbb{R}^{m\times n}} \|\mathbf{A}\|_* \\ &\text{s.t. } \mathcal{R}\left(\mathbf{A}\right) = \mathbf{b}. \end{aligned} \tag{20}$$

In this paper, the authors study whether the nuclear norm heuristic can recover an impulse response generated by a stable linear system, if elements of the upper-triangle of the associated Hankel matrix were given.

For the case of a single real pole, the result is as follows. Given $-1 < h < 1$, which is the real pole of the system, define the truncated impulse response vector $\mathbf{h} \in \mathbb{R}^n$ as $\mathbf{h} = [1, h, h^2, \ldots, h^{n-1}]^T$, and the associated Hankel matrix as $G_0 = \mathbf{h}\mathbf{h}^T$. It is evident that G_0 is of rank one. Then the nuclear norm heuristic to recover the lower triangle part of the related Hankel matrix is applied as follows.

$$\begin{aligned} \hat{G}_0 \triangleq\, &\arg\min_{G\in\mathbb{R}^{n\times n}} \|G\|_* \\ \text{s.t.} \quad & G(i,j) = G_0(i,j), \forall\ (i+j) \leq n+1, \\ & G \text{ is Hankel.} \end{aligned} \tag{21}$$

Then it gives that $\hat{G}_0 = G_0$, i.e. (21) reconstructs G_0.

The conventional way to prove matrix recovery (or completion) result is by building a certain certificate, for instance the way used in [5] and the iterative golfing scheme advocated by [17]. These approaches all need stochastic assumptions and matrix concentration inequalities [22] in order to construct the desired certificate.

Since the setting is deterministic and the hidden matrix G_0 is a structured matrix, those techniques are not applicable to this situation. The key challenge for the proof is to construct the certificate by exploring the structural information of G_0.

Experimental illustrations of slightly more complicated case, i.e. the case of two real poles, are also conducted. We observe that the nuclear norm heuristic (21) will not always succeed in completing the associated Hankel matrix in this case.

2.3 Paper III

This paper considers a preprocessing stage when solving the BPDN problem, i.e. the *screening test*. More precisely, the test tries to identify the elements of $\hat{\mathbf{x}}_b$ which are equal to zero with a small computational cost before actually solving the entire optimization problem. When those elements are identified, by screening them out, the optimization problem will be reduced to a lower dimensional optimization problem which gives the same solution to the original problem.

One may wonder how it is possible to do the screening test. This could be motivated by considering the following special case of the solution to the BPDN. That is, for the problem described in (11)

$$\hat{\mathbf{x}}_b(\lambda) = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min}\ \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{22}$$

it can be shown [23] that if $\lambda > \max_i|\mathbf{A}^T\mathbf{y}|$ holds, then one has that $\hat{\mathbf{x}}_b(\lambda) = 0$, which means that all the variables are shrunken out by cheap calculations (basically the vector inner products). Despite the fact that this example is special, it gives an intuition of why the screening idea is useful and possible.

This paper then presents an approach to implement the screening test. The main idea of the proposed method is to make an ellipsoid approximation of the feasible region of the related dual problem. The benefit of the ellipsoid approximation is that by doing so, an efficient ellipsoid update rule can be applied to shrink the ellipsoid. Such ellipsoid algorithm is inspired by the ideas from its application in membership set identification [19].

A comparative experiment indicates that, by making use of the proposed scheme, a smaller overall time complexity (including the time for the screening test and the time for solving the reduced size optimization problem) can be achieved compared to other known screening tests.

2.4 Paper IV

The Randomized Kaczmarz Algorithm (RKA) [30] is a method which solves a system of consistent overdetermined linear equations. That is, solve $\mathbf{x}_0 \in \mathbb{R}^n$

from:

$$\mathbf{A}\mathbf{x}_0 = \mathbf{b}, \tag{23}$$

where matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$ is of full column rank, and $\mathbf{b} \in \mathbb{R}^m$. Let $\mathbf{a}_i^T$ denotes the i-th row of matrix $\mathbf{A}$, and b_i is the i-th element of $\mathbf{b}$. Given the initial estimate $\mathbf{x}^0$, the process of the RKA reads as

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \frac{b_i - \mathbf{a}_{i(k)}^T \mathbf{x}^k}{\|\mathbf{a}_{i(k)}\|_2^2} \mathbf{a}_{i(k)} \tag{24}$$

for $k = 1, 2\cdots$, where $i(k)$ is chosen randomly, such that $i(k) = j$ with probability $\frac{\|\mathbf{a}_j\|_2^2}{\|\mathbf{A}\|_F^2}$. The following convergence result was established in [30]:

$$\mathbb{E}(\|\mathbf{x}^k - \mathbf{x}_0\|_2^2) \leq \left(1 - \frac{1}{\|\mathbf{A}\|_F \|\mathbf{A}^\dagger\|_2}\right)^k \|\mathbf{x}^0 - \mathbf{x}_0\|_2^2, \tag{25}$$

in which $\mathbb{E}$ takes the expectation with respect to the random choices of $\{i(l)\}_{l=1}^k$.

As discussed in the literature, whether the probability as suggested in [30] is the optimal choice is unknown. This paper shows that it is possible to find a better probability distribution for the RKA.

The key idea is to derive a tight upper bound to the convergence rate of the RKA, and then this upper bound is optimized. It turns out that optimizing the upper bound leads to a Semi-Definite Programming (SDP) problem - which is a convex and hence computationally tractable. As indicated by Theorem 3 in the paper, optimizing $\|\mathbf{A}\|_F \|\mathbf{A}^\dagger\|_2$ in (25) with respect to the row norms of $\mathbf{A}$ will lead to the same SDP problem. Conversely, this gives that a probability distribution resulting in a faster convergence for the RKA than the one suggested in [30] can always be found using this approach.

Remark that: 1) Solving the resulting SDP problem could be more time consuming than solving the original system of linear equations. Considering this, we spent one section in the paper to suggest two ways to approximate the SDP problem, which can be solved with less computational cost; 2) The resulting SDP formulation is closely related with the SDP formulations arising in the optimal input design problems [31].

References

[1] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] Zvika Ben-Haim and Yonina C. Eldar. The Cramér-Rao bound for estimating a sparse parameter vector. *IEEE Transactions on Signal Processing*, 58(6):3384–3389, 2010.

[3] Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[4] Emmanuel J. Candés. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[5] Emmanuel J. Candés and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[6] Emmanuel J. Candés, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[7] Emmanuel J. Candés and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[8] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[9] Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

[10] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11):1413–1457, 2004.

[11] David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via $l1$ minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[12] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

[13] Yonina C. Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[14] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

[15] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[16] Dongdong Ge, Xiaoye Jiang, and Yinyu Ye. A note on the complexity of l_p minimization. *Mathematical programming*, 129(2):285–299, 2011.

[17] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[18] Elaine T. Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for l_1 minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

[19] Robert L. Kosut, Ming K Lau, and Stephen P. Boyd. Set-membership identification of systems with parametric and nonparametric uncertainty. *IEEE Transactions on Automatic Control*, 37(7):929–941, 1992.

[20] Welch, L. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3): 397-399, 1974.

[21] Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[22] Pascal Massart and Jean Picard. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

[23] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.

[24] R. Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[25] Joel A. Tropp, Inderjit S. Dhillon, Robert W. Heath, and Thomas Ströhmer. Designing structured tight frames via an alternating projection method. *IEEE Transactions on Information Theory*, 51(1):188–209, 2005.

[26] Joel A. Tropp and Stephen J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

[27] Michael Unser. Sampling-50 years after shannon. *Proceedings of the IEEE*, 88(4):569–587, 2000.

[28] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[29] Yin Zhang. Theory of compressive sensing via l_1 minimization: a non-RIP analysis and extensions. *Journal of the Operations Research Society of China*, 1(1):79–105, 2013.

[30] Strohmer, Thomas and Vershynin, Roman. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[31] Yaming Yu. Monotonic convergence of a general algorithm for computing optimal designs. *The Annals of Statistics*, 38(3): 1593-1606, 2010.

Paper I

Sparse Estimation From Noisy Observations of an Overdetermined Linear System

Liang Dai [a] and Kristiaan Pelckmans [a]

[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 $\mathbf{y} \in \mathbb{R}^N$ obeys the following system:

$$\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathbf{v}, \tag{1}$$

where the elements of the vector $\mathbf{x}^0 \in \mathbb{R}^n$ are considered to be the fixed but unknown parameters of the system. Moreover, it is assumed that $\mathbf{x}_0$ 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 $\mathbf{x}^0$ (i.e. $\mathbf{x}_i^0 = 0 \Leftrightarrow i \notin \mathcal{T}$). Let $\mathcal{T}^c$ be the complement of $\mathcal{T}$, i.e. $\mathcal{T} \bigcup \mathcal{T}^c = \{1, 2, \cdots, n\}$ and $\mathcal{T} \bigcap \mathcal{T}^c = \emptyset$. The elements of the vector $\mathbf{v} \in \mathbb{R}^N$ are assumed to follow the following distribution

$$\mathbf{v} \sim \mathcal{N}(0, cI_N), \tag{2}$$

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 $\mathbf{A} \in \mathbb{R}^{N \times n}$ with $N > n$ is a deterministic 'sensing' or 'regressor' matrix. Such a setting ($\mathbf{A}$ 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); Candés and Wakin (2008).

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

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{3}$$

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

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

$\mathbf{A} \in \mathbb{R}^{N\times n}$ *obey*

$$c_1\sqrt{N} \leq \sigma_1(\mathbf{A}) \leq \sigma_2(\mathbf{A}) \leq \ldots \leq \sigma_n(\mathbf{A}) \leq c_2\sqrt{N}, \quad (4)$$

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 (2006), 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:

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{A}^T\mathbf{A} = \mathbf{D} \succ 0. \quad (5)$$

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 $\frac{1}{\sqrt{N}}\mathbf{A}$ 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 $\mathbf{x}^0$ is the LSE under certain assumptions on the noise term. For the Gauss-Markov theorem, please refer to Plackett (1950). However, the normal LSE does not utilize the 'sparse' information of $\mathbf{x}^0$, which raises the question that whether it is possible to improve on the normal LSE by exploiting this information. In the literature, several approaches have been suggested, which can perform as if the true support set of $\mathbf{x}^0$ were known. Such property is termed as the ORACLE property in Fan and Li (2001). In Fan and Li (2001), the SCAD (Smoothly Clipped Absolute Deviation) estimator is presented, which turns out to solve a non-convex optimization problem; later in Zou (2006), the ADALASSO (Adaptive Least Absolute Shrinkage and Selection Operator) estimator is presented. The ADALASSO estimator consists of two steps, which implements a normal LSE in the first step, and then solves a reweighed Lasso optimization problem, which is convex. Recently, in Rojas and Hjalmarsson (2011), two LASSO-based estimators, namely the 'A-SPARSEVA-AIC-RE' method and the 'A-SPARSEVA-BIC-RE' method, are suggested. Both methods need to do the LSE in the first step, then solve a Lasso optimization problem, and finally redo the LSE estimation.

Remark 1 *This note concerns the case that $\mathbf{x}^0$ is a fixed sparse vector. However, when sparse estimators are applied to estimate non-sparse vectors, erratic phenomena could happen. For details, please see the discussions in Leeb and Pötscher (2008); Kale (1985).*

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 $\mathbf{x}^{ls} = \mathbf{V}\Sigma^{-1}\mathbf{U}^T\mathbf{y} = \mathbf{A}^\dagger\mathbf{y}$.
- *LP:* Choose $0 < \epsilon < 1$ and solve the following Linear Programming problem:

  $$\mathbf{x}^{lp} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{x} - \mathbf{x}^{ls}\|_\infty \leq \lambda, \quad (6)$$

  where $\lambda = \sqrt{\frac{2n}{N^{1-\epsilon}}}$. 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 $\mathbf{x}^{rels}$ is given by $\mathbf{x}^{rels}_{\mathcal{T}^{lp}} = \mathbf{A}^\dagger_{\mathcal{T}^{lp}}\mathbf{y}$, and $\mathbf{x}^{rels}_{\mathcal{T}^{lp C}} = \mathbf{0}$, in which ${\mathcal{T}^{lp}}^C$ denotes the complement set of $\mathcal{T}^{lp}$.

Note that the LP problem has an analytical solution. Writing the ∞ norm constraint explicitly as

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

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 observation, the solution of the LP problem is given as

$$x_i^{lp} = \begin{cases} 0, & \text{if } |x_i^{ls}| \le \lambda \\ x_i^{ls} - \lambda, & \text{if } x_i^{ls} > \lambda \\ x_i^{ls} + \lambda, & \text{if } x_i^{ls} < -\lambda \end{cases}$$

for $i = 1, 2, \cdots, n$. Such a solution $\mathbf{x}^{lp}$ is also referred to as an application of the soft-thresholding operation to $\mathbf{x}^{ls}$, see e.g. Donoho and Johnstone (1995). Several remarks related to the algorithm are given as follows.

Remark 2 *Note that the tuning parameter λ chosen as $\lambda^2 = \frac{2n}{N^{1-\epsilon}}$ is very similar to the one (which is proportional to $\frac{2n}{N}$) as given in Rojas and Hjalmarsson (2011) based on the Akaike's Information Criterion (AIC).*

Remark 3 *The* order *of λ chosen as $-\frac{1}{2} + \frac{\epsilon}{2}$ is essential to make the asymptotical oracle property hold. Intuitively speaking, such a choice can make the following two facts hold.*

(1) Whenever $\epsilon > 0$, $\mathbf{x}^0$ 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 $\mathbf{V}\Sigma^{-1}\mathbf{U}^T\mathbf{v}$. With such a choice, it is possible to get a good approximation of the support set of $\mathbf{x}^0$ in the second step.

Remark 4 *Though the formulation of Eq. (6) is inspired by the Dantzig selector in Candés 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

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \ s.t. \ \ \|\mathbf{W}(\mathbf{x} - \mathbf{x}^{ls})\|_\infty \le \lambda. \qquad (8)$$

If $\mathbf{W}$ is chosen as the identity matrix, we obtain the proposed method; If $\mathbf{W}$ is chosen as $\mathbf{A}^T\mathbf{A}$, 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 regularization parameter. However, the solution path of Eq. (6) with respect to the value of λ behaves regularly, which is due to the fact that, given λ, the solution to Eq. (6) is given by the application of the soft-thresholding operation to the LSE estimation. When λ 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 $\mathbf{x}^{ls} = [2, 0.5, -1, -1.5]^T$, then the solution path to Eq. (6) w.r.t. λ is given as in Fig. (1).

Table 1
Computational steps needed for different methods

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

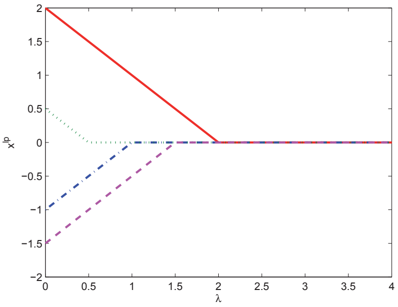


Fig. 1. An illustration of the solution path to Eq. (6) w.r.t. λ. When λ equals zero, the solution to Eq. (6) is $\mathbf{x}^{ls}$; when λ increases, the solution trajectory shrinks linearly to zero and then remains zero.

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.*

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 (2011) and Zou (2006).*

3 Analysis of the algorithm

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

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

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

Lemma 3 *Given $d > 0$, then*

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

In the following, we will first analyze the probability that $\mathbf{x}^0$ 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 $\mathbf{x}^0$ by Eq. (6), which will lead to the asymptotic ORACLE property of the proposed estimator.

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

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

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

$$\begin{aligned} &\mathbb{P}\left(\|\mathbf{V}^T\mathbf{x}^{ls} - \mathbf{V}^T\mathbf{x}^0\|_\infty > \frac{\lambda}{\sqrt{n}}\right) \\ &\quad \leq \mathbb{P}\left(\|\Sigma\mathbf{V}^T\mathbf{x}^{ls} - \Sigma\mathbf{V}^T\mathbf{x}^0\|_\infty > \frac{\lambda\sigma}{\sqrt{n}}\right) \\ &\quad = \mathbb{P}\left(\|\mathbf{b}\|_\infty > \frac{\lambda\sigma}{\sqrt{n}}\right) \\ &\quad = \mathbb{P}\left(\exists i, such\ \ that\ \ |b_i| > \frac{\lambda\sigma}{\sqrt{n}}\right) \\ &\quad \leq \sum_{i=1}^{i=n} \mathbb{P}\left(|b_i| > \frac{\lambda\sigma}{\sqrt{n}}\right). \end{aligned}$$

Application of Lemma 3 gives the desired result. □

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

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

$$|\mathbf{V}_i\mathbf{c}| \leq \|\mathbf{c}\|_2 \leq \|\mathbf{c}\|_\infty\sqrt{n} \leq \lambda.$$

The first inequality is by definition, the second inequality comes from the Cauchy inequality, the last inequality is due to the assumption of the lemma. □

Combining the previous two lemmas gives

Lemma 6 $\mathbb{P}(\|\mathbf{x}^{ls} - \mathbf{x}^0\|_\infty \leq \lambda) \geq 1 - ne^{-\frac{\lambda^2\sigma^2}{2n}}$.

Proof The proof goes as follows

$$\begin{aligned} &\mathbb{P}\left(\|\mathbf{x}^{ls} - \mathbf{x}^0\|_\infty \leq \lambda\right) \\ &\quad \geq \mathbb{P}\left(\|\mathbf{V}^T\mathbf{x}^{ls} - \mathbf{V}^T\mathbf{x}^0\|_\infty \leq \frac{\lambda}{\sqrt{n}}\right) \\ &\quad = 1 - \mathbb{P}\left(\|\mathbf{V}^T\mathbf{x}^{ls} - \mathbf{V}^T\mathbf{x}^0\|_\infty > \frac{\lambda}{\sqrt{n}}\right) \\ &\quad \geq 1 - ne^{-\frac{\lambda^2\sigma^2}{2n}} \end{aligned}$$

The first inequality comes from Lemma 5, and the second inequality follows from Lemma 4. □

The above lemma tells us that $\mathbf{x}^0$ will lie inside the feasible set of the LP problem as given in Eq. (6) with high probability. By a proper choice of λ, the following result is concluded.

Theorem 1 *Given $0 < \epsilon < 1$, and let $\lambda^2 = \frac{2n}{N^{1-\epsilon}}$, we have that*

$$\mathbb{P}\left(\|\mathbf{x}^{ls} - \mathbf{x}^0\|_\infty \leq \lambda\right) \geq 1 - ne^{-c_1^2 N^\epsilon}.$$

Next, we will derive an error bound (in the l_2- norm) of the estimator given by the LP formulation. Define

$$\mathbf{h} = \mathbf{x}^{lp} - \mathbf{x}^0,$$

as the error vector of LP formulation. We have that the error term $\mathbf{h}$ is bounded as follows:

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

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

Next we consider the error vector on the support $\mathcal{T}$, denoted as $\mathbf{h}_{\mathcal{T}}$. From the property of the soft-thresholding operation, it follows that $\|\mathbf{x}^{ls}_{\mathcal{T}} - \mathbf{x}^{lp}_{\mathcal{T}}\|_\infty \leq \lambda$. Then we have that $\|\mathbf{x}^0_{\mathcal{T}} - \mathbf{x}^{lp}_{\mathcal{T}}\|_\infty \leq \|\mathbf{x}^{ls}_{\mathcal{T}} - \mathbf{x}^{lp}_{\mathcal{T}}\|_\infty + \|\mathbf{x}^{ls}_{\mathcal{T}} - \mathbf{x}^0_{\mathcal{T}}\|_\infty \leq 2\lambda$

Combining both statements gives that $\|\mathbf{h}\|_2^2 = \|\mathbf{h}_{\mathcal{T}}\|_2^2 + \|\mathbf{h}_{\mathcal{T}^c}\|_2^2 \leq |T|\|\mathbf{h}_{\mathcal{T}}\|_\infty^2 \leq 4s\lambda^2$. □

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

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

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

and

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

Proof Let the vector $\bar{\mathbf{v}}$ denote $\bar{\mathbf{v}} = \mathbf{V}\Sigma^{-1}\mathbf{U}^T\mathbf{v}$. Since $\mathbf{x}^{ls} = \mathbf{x}^0 + \mathbf{V}\Sigma^{-1}\mathbf{U}^T\mathbf{v}$, one has that $\mathbf{x}^{ls} = \mathbf{x}^0 + \bar{\mathbf{v}}$, in which $\bar{\mathbf{v}}$ follows a normal distribution $\mathcal{N}(0, \mathbf{V}\Sigma^{-2}\mathbf{V}^T)$. Without loss of generality, assume that $x_1^0, x_2^0, \ldots, x_s^0$ are the nonzero elements of $\mathbf{x}^0$ and their values are positive. Since λ decreases when N increases, so there exist a number $N_1 \in \mathbb{N}$, such that $\lambda < \frac{x_0}{2}$ for all $N \geq N_1$. In the following derivations, we use $v_{i,j}$ to denote the element in the ith row, jth column of $\mathbf{V}$ and $\bar{v}_i$ denotes the ith element of $\bar{\mathbf{v}}$. When $N > N_1$, we have the following bound of $\mathbb{P}(\mathcal{T} \neq \mathcal{T}^{lp}(N))$:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{T} \neq \mathcal{T}^{lp}(N)\right) \\
&= \mathbb{P}\left(|x_1^0 + \bar{v}_1| < \lambda, or\ \ |x_2^0 + \bar{v}_2| < \lambda, \ldots, or\ \ |x_s^0 + \bar{v}_s| < \lambda;\right. \\
&\qquad\qquad \left. or\ \ |\bar{v}_{s+1}| > \lambda, or\ \ |\bar{v}_{s+2}| > \lambda, \ldots, or\ \ |\bar{v}_N| > \lambda\right) \\
&\leq \sum_{i=1}^{s} \mathbb{P}(-\lambda - x_i^0 < \bar{v}_i < \lambda - x_i^0) + \sum_{i=s+1}^{N} \mathbb{P}(|\bar{v}_i| > \lambda) \\
&\leq \sum_{i=1}^{s} \frac{2\lambda e^{-(2\sum_{j=1}^n \sigma_j^{-2} v_{ij}^2)^{-1}(-x_i^0+\lambda)^2}}{\sqrt{2\pi(\sum_{j=1}^n \sigma_j^{-2} v_{ij}^2)}} \\
&\qquad\qquad + \sum_{i=s+1}^{N} e^{-(2\sum_{j=1}^n \sigma_j^{-2} v_{ij}^2)^{-1}\lambda^2} \\
&\leq \sum_{i=1}^{s} \frac{2c_2\sqrt{N}\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}c_1^2 N(-x_i^0+\lambda)^2} + \sum_{i=s+1}^{N} e^{-\frac{1}{2}c_1^2 N\lambda^2} \\
&\leq 2c_2 s\sqrt{n} N^{\frac{\epsilon}{2}} e^{-\frac{1}{8}(c_1 x_0)^2 N} + Ne^{-c_1^2 nN^\epsilon} \\
&= CN^{\frac{\epsilon}{2}} e^{-\frac{1}{8}(c_1 x_0)^2 N} + Ne^{-c_1^2 nN^\epsilon}, \qquad (9)
\end{aligned}$$

where $C = 2c_2 s\sqrt{n}$. The second inequality in the chain holds due to the fact that the probability distribution function of $\bar{v}_i$ is monotonically increasing in the interval $[-\lambda - x_i^0, \lambda - x_i^0]$, 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. $\lim_{N\to\infty} \mathbb{P}(\mathcal{T}^{lp}(N) = \mathcal{T}) = 1$. For the proof of (b), notice the following relation

$$\mathbb{P}\left(\mathbf{x}^{rels}(N) = \mathbf{x}^{ls-or}(N)\right) \geq \mathbb{P}\left(\mathcal{T}^{lp}(N) = \mathcal{T}\right).$$

From (a) we know that the right hand side will tend to 1 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 $\mathbf{A}$ has singular values which satisfies Eq. (4), with* *constants c_1, c_2 as given there. Let $x_0 \triangleq \min\{|x_i^0|, i \in \mathcal{T}\} \in \mathbb{R}^+$, and $\lambda^2 = \frac{2n}{N^{1-\epsilon}}$, then there exists a finite number $N' \in \mathbb{N}$, such that*

$$(a): \mathbb{P}\left(\cap_{N=N'}^{\infty}\{\mathcal{T}^{lp}(N) = \mathcal{T}\}\right) = 1$$

and

$$(b): \mathbb{P}\left(\cap_{N=N'}^{\infty}\{\mathbf{x}^{rels}(N) = \mathbf{x}^{ls-or}(N)\}\right) = 1.$$

Proof From the proof in the previous lemma, we have that when $N > N_1$

$$\begin{aligned}
&\mathbb{P}(\mathcal{T} \neq \mathcal{T}^{lp}(N)) \\
&\leq CN^{\frac{\epsilon}{2}} e^{-\frac{1}{8}(c_1x_0)^2N} + Ne^{-c_1^2nN^\epsilon} \\
&= Ce^{-\frac{1}{8}(c_1x_0)^2N + \frac{\epsilon}{2}ln(N)} + e^{\ln(N) - c_1^2nN^\epsilon} \\
&= Ce^{(c_1x_0)^2N(\frac{\epsilon ln(N)}{2(c_1x_0)^2N} - \frac{1}{8})} + e^{c_1^2nN^\epsilon(\frac{\ln(N)}{c_1^2nN^\epsilon} - 1)}.
\end{aligned}$$

Since $0 < \epsilon < 1$ and $x_0 > 0$, one has that $\frac{\epsilon ln(N)}{2(c_1x_0)^2N}$ and $\frac{\ln(N)}{c_1^2nN^\epsilon}$ will tend to zero if $N \to \infty$. Hence there exists a number $N_2 \in \mathbb{N}$ such that for all $N > N_3 \triangleq \max(N_1, N_2)$ one has that $\frac{\epsilon ln(N)}{2(c_1x_0)^2N} < \frac{1}{16}$ and $\frac{\ln(N)}{c_1^2nN^\epsilon} < \frac{1}{2}$. Hence

$$\begin{aligned}
&\sum_{N=N_3}^{\infty} \mathbb{P}(\mathcal{T}^{lp}(N) \neq \mathcal{T}) \\
&\leq \sum_{N=N_3}^{\infty} Ce^{-\frac{1}{16}(c_1x_0)^2N} + \sum_{N=N_3}^{\infty} e^{-\frac{1}{2}c_1^2nN^\epsilon} \\
&\leq \int_{N=N_3-1}^{\infty} Ce^{-\frac{1}{16}(c_1x_0)^2t}dt + \int_{N_3-1}^{\infty} e^{-\frac{1}{2}c_1^2nt^\epsilon}dt \\
&= A + B.
\end{aligned}$$

Furthermore, it can be seen that

$$A = \int_{N=N_3-1}^{\infty} Ce^{-\frac{1}{16}(c_1x_0)^2t}dt < \infty.$$

In the following, we will show that $B = \int_{N_3-1}^{\infty} e^{-\frac{1}{2}c_1^2nt^\epsilon}dt < \infty$. By a change of variable using $x = \frac{1}{2}c_1^2nt^\epsilon$, we have that

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

with Γ 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 $\{\mathcal{T} = \mathcal{T}^{lp}(N)\}_{N=N'}^{\infty}$ will hold. Hence (a) and (b) are proven. □

4 Illustrative Experiments

This section supports the findings in the previous section with numerical examples and make the comparisons with the other algorithms which possess the ORACLE property in the literatures.

4.1 Experiment 1

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

- $\mathbf{x}^0$ 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 $\mathbf{v} \in \mathbb{R}^N$ 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 $\mathbf{x}^0$. In this experiment, ϵ for the proposed method is set to $\frac{1}{3}$; λ_N for 'ADALASSO' is chosen as $N^{1/2-\gamma/4}$ (this choice satisfies all the assumptions in Theorem 2 in Zou (2006)), and γ 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.

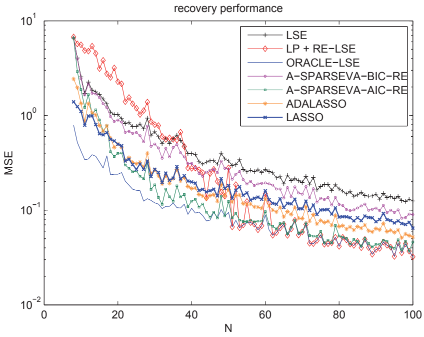


Fig. 2. Performance of the different estimators from N observations to estimate $\mathbf{x}^0$. This picture indicates that the proposed estimator will give exactly the same performance as the ORACLE estimator for a large N ($N \approx 75$).

Fig. 3 demonstrates the efficacy of support recovery of the LP formulation in Eq. (6) for different choices of ϵ. 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.

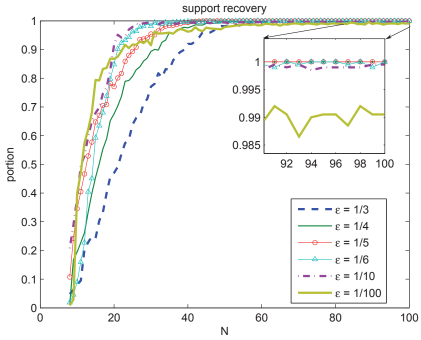


Fig. 3. Support recovery performance of Eq. (6) for different choices of ϵ. Empirically, we observe that: 1) When ϵ is chosen to be small, the ratio for successful support recovery will be larger when N is small; but when N is large, the ratio for successful support recovery will converge slower to 100% and oscillation exists. This can be observed in the zoomed-in part. 2) When ϵ is chosen to be large, the ratio for successful support recovery will be smaller when N is small; but when N is large, the ratio for successful support recovery will go faster to 100% and no oscillation exists, see also the zoomed-in part in the figure.

In practice, cross validation technique could be exploited to choose the tuning parameters. In the following, we will take the ADALASSO and the proposed method for granted to illustrate the idea and compare the performances for both methods when the parameters are obtained by the cross validation technique. In the ADALASSO algorithm and the proposed algorithm, there are two tuning parameters, namely γ for the ADALASSO, and ϵ for the proposed method. In the following part, we will apply the 5-fold cross-validation method (see Trevor, Hastie, Tibshirani and Friedman (2005)) 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 $\|\hat{\mathbf{x}} - \mathbf{x}^0\|_2^2$, where $\hat{\mathbf{x}}$ denotes the estimate obtained by the estimator. ϵ are selected from $\{1/8, 1/4, 1/2\}$, γ 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:

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

Here both $\{w_{i_{k'}}\}_{k'}$ and $\{c_{i_{k'}}\}_{k'}$ 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

$$\mathbf{y} = \mathbf{A}\mathbf{c}^0 + \mathbf{v}, \tag{10}$$

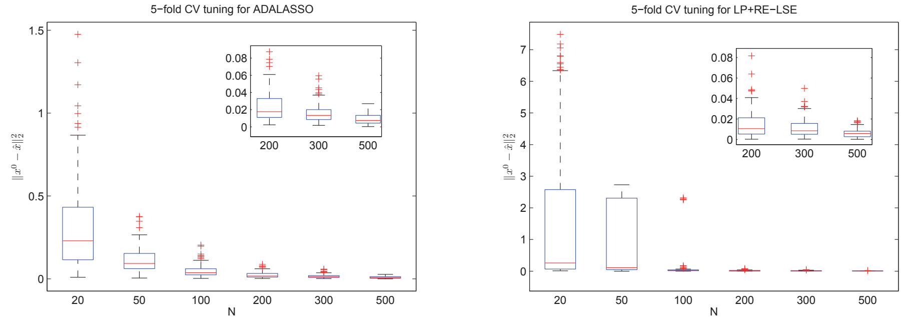
where $\mathbf{y} = [y(t_s), \cdots, y(Nt_s)]^T$. The matrix $\mathbf{A} \in \mathbb{R}^{N \times n}$ is defined as follows. The i-th row of $\mathbf{A}$ is given by

$$\mathbf{A}_i = [\sin(iw_1 t_s), \sin(iw_2 t_s), \ldots, \sin(iw_n t_s)], \tag{11}$$

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

In this experiment, $n = 10$ and $\mathbf{c}^0 = (1, 1, 1, 0, \cdots, 0)^T$, $w_k = k$ for $k = 1, 2, \cdots, n$. We increase N up to 500 and the noise vector $\mathbf{v}$ satisfies $\mathbf{v} \backsim \mathcal{N}(0, I_N)$. 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 $\mathbf{x}^0$ is displayed in Fig. 5. It is again clear that the proposed



(a)

(b)

Fig. 4. This figure demonstrates the boxplots of the recovery error obtained through the ADALASSO estimator and the proposed estimator when the tuning parameters are chosen by the 5-fold cross validation method. From this figure, we can see that performances of both methods are similar when N is large, see the zoomed-in part in the figures. It can also be observed that when N is small, the ADALASSO method has smaller recovery error compared with the proposed method.

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.

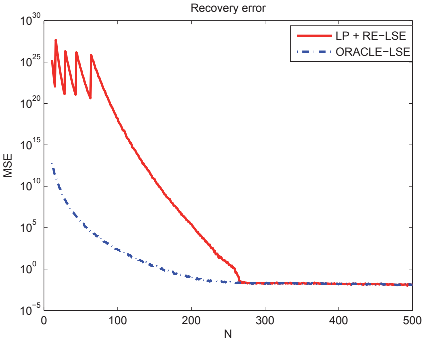


Fig. 5. Performance of applying the proposed estimator to recovery sinusoids functions from N observations in Experiment 2. This example also indicates that after a finite number the estimate is exactly equal to the ORACLE estimator.

This is indeed predicted by the theory above since the $\mathbf{A}$ in Eq. (10) obeys the assumption of Eq. (4). This follows from the proposition given as:

Proposition 1 *There exist constants* $\{C_{i,j}\}_{0\leq i,j\leq n}$ *which do not depend on* N*, such that the following results hold. For any* $1 \leq i \neq j \leq n$*, one has that:*

$$\left|(A^T A)_{i,j}\right| = \left|\sum_{t=1}^{N} \sin(tw_i t_s)\sin(tw_j t_s)\right| \leq C_{i,j} \qquad (12)$$

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

$$(A^T A)_{i,i} = \sum_{t=1}^{N} (\sin(tw_i t_s))^2 \geq \frac{N}{2} - C_{i,i}. \qquad (13)$$

The proof is given in Appendix A. With this proposition, an application of Geršgorin circle theorem implies that the eigenvalues of $\mathbf{A}^T\mathbf{A}$ 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 $\mathbf{x}^0$ 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 $\mathbf{W}$ and how to select the λ in Eq. (8), in order to make Eq. (8) a practical improvement of the proposed method which also maintains the ORACLE property.

References

Rojas, C. & Hjalmarsson, H. (2011). Sparse estimation based on a validation criterion. *The 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC11)*, Orlando, USA.

Donoho, D. L. (2006). Compressed sensing. *Information Theory, IEEE Transactions on,* 52(4), 1289-1306. Germany: De Gruyter.

Plackett, R. L. (1950). Some theorems in least squares. Biometrika, 37(1/2), 149-157.

Rick, D. (2005). *Probability: Theory and Examples (2nd Edition).* Duxbury press.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432), 1200-1224.

Söderström, T. & Stoica, P. (1989). *System Identification.* UK: Prentice-Hall International.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348-1360.

Candés, E. & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6), 2313-2351.

Kump, P., Bai, Er-W., Chan, K. S., Eichinger, B. & Li K. (2012). Variable selection via RIVAL (removing irrelevant variables amidst Lasso iterations) and its application to nuclear material detection. *Automatica*, 48(9), 2107-2115.

Kukreja, S. L. (2009). Application of a least absolute shrinkage and selection operator to aeroelastic flight test data. *International Journal of Control*, 82(12), 2284-2292.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 73(3), 267-288.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418-1429.

Efron, B., Hastie, T. & Tibshirani R. (2007). Discussion of 'the Dantzig selector'. *The Annals of Statistics*, 35(6), 2358-2364.

Stoica, P. & Moses, R. L. (1997). *Introduction to spectral analysis (Vol. 89).* New Jersey: Prentice hall.

Kale, B. K. (1985). A note on the super efficient estimator. Journal of Statistical Planning and Inference, 12, 259-263.

Leeb, H. & Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges's estimator. *Journal of Econometrics,* 142(1), 201-211.

Candés, E. J. & Wakin, M. B. (2008). An introduction to compressive sampling. *Signal Processing Magazine, IEEE,* 25(2), 21-30.

Trevor, J., Hastie, T., Tibshirani, R. J. & Friedman, J. H. (2005). *The elements of statistical learning: data mining, inference, and prediction.* Springer.

A Proof of Proposition 4

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

$$\begin{aligned}&\left|\sum_{t=1}^{N}\sin(tw_it_s)\sin(tw_jt_s)\right|\\ &=\frac{1}{2}\left|\sum_{t=1}^{N}\left(\cos(t(w_i-w_j)t_s)-\cos(t(w_i+w_j)t_s)\right)\right|\\ &\le\frac{1}{2}\left|\sum_{t=1}^{N}\cos(t(w_i-w_j)t_s)\right|+\frac{1}{2}\left|\sum_{t=1}^{N}\cos(t(w_i+w_j)t_s)\right|.\end{aligned}$$

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.

$$\begin{aligned}&\left|\sum_{t=1}^{N}\cos(t(w_i-w_j)t_s)\right|\\ &=\left|\mathrm{Re}\left(\frac{1-e^{j(N+1)(w_i-w_j)t_s}}{1-e^{j(w_i-w_j)t_s}}\right)-1\right|\\ &\le\left|\frac{1-e^{j(N+1)(w_i-w_j)t_s}}{1-e^{j(w_i-w_j)t_s}}\right|+1\\ &\le\frac{2}{\left|1-e^{j(w_i-w_j)t_s}\right|}+1,\end{aligned}$$

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

In order to prove inequality (13), observe that

$$\begin{aligned}\sum_{t=1}^{N}(\sin(tw_it_s))^2&=\frac{1}{2}\sum_{t=1}^{N}(1-\cos(2tw_it_s))\\ &\ge\frac{N}{2}-\frac{1}{2}\left|\sum_{t=1}^{N}\cos(2tw_it_s)\right|.\end{aligned}$$

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

Paper II

On the Nuclear Norm Heuristic for a Hankel Matrix Completion Problem

Liang Dai [a] and Kristiaan Pelckmans [a]

[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 addresses the question if and why the nuclear norm heuristic can recover an impulse response generated by a stable single-real-pole system, if elements of the upper-triangle of the associated Hankel matrix were given. Since the setting is deterministic, theories based on stochastic assumptions for low-rank matrix recovery do not apply here. A 'certificate' which guarantees the completion is constructed by exploring the structural information of the hidden matrix. Experimental results and discussions regarding the nuclear norm heuristic applied to a more general setting are also given.

Key words: System Identification; Matrix Completion.

1 Introduction

Techniques of convex relaxation using the nuclear norm heuristic have become increasingly popular in the systems and control society, see e.g. the examples reported in [1], [2] and the discussions therein. This note provides a novel theoretical justifications for the usage of the nuclear norm heuristic in an fundamental task in systems theory, i.e. to recover the impulse response of a system from the first few entries of the related series. Precisely, we make the following assumptions throughout the note: (1) the provided entries are exact, i.e. there is no noise present, (2) the first n entries of the impulse response are provided while the last $n-1$ entries are to be completed. Note that this setting has *no stochastic quantities involved*.

The problem considered here can be casted as a special case of the 'matrix completion' problem [5,6,7]. However, in the problem considered in this work, the sampled entries are given deterministically, while 'matrix completion' problems are typically analyzed using random sampling patterns. And also in the current work, the underlying matrix is a structural (Hankel) matrix. These differences make the theories in the literature not applicable to this problem. While this task can be easily solved using standard techniques [1,9], the rationale for this work is that to provide a complete picture for understanding how the nuclear norm heuristic performs on this fundamental problem by some new proof techniques.

This contribution is organized as follows. The main theorem is given in Section II. The proof of the result will be given in section III. Section IV gives more discussions in a more general matrix completion case and conclude the note.

The following notational conventions will be used. Vectors are denoted in boldface, scalars are denoted in lowercase, matrices as capital letters, and sets are represented as calligraphic letters. $\mathcal{H}_n$ denotes the set of $n \times n$ Hankel matrices, I_n denotes the identity matrix of size $n \times n$, $\mathbf{e}_i$ denotes the unit vector with only the i-th element to be one and all the other elements zero, $\|\cdot\|_*$ represents the nuclear norm (sum of all the singular values) of a matrix, $\|\cdot\|_2$ represents the spectral norm of a matrix, and $\|\cdot\|_F$ represents the Frobenius norm of a matrix.

2 Results

The following theorem states the finding formally.

Theorem 1 Given $-1 < h < 1$, define vector $\mathbf{h} \in \mathbb{R}^n$ as $\mathbf{h} = [1, h, h^2, \ldots, h^{n-1}]^T$, and matrix $G_0 \in \mathcal{H}_n$ as $\mathbf{h}\mathbf{h}^T$. Consider the following application of the nuclear norm heuristic:

$$\hat{G}_0 \triangleq \underset{G \in \mathcal{H}_n}{\arg\min} \|G\|_* \quad \text{s.t. } G(i,j) = G_0(i,j), \forall\ (i+j) \leq n+1. \tag{1}$$

Then $\hat{G}_0 = G_0$ and consequently, we say that (1) reconstructs G_0.

Remark 1 *Since the true matrix G_0 is of low rank (of rank one), successful reconstruction of G_0 up to the first $2n-1$ entries is given by the solution of*

$$\begin{aligned}\tilde{G} \triangleq &\arg\min_{G\in\mathcal{H}_n} \ \textit{rank}(G) \\ &\textit{s.t.} \ \ G(i,j) = G_0(i,j), \forall (i+j) \le n+1.\end{aligned} \tag{2}$$

The solution to this particular problem can lead to an exact recovery of G_0, see the results in [3,8,9].

2.1 Proof of Theorem 1

The *sketch* for the whole proof of the Theorem 1 is as follows: Lemma 1 gives a sufficient condition for the recovery of G_0 by solving eq. (1). Lemma 2 and Lemma 3 are devoted to build a 'certificate' which can guarantee such condition is always satisfied under the assumptions made in Theorem 1.

For the matrices G_0 and G as defined in Theorem 1, define:

$$H = G_0 - G, \tag{3}$$

Notice that all the entries of H in the upper triangle part will be zero by construction, so H can always be decomposed as

$$H = \sum_{i=1}^{n-1} v_i G_i, \tag{4}$$

where $\{G_i\}_{i=1}^{n-1}$ are the basis matrices with the elements of the ith lower anti-diagonal equal to 1 and the others equal to zero and $v_i \in \mathbb{R}, \forall i = 1, \cdots, n-1$. For instance, when $n = 3$, one has that

$$H = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & v_1 \\ 0 & v_1 & v_2 \end{pmatrix} = v_1 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

For the notational convenience, we define the following two projection matrices, namely

$$P = \frac{G_0}{\|\mathbf{h}\|_2^2}$$

and the complement projection matrix of P as

$$Q = I_n - P.$$

Proposition 1 will be used in the following discussions, which tells that the nuclear norm is the dual norm of the spectral norm for a given matrix [7].

Proposition 1 *Given $A \in \mathbb{R}^{n\times n}$ matrix, then*

$$\|A\|_* = \sup\{\mathrm{tr}(MA) : \|M\|_2 \le 1, M \in \mathbb{R}^{n\times n}\}. \tag{6}$$

We also need the following fact.

Proposition 2 *Given H as defined in eq. (3), if $H \neq 0$, then $QHQ \neq 0$.*

Proof We prove that the only possibility for $QHQ = 0$ to hold is when $H = 0$. Notice that $H = (P+Q)H(P+Q)$, expanding this equality, we have that

$$H = PHP + PHQ + QHP + QHQ.$$

Hence if $QHQ = 0$, we have that

$$\begin{aligned} H &= PHP + PHQ + QHP \\ &= PH + QHP. \end{aligned}$$

As $P = \frac{\mathbf{h}\mathbf{h}^T}{\|\mathbf{h}\|_2^2}$, then the previous relation implies that H can be represented as $\mathbf{h}\mathbf{a}^T + \mathbf{b}\mathbf{h}^T$ where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Due to the fact that H is symmetric, so we have that

$$\mathbf{h}\mathbf{a}^T + \mathbf{b}\mathbf{h}^T = \mathbf{a}\mathbf{h}^T + \mathbf{h}\mathbf{b}^T,$$

or equivalently

$$\mathbf{h}(\mathbf{b}-\mathbf{a})^T = (\mathbf{b}-\mathbf{a})\mathbf{h}^T. \tag{7}$$

Given the fact in eq. (7), the two *rank-one* matrices $\mathbf{h}(\mathbf{b}-\mathbf{a})^T$ and $(\mathbf{b}-\mathbf{a})\mathbf{h}^T$ will have the same row space and column space, which implies that $\mathbf{b}-\mathbf{a} = k\mathbf{h}$, for some $k \in \mathbb{R}$.

This gives that H can be written as

$$H = \mathbf{h}\mathbf{a}^T + \mathbf{b}\mathbf{h}^T = \mathbf{h}\mathbf{a}^T + \mathbf{a}\mathbf{h}^T + k\mathbf{h}\mathbf{h}^T,$$

i.e.

$$H = (\mathbf{a} + \frac{k}{2}\mathbf{h})\mathbf{h}^T + \mathbf{h}(\mathbf{a} + \frac{k}{2}\mathbf{h})^T.$$

Let $\mathbf{c} = (c_1, c_2, \cdots, c_n)^T = \mathbf{a} + \frac{k}{2}\mathbf{h}$. Notice that the i-th element of the first column of H equals $h^{i-1}c_1 + c_i$. By construction, the first column of H is a zero vector, hence for $i = 1$, we have that $2c_1 = 0$, which gives that $c_1 = 0$. Thus the i-th element of the first column of H equals c_i which implies that $c_2 = \cdots = c_n = 0$, i.e. $\mathbf{c} = 0$. This implies that $H = 0$ as desired. □

Lemma 1 provides a sufficient condition for Theorem 1 to hold.

Lemma 1 *If for any H as in eq. (3), one has that*

$$|\operatorname{tr}(PH)| < \|QHQ\|_*, \tag{8}$$

then the optimization problem (1) recovers G_0 exactly.

In other words, this lemmas says that, for any nonzero deviation H from G_0, one has that the nuclear norm of $G_0 + H$ will become strictly larger when eq. (8) holds.

Proof Let $V \in \mathbb{R}^{n\times(n-1)}$ be a matrix which satisfies $VV^T = Q$ and $V^TV = I_{n-1}$. Hence the sub-gradients of $\|\cdot\|_*$ at G_0 are given as the set (see e.g. [7]):

$$\mathcal{S}_{\mathbf{h}} = \left\{P + VBV^T \ : \|B\|_2 \leq 1\right\}. \tag{9}$$

By the property of sub-gradient, we have that for any H as in eq. (3),

$$\|G_0 + H\|_* \geq \|G_0\|_* + \langle H, F\rangle,$$

where $F \in \mathbb{R}^{n\times n}$ is any matrix which belongs to $\mathcal{S}_{\mathbf{h}}$.

Hence, for any H as in eq. (3), if there exists one element in $\mathcal{S}_{\mathbf{h}}$, i.e. one B with $\|B\|_2 \leq 1$, such that

$$\left\langle H, P + VBV^T\right\rangle > 0$$

or equivalently

$$\operatorname{tr}(HP) > \langle V^THV, -B\rangle, \tag{10}$$

then we have $\|G_0+H\|_* > \|G_0\|_*$, which implies the claims in Theorem 1. Hence, we are left to find a matrix which satisfies inequality (10) given the assumption (8).

From eq. (8), we have that

$$|\operatorname{tr}(HP)| < \|QHQ^T\|_*,$$

and since by definition of Q as a projection matrix onto an $n-1$ dimensional subspace, one has that

$$\|QHQ^T\|_* = \|V^THV\|_*,$$

it follows that

$$|\operatorname{tr}(HP)| < \|V^THV\|_*.$$

Furthermore, it follows from Proposition 1 that there exists a matrix B_1 with $\|B_1\|_2 \leq 1$, such that

$$\|V^THV\|_* = \left\langle V^THV, B_1\right\rangle,$$

therefore,

$$|\operatorname{tr}(HP)| < \left\langle V^THV, B_1\right\rangle.$$

Conversely, one has

$$\operatorname{tr}(HP) > -\left\langle V^THV, B_1\right\rangle = \langle V^THV, -B_1\rangle,$$

and hence, the inequality (10) holds for B_1, which proves the result. □

Next, we prove that the condition in Lemma 1 will always hold whenever $H \neq 0$. Lemma 2 constructs a matrix M_0 which will be used in Lemma 3, while Lemma 3 constructs a 'certificate' M_1 explicitly which guarantees the satisfaction of eq. (8).

The *sketch* for proving Lemma 2 is as follows. First, we construct two matrices, namely the Q_1 and the Q_2, by considering two related linear equations. Then, we can have four related properties about Q_1 and Q_2, i.e. *Fact 1, Fact 2, Fact 3 and Fact 4*, which are very useful in the derivation. Finally, we construct a matrix M_0 based on Q_1 and Q_2.

Lemma 2 *Given the matrices $G_i, P, Q \in \mathbb{R}^{n\times n}$ defined as before, there exists a matrix $M_0 \in \mathbb{R}^{n\times n}$ with $\|M_0\|_2 < 1$, such that*

$$\operatorname{tr}(QG_iQM_0 - G_iP) = 0,\ \forall i = 1, 2, \ldots, n-1. \tag{11}$$

Proof We will give a construction of such matrix M_0. Let $r > 0$ denote the norm of vector $\mathbf{h}$, which clearly satisfies

$$r^2 = \|\mathbf{h}\|_2^2 = 1 + h^2 + \cdots + h^{2n-2}.$$

We construct two matrices $Q_1 \in \mathbb{R}^{n\times n}$ and $Q_2 \in \mathbb{R}^{n\times n}$ which satisfy the following two equations:

$$\begin{aligned} r^2(Q_1 + Q_2) &= r^2Q = r^2I_n - G_0 \\ &= r^2I_n - \begin{bmatrix} 1 & h & h^2 & \cdots & h^{n-1} \\ h & h^2 & h^3 & \cdots & h^n \\ h^2 & h^3 & h^4 & \cdots & h^{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h^{n-1} & \ldots & \ldots & \ldots & h^{2n-2} \end{bmatrix}, \end{aligned} \tag{12}$$

and $r^2(Q_1 - Q_2) \in \mathbb{R}^{n\times n}$, which is defined in eq. (13).

The defined matrices Q_1, Q_2 will have the following properties:

- *Fact 1:*

$$(Q_1 + Q_2)(Q_1 + Q_2) = (Q_1 - Q_2)(Q_1 - Q_2). \tag{14}$$

- *Fact 2:*

$$(Q_1 + Q_2)(Q_1 - Q_2) = (Q_1 - Q_2)(Q_1 + Q_2).$$

$$r^2(Q_1 - Q_2) = \tag{13}$$
$$\begin{bmatrix} -h^{n} & -h^{n+1} & -h^{n+2} & \cdots & -h^{2n-2} & h+h^3+\cdots+h^{2n-3} \\ -h^{n+1} & -h^{n+2} & \vdots & -h^{2n-2} & h+h^3+\cdots+h^{2n-3} & -1 \\ -h^{n+2} & \vdots & -h^{2n-2} & h+h^3+\cdots+h^{2n-3} & -1 & -h \\ \vdots & -h^{2n-2} & h+h^3+\cdots+h^{2n-3} & -1 & \vdots & \vdots \\ -h^{2n-2} & h+h^3+\cdots+h^{2n-3} & -1 & \vdots & \vdots & -h^{n-3} \\ h+h^3+\cdots+h^{2n-3} & -1 & -h & \cdots & -h^{n-3} & -h^{n-2} \end{bmatrix}.$$

- *Fact 3:*
$$Q_1Q_2 = Q_2Q_1 = 0.$$
- *Fact 4:*
$$Q_1^2 = Q_1, \quad Q_2^2 = Q_2.$$

Fact 3 and Fact 4 basically imply that the matrices Q_1 and Q_2 are projection matrices and moreover are orthogonal to each other. Due to this observation, the matrix $(Q_1 - Q_2)$ will have eigenvalues in the set $\{1, -1, 0\}$, which clearly means that the spectral norm of $(Q_1 - Q_2)$ is 1.

These discussions lead us to consider the following choice of M_0:
$$M_0 = -h^n(Q_1 - Q_2). \tag{15}$$
Now we can prove that matrix M_0 satisfies all the equalities given in eq. (11) based on these Facts. First, notice that the equalities in eq. (11) are equivalent to the following equalities

$$\operatorname{tr}\left(G_i(QM_0Q - P)\right) = 0, \ \forall i = 1, 2, \ldots, n-1. \tag{16}$$

The term $QM_0Q - P$ in eq. (16) can be calculated out as follows:

$$\begin{aligned} QM_0Q - P &= -h^n(Q_1 + Q_2)(Q_1 - Q_2)(Q_1 + Q_2) - P \\ &= -h^n(Q_1 - Q_2)(Q_1 + Q_2) - P \\ &= -h^n(Q_1 - Q_2) - P \\ &= M_0 - P. \end{aligned}$$

In the previous derivations, we have made use of the fact that

$$(Q_1 + Q_2)(Q_1 - Q_2) = (Q_1 - Q_2),$$

which could be verified by expanding the left hand side and using the fact that both Q_1 and Q_2 are projection matrices. So, to prove the the equalities in eq. (16) is equivalent to prove that

$$\operatorname{tr}\left(G_i(M_0 - P)\right) = 0, \ \forall i = 1, 2, \ldots, n-1. \tag{17}$$

Notice that M_0 has the same elements as P in the lower anti-diagonal part, so eq. (17) holds, which in turn makes eq. (11) hold. Notice that $\|M_0\|_2 = |h|^n$, which is less than 1. This concludes that M_0 is the desired matrix.

Remark 2 *Here we are not constructing the matrices Q_1 and Q_2 explicitly. But from eq. (12) and eq. (13), the matrices Q_1 and Q_2 can be reconstructed, and from which, we can see that both Q_1 and Q_2 are all nonzero matrices.*

We are left to prove the Facts 1, 2, 3 and 4. In order to keep clarity of the note, we leave the detailed verifications to the appendix. □

Based on the constructed M_0 in Lemma 2, we can certify that:

Lemma 3 *For any H as given in eq. (3), we have that*

$$|\operatorname{tr}(PH)| < \|QHQ\|_*. \tag{18}$$

Proof We distinguish between two cases, namely

$$\operatorname{tr}(PH) < \|QHQ\|_*, \tag{19}$$

and
$$-\operatorname{tr}(PH) < \|QHQ\|_*.$$
We will give a derivation of eq. (19), the latter inequality follows along the same lines. With the application of Proposition 1, it follows that to prove eq. (19) is equivalent to prove that

$$\sup_{\|M\|_2 \le 1} \operatorname{tr}\left(QHQM - HP\right) > 0. \tag{20}$$

Notice that $H = \sum_{i=1}^{n-1} v_i G_i$, and that by construction of M_0 in Lemma 2, we have that

$$\operatorname{tr}\left(QHQM_0 - HP\right) = \sum_{i=1}^{n-1} v_i \operatorname{tr}(QG_iQM_0 - G_iP) = 0.$$

Next, observe that M_0 is strictly inside the ball $\|M\|_2 \le 1$, hence there exists a small value $\delta > 0$ such that

$$M_1 = M_0 + \delta(QHQ), \tag{21}$$

will also be inside the unit ball $\|M\|_2 \leq 1$. Since $H \neq 0$, it follows that $QHQ \neq 0$, which implies that

$$\operatorname{tr}(QHQM_1 - HP) = \delta \operatorname{tr}(QHQQHQ) = \delta \|QHQ\|_F^2$$

is positive. This certifies eq. (20) and hence ineq. (19), which in turn concludes the proof of Lemma 3. □

In conclusion, application of Lemmas 1, 2 and 3 gives Theorem 1.

3 Discussions

The previous sections studies a low rank matrix completion problem where the matrix to be recovered is known to be Hankel and the revealed entries follow a deterministic pattern. It is shown that the nuclear norm approach gives the correct answer in case the exact (noiseless) entries of the upper triangular part of this matrix are provided and the system is a single real pole stable system.

It is natural to raise the question whether the nuclear norm heuristic will still work when the rank of the matrix G_0 is larger than 1. The answer is generally negative. We will provide an numerical illustration of this finding. Consider the following example of a second order system. Let $h_1, h_2 \in \mathbb{R}$ be the two poles which satisfy $-1 < h_1, h_2 < 1$. We further assume that the impulse response of the system is given by the sequence $\{h_1^{i-1} + h_2^{i-1}\}_{i=1}^{\infty}$. let $n = 10$, i.e., the matrix G_0 is of size 10×10, then the completion problem based on the nuclear norm heuristic is given as

$$\hat{G} = \underset{G \in \mathcal{H}_{10}}{\arg\min} \|G\|_* \qquad (22)$$
$$\text{s.t. } G(i,j) = G_0(i,j), \forall i + j \leq 11.$$

Now we can compare the value of the nuclear norm of the optimum $\hat{G}$ with the value of the nuclear norm of G_0 obtained by filling out the remaining entries using the specification of the system. Figure (1) displays the difference between the nuclear norm of G_0 and $\hat{G}$ for different choices of h_1 and h_2, which are chosen as $h_1 = -0.94 : 0.05 : 0.94$ and $h_2 = -0.94 : 0.05 : 0.94$. From this experiment, it becomes clear that G_0 does not always has minimal nuclear norm, and recovery by eq. (22) will not necessarily succeed. However, it is worthwhile to mention that, in most cases, the nuclear norm heuristic gives the correct recovery.

Hence we conclude the article with the following open questions which are left for future work: (1) A rigorous characterization of when the nuclear norm heuristic will work in the stable multiple-pole system case is in order. By inspecting the proofs for the single pole case in this note, we can see that the Lemmas 1 and 3 are also applicable in such case. More precisely, When the matrix M_0 in Lemma 2 is constructed, then a 'certificate' to guarantee the successful completion in this general case can be constructed the same way as in Lemma 3. However, it is evident that such a construction for M_0 is more complicated. So, how to construct this matrix is left as an open question for future research. (2) Another open question is that when the nuclear norm heuristic doesn't work, see the cases in the previous example, it is interesting to find out which additional assumptions could assist the heuristic to work. (3) Thirdly, the results in this article assumes noiseless data, it is not clear how this assumption can be relaxed in the noisy case.

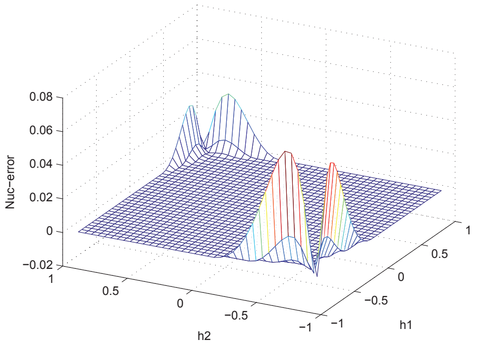


Fig. 1. This figure displays $\|G_0\|_*$- $\|\hat{G}\|_*$ for a range of 2 real poles. It is seen that the nuclear norm objective value is not always minimal for the true system G_0 for many choices of h_1 and h_2, implying that the heuristic will not always work for such systems. Note that the difference is exactly equal to zero for the case where $h_1 = h_2$ as confirmed by Theorem 1.

References

[1] L. Vandenberghe, Convex optimization techniques in system identification, *16th IFAC Symposium on System Identification*, Brussels, Belgium, July 2012.

[2] I. Markovsky, How effective is the nuclear norm heuristic in solving data approximation problems?, *16th IFAC Symposium on System Identification*, Brussels, Belgium, July 2012.

[3] M. Fazel, H. Hindi, and S. Boyd, Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices, *Proceeding of American Control Conference*, Denver, Colorado, June 2003.

[4] M. Fazel, H. Hindi, and S. Boyd, A rank minimization heuristic with application to minimum order system approximation, *Proceeding of American Control Conference*, Arlington, Virginia, June 2001.

[5] E. J. Candés and B. Recht, Exact matrix completion via convex optimization, *Foundation of Computational Mathematics*, Vol 9, pp. 717-772, 2009.

[6] D. Gross, Recovering low-rank matrices from few coefficients in any basis, *IEEE Transaction on Information Theory*, Vol 57, pp. 1548-1566, 2011.

[7] B. Recht, M. Fazel and P. A. Parrilo, Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization, *SIAM Review*, Vol 52, no 3, pp. 471-501, 2010.

[8] Z. Liu, L. Vandenberghe, Semidefinite programming methods for system realization and identification, *Proceedings of the 48th IEEE Conference on Decision*, Shanghai, China, 2009.

[9] A. Tether, Construction of minimal linear state-variable models from finite input-output data, *IEEE Transactions on Automatic Control*, Vol 15, pp. 427-436, 1970.

A Proofs of Fact 1, 2, 3 and 4

Fact 1 is proven as follows: Notice that to prove (14) is equivalent to prove

$$r^2(Q_1+Q_2)r^2(Q_1+Q_2) = r^2(Q_1-Q_2)r^2(Q_1-Q_2).$$

By definition, one has that

$$r^2(Q_1{+}Q_2)r^2(Q_1{+}Q_2) = (r^2I_n{-}G_0)^2 = r^4I_n{-}r^2(\mathbf{h}\mathbf{h}^T).$$

So, we are left to prove that

$$r^2(Q_1-Q_2)r^2(Q_1-Q_2) = r^4I_n - r^2(\mathbf{h}\mathbf{h}^T).$$

We will give expression of the elements of $r^2(Q_1 - Q_2)r^2(Q_1-Q_2)$ by calculating out the off-diagonal entries and the on-diagonal entries separately.

- *Off-diagonal elements (except for the last column and last row).* Take for any $1 \le k_1, k_2 < n$ the corresponding columns from the matrix $\Delta = r^2(Q_1-Q_2)$ as (assume that $k_1 < k_2$ without loss of generality):

$$\begin{aligned}\Delta_{k_1} = \big[&-h^{n+k_1-1}, -h^{n+k_1}, \cdots,\\ &-h^{2n-2}, -h^{2n-1}, -1, \cdots, -h^{k_1-2}\big]^T + hr^2\mathbf{e}_{n-k_1+1},\end{aligned}$$

and

$$\begin{aligned}\Delta_{k_2} = \big[&-h^{n+k_2-1}, -h^{n+k_2}, \cdots,\\ &-h^{2n-2}, -h^{2n-1}, -1, \cdots, -h^{k_2-2}\big]^T + hr^2\mathbf{e}_{n-k_2+1}.\end{aligned}$$

So, the (k_1, k_2) and (k_2, k_1) elements of $r^2(Q_1 - Q_2)r^2(Q_1-Q_2)$ are given by the inner-product:

$$\begin{aligned}&\Delta_{k_1}^T\Delta_{k_2} =\\ &\big[-h^{n+k_1-1}, -h^{n+k_1}, \cdots, -h^{n+k_1-k_2-1}\big]\\ &\quad\big[-h^{n+k_2-1}, -h^{n+k_2}, \cdots, -h^{2n-1}\big]^T\\ &+\big[-h^{2n+k_1-k_2}, -h^{n+k_1-k_2+1}, \cdots, -h^{2n-1}\big]\\ &\quad\big[-1, -h, \cdots, -h^{k_2-k_1-1}\big]^T\\ &+\big[-1, -h, \cdots, -h^{k_1-2}\big]\\ &\quad\big[-h^{k_2-k_1}, -h^{k_2-k_1+1}, \cdots, -h^{k_2-2}\big]^T\\ &+hr^2(h^{2n+k_1-k_2-1}+h^{k_2-k_1-1}).\end{aligned}$$

Reorganizing the equation, we have that

$$\begin{aligned}\Delta_{k_1}^T\Delta_{k_2} &= -(h^{k_2+k_1-2}+h^{k_2+k_1}+\cdots+h^{2n+k_2+k_1-4})\\ &= -h^{k_1+k_2-2}r^2,\end{aligned}$$

which is as desired.

- *Off-diagonal elements (in the last column and last row).* Take for any $1 \le k < n$ and n the corresponding columns from the matrix $\Delta = r^2(Q_1-Q_2)$ as:

$$\begin{aligned}\Delta_k = \big[&-h^{n+k-1}, -h^{n+k}\cdots,\\ &-h^{2n-2}, \sum_{i=1}^{n-1}h^{2i-1}, -1, \cdots, -h^{k-2}\Bigg]^T,\end{aligned}$$

and

$$\begin{aligned}\Delta_n = \Bigg[&\sum_{i=1}^{n-1}h^{2i-1}, -1, \cdots,\\ &-h^{n-k-3}, -h^{n-k-2}, -h^{n-k-1}, \cdots, -h^{n-2}\big]^T.\end{aligned}$$

So, the (k, n) and (n, k) elements of $r^2(Q_1{-}Q_2)r^2(Q_1{-}Q_2)$ are given by the inner-product

$$\begin{aligned}\Delta_k^T\Delta_n = \quad &-(h^{n+k-1}+h^{n+k}+\cdots+h^{3n+k-4})\\ &+(h^{n+k}+h^{n+k+2}+\cdots+h^{3n-k-5})\\ &-(h^{n-k-1}+h^{n-k+1}+\cdots+h^{3n-k-5})\\ &+(h^{n-k-1}+h^{n-k+1}+\cdots+h^{n+k-4}).\end{aligned}$$

Reorganizing the equation, we have that

$$\begin{aligned}\Delta_k^T\Delta_n &= -(h^{n+k-2}+h^{n+k-1}+\cdots+h^{3n+k-4})\\ &= -h^{n+k-2}r^2,\end{aligned}$$

as desired.

- *The (k,k) entries where* $1 \le k < n$. We need to verify the following equation:

$$\begin{aligned}&(h^{2n+2k-2}+h^{2n+2k}+\cdots+h^{4n-4})\\ &\quad+(h+h^3+\cdots+h^{2n-3})^2+(1+h^2+\cdots+h^{2k-4})\\ &\quad=(1+h^2+\cdots+h^{2n-2})^2-(1+h^2+\cdots+h^{2n-2})h^{2k-2},\end{aligned}$$

which is equivalent to verify that

$$\begin{aligned}(1+h^2+\cdots+h^{4n-4})+(h+h^3+\cdots+h^{2n-3})^2\\ =(1+h^2+\cdots+h^{2n-2})^2.\end{aligned}\tag{A.1}$$

This can be verified by the following:

$$\begin{aligned}&\Leftrightarrow \frac{(1-h^2)(1-h^{4n-2})+h^2(1-h^{2n-2})^2}{(1-h^2)^2} = \frac{(1-h^{2n})^2}{(1-h^2)^2}\\ &\Leftrightarrow 1+h^{4n}-2h^{2n} = (1-h^{2n})^2.\end{aligned}$$

- *The (n,n) entry.* We need to verify the following equation:

$$\left(\sum_{i=1}^{n-1} h^{2i-1}\right)^2 + (1+h^2+\cdots+h^{2n-4})$$
$$= (1+h^2+\cdots+h^{2n-2})^2 - (1+h^2+\cdots+h^{2n-2})h^{2n-2},$$

  which is equivalent to verify that

$$(1+h^2+\cdots+h^{4n-4}) + (h+h^3+\cdots+h^{2n-3})^2$$
$$= (1+h^2+\cdots+h^{2n-2})^2.$$

  This equality has been verified in eq. (A.1). In sum, we have proved the Fact 1.

For the proof of Fact 2, we need to notice that the vector $\mathbf{h}$ lies in the null space of the matrix $\Delta = r^2(Q_1 - Q_2)$. As matrix Δ is a square Hankel matrix, so it is to prove that any column of Δ is orthogonal with $\mathbf{h}$.

- *The first (n-1) columns (except for the last column).* Take for any $1 \leq k < n$ the corresponding column from the matrix Δ, i.e.

$$\Delta_k = \Big[-h^{n+k-1}, -h^{n+k}, \cdots,$$
$$-h^{2n-2}, \sum_{i=1}^{n-1} h^{2i-1}, -1, \cdots, -h^{k-2}\Big]^T.$$

  We have that:

$$\Delta_k^T\mathbf{h} = -\left(h^{n+k-1} + h^{n+k+1} + \cdots + h^{3n-k-3}\right)$$
$$+ \left(h^{n-k+1} + h^{n-k+3} + \cdots + h^{3n-k-3}\right)$$
$$- \left(h^{n-k+1} + h^{n-k+3} + \cdots + h^{n+k-3}\right),$$

  which is zero. This proves that all the first (n-1) columns of matrix Δ are orthogonal to vector $\mathbf{h}$.
- *The n-th column.* The n-th column of matrix Δ is also orthogonal to vector $\mathbf{h}$, which is certified by the following calculation.

$$\Delta_n^T\mathbf{h} = (h+h^3+\cdots+h^{2n-3}) - (h+h^3+\cdots+h^{2n-3}) = 0$$

Since $(Q_1 - Q_2)\mathbf{h} = \mathbf{h}(Q_1 - Q_2) = 0$, we have that $(Q_1 - Q_2)P = P(Q_1 - Q_2) = 0$. With this observation, the Fact 2 can be concluded by the following:

$$(Q_1+Q_2)(Q_1-Q_2) = (Q_1-Q_2)(Q_1+Q_2)$$
$$\Leftrightarrow (I_n - P)(Q_1-Q_2) = (Q_1-Q_2)(I_n - P)$$
$$\Leftrightarrow P(Q_1-Q_2) = (Q_1-Q_2)P.$$

This concludes the proof of Fact 2.

For the proof of Fact 3, the reasoning goes as follows: By expanding the equations in Fact 1, and canceling the common terms in both sides, we have that

$$Q_2Q_1 + Q_1Q_2 = -Q_2Q_1 - Q_1Q_2.$$

By the same operation on the equation in Fact 2, we have that

$$Q_2Q_1 - Q_1Q_2 = -Q_2Q_1 + Q_1Q_2.$$

From these two equations, we can calculate out $Q_1Q_2 = Q_2Q_1 = 0$, which concludes the proof of the Fact 3.

For the proof of the Fact 4, we need to notice that as vector $\mathbf{h}$ lies in the null space of $Q_1 - Q_2$, so we have that

$$(Q_1 - Q_2)P = 0.$$

Together with $(Q_1+Q_2)P = 0$, we have that $Q_1P = 0$ and $Q_2P = 0$. So we have that

$$Q_1 = Q_1(Q_1+Q_2+P) = Q_1^2 + Q_1Q_2 + Q_1P = Q_1^2,$$
$$Q_2 = Q_2(Q_1+Q_2+P) = Q_1Q_2 + Q_2^2 + Q_2P = Q_2^2,$$

which concludes the proof of Fact 4.

Paper III

AN ELLIPSOID BASED, TWO-STAGE SCREENING TEST FOR BPDN

Liang Dai, Kristiaan Pelckmans

Uppsala University,
Institute of Information Technology,
Division of Systems and Control, Uppsala, Sweden

ABSTRACT

Consider the Basis Pursuit De-Noising (BPDN) estimator for recovery of unknown, sparse parameters. This note presents an ellipsoid-based, two-stage screening test method which aims to reduce a-priori the dimensionality of the resulting optimization problem. The new elements of the proposed method are given by (i) using an efficient ellipsoid approximation scheme in both stages and (ii) making better use of the information which has been calculated during the first stage. A comparative experiment indicates that this procedure can lead to better overall time complexity compared to known screening tests, while *screening away* more irrelevant variables in a preprocessing stage.

1. INTRODUCTION

We will first introduce the Basis Pursuit De-Noising (BPDN) estimator briefly in the following. Let $n \in \mathbb{N}$ denotes the number of observations, and $m \gg n$ denotes the dimensionality of the problem. Given a measurement vector $\mathbf{x} \in \mathbb{R}^n$, m dictionary vectors(atoms) $\mathbf{b}_i \in \mathbb{R}^n$. Let these atoms be organized in a matrix $\mathbf{B} \in \mathbb{R}^{n\times m}$ such that the ith column of $\mathbf{B}$ equals $\mathbf{b}_i$. Assume that there is an (unknown) vector $\mathbf{w}_0 \in \mathbb{R}^m$ (which is assumed to be sparse) and a vector $\mathbf{e} \in \mathbb{R}^n$ (which represents noise), such that

$$\mathbf{x} = \mathbf{B}\mathbf{w}_0 + \mathbf{e}.$$

The task is to recover $\mathbf{w}_0$ from $\mathbf{x}$ and $\mathbf{B}$. A survey of techniques applicable to this task is given in [3, 4].

A reasonable estimate $\mathbf{w} = (w_1, \ldots, w_m)^T$ of $\mathbf{w}_0$ is given by solving the following problem for a given $\lambda > 0$:

$$\min_{\mathbf{w}\in\mathbb{R}^m} \frac{1}{2}\|\mathbf{x} - \mathbf{B}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_0, \tag{1}$$

where $\|\mathbf{w}\|_0 = \sum_{d=1}^m I(w_d \neq 0)$, with the indicator $I(z)$ equals to one iff z holds true, and zero otherwise. Here the parameter $\lambda > 0$ regulates the tradeoff between the data fit and representation complexity.

While (1) is non-smooth, non-convex, and strictly NP-hard [8], one often resorts to solving the convex BPDN which serves as tractable proxy to (1) by solving

$$\bar{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^m} \frac{1}{2}\|\mathbf{x} - \mathbf{B}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1, \tag{2}$$

where the convex L_1-norm is defined as $\|\mathbf{w}\|_1 = \sum_{d=1}^m |w_d|$. Here, the norm $\|\cdot\|_1$ is regarded as the convex envelope to the non-convex $\|\cdot\|_0$. This estimator (2) is sometimes referred to as to the Least Absolute Shrinkage and Selection Operator (LASSO) estimator. As in [1], we also assume that $\|\mathbf{x}\|_2 = 1$ and $\|\mathbf{b}_i\|_2 = 1$ for $i = 1 \ldots m$. The formulation of BPDN has found many interesting applications and theoretical results, in particular because of the facts that:

- Since the BPDN boils down to a convex optimization problem, it can be solved efficiently with well-known tools as the Interior Point Method (IPM) [5]. This is a general numerical solver for problems of convex optimization. Extensive research on this particular problem resulted in a wide variety of numerical solvers which obtain better practical as well as theoretical performance by exploiting more structure information of the problem. For an up-to-date collection of such methods, please consult[1].
- Theoretical excitement stems from the fact that recoverability(such as the support of $\mathbf{w}_0$,or some 'good' estimations of $\mathbf{w}_0$) of $\mathbf{w}_0$ can be guaranteed under certain conditions of the measurement matrix $\mathbf{B}$ (Restricted Isometry Property, Null Space Property, Spherical Section Property, etc, see [3, 4]) and the sparsity level of $\mathbf{w}_0$. Such guarantees come in different forms as surveyed in [3, 4] and citations therein.

According to the reference [1], the Lagrangian dual to problem (2) is given as follows:

$$\bar{\theta} = \arg\max_{\theta\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\left\|\theta - \frac{\mathbf{x}}{\lambda}\right\|_2^2 \qquad \text{s.t. } |\theta^T\mathbf{b}_i| \leq 1, \forall i = 1, 2, \ldots, m. \tag{3}$$

The optimal solutions $\bar{\mathbf{w}}$ to problem (2) and $\bar{\theta}$ to problem (3) are connected through eq. (4) and eq. (5). We refer the readers to the reference[1] for details.

$$\mathbf{x} = \sum_{i=1}^m \bar{w_i}\mathbf{b}_i + \lambda\bar{\theta}, \tag{4}$$

and

$$\bar{\theta}^T\mathbf{b}_i \in \begin{cases} \text{sign}(\bar{w_i}) & \text{iff } \bar{w_i} \neq 0 \\ [-1, 1] & \text{iff } \bar{w_i} = 0. \end{cases} \tag{5}$$

Define the halfspace $H(\mathbf{y})$ for $\mathbf{y} \in \mathbb{R}^n$ as

$$H(\mathbf{y}) = \left\{\mathbf{z} : \mathbf{z}^T\mathbf{y} \leq 1\right\} \subset \mathbb{R}^n.$$

Let $L(\mathbf{y})$ be the corresponding hyperplane

$$L(\mathbf{y}) = \left\{\mathbf{z} : \mathbf{z}^T\mathbf{y} = 1\right\} \subset \mathbb{R}^n.$$

The reasoning behind the construction of a screening test goes as follows. From eq. (3), (4) and (5), we can see that if $\bar{\theta}$ is not on

[1] http://ugcs.caltech.edu/˜srbecker/wiki/Category:Solvers

$L(\mathbf{b_i})$ nor $L(-\mathbf{b_i})$, then $\bar{w}_i$ will be zero. This is a crucial observation for screening tests as pointed out in [1, 2, 6]. The idea is then to make a set $Q \subset \mathbb{R}^n$ which contains $\bar{\theta}$, and check for $i = 1, \ldots, m$, whether $L(\mathbf{b}_i)$ or $L(-\mathbf{b_i})$ intersects Q or not. If for i, no intersection takes place, one can conclude that $\bar{w}_i$ is zero, and it doesn't need to be included in later stages of the optimization problem. That is, this corresponding dictionary $\mathbf{b}_i$ is *screened away* in the subsequent optimization problem.

The aim of this paper is to reduce m before actually solving (2). That is, we aim to filter out (or *screen out*) as many different columns of $\mathbf{B}$ as possible, before performing the convex optimization problem (2) completely. Such preprocessing stage could then lead to subsequent less time and memory intensive optimization procedures since m could be reduced severely. An important point is that such *screening* stage should not be too computationally involved to perform.

Some test methods have been devised already in [1, 2, 6] with different levels of effectiveness. This note introduces a two-stage ellipsoid based screening test which further improves the screening performance. This means that in total, the computational cost including the cost for the screening test and the cost for the subsequent optimization will be reduced. Our strategy is composed of two stages, which in general are:

1. Approximate the basic potential region Q for $\bar{\theta}$ with an ellipsoid, and then perform the 'intersection test'. If neither $L(\mathbf{b}_i)$ nor $L(-\mathbf{b}_i)$ intersect with this ellipsoid, then the corresponding $\bar{\mathbf{w}}_i$ is set to zero. This is similar to the tests as performed in the traditional screening tests.
2. In the second stage, a new approximation of the potential region of $\bar{\theta}$ based on the information which is obtained earlier (we only choose one halfspace which shrink the volume most, details are given in section 3). Then another round screening test is obtained based on this updated ellipsoid. Note that, this stage, the test is only carried out on those atoms with haven't been determined to be screened out in the first stage.

Our method is motivated as follows: (1) the update rule of the ellipsoid approximation is simple; (2) while performing the 'intersection test' in the first round, information can also be used for obtaining a tighter approximation of the potential region of $\bar{\theta}$; (3) the 'intersection test' in every round also requires low time cost.

We will use the following notational conventions throughout. A lower-case letter denotes a scalar, a boldface lowercase denotes a vector and a boldface capital denotes a matrix. This paper is organized as follows. Section II describes the ellipsoid related results, including the update rule, and some related geometrical results. Section III describes our algorithm in detail. Section IV gives experimental results indicating the efficacy of the method, and compares to existing approaches. Section IV concludes this paper and points towards interesting open avenues for further research.

2. ELLIPSOID RELATED RESULTS

In this section, we will give the ellipsoid update rule and some related results. These results will be used in the forming of our proposed algorithm in the following sections.

2.1. Ellipsoid Update Rule

Given a halfspace represented as

$$H_h(\mathbf{x}_p, \mathbf{g}) = \left\{ \mathbf{z} \in \mathbb{R}^n : \mathbf{g}^T(\mathbf{z} - \mathbf{x}_p) + h \leq 0 \right\},$$

and the corresponding hyperplane as

$$L_h(\mathbf{x}_p, \mathbf{g}) = \left\{ \mathbf{z} \in \mathbb{R}^n : \mathbf{g}^T(\mathbf{z} - \mathbf{x}_p) + h = 0 \right\},$$

where $h \geq 0, \mathbf{g}, \mathbf{x}_p \in \mathbb{R}^n$ are given, and an ellipsoid

$$E(\mathbf{x}_p, \mathbf{P}_p) = \{\mathbf{z} \in \mathbb{R}^n : (\mathbf{z} - \mathbf{x}_p)^T \mathbf{P}_p^{-1} (\mathbf{z} - \mathbf{x}_p) \leq 1\},$$

where $\mathbf{P}_p \in \mathbb{R}^{n \times n}$ and $\mathbf{P}_p \succeq 0$. Then the ellipsoid with the minimum volume which contains the intersection of $H_h(\mathbf{x}_p, \mathbf{g})$ and $E(\mathbf{P}_p, \mathbf{x}_p)$ could be represented as

$$E(\mathbf{x}_u, \mathbf{P}_u) = \left\{ \mathbf{z} \in \mathbb{R}^n : (\mathbf{z} - \mathbf{x}_u)^T \mathbf{P}_u^{-1} (\mathbf{z} - \mathbf{x}_u) \leq 1 \right\}.$$

Here we define

$$\begin{cases} \mathbf{x}_u = \mathbf{x}_p - \frac{1+\alpha n}{n+1} \mathbf{P}_p \bar{\mathbf{g}} \\ \mathbf{P}_u = \frac{n^2(1-\alpha^2)}{n^2-1} \left(\mathbf{P}_p - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)} \mathbf{P}_p \bar{\mathbf{g}} \bar{\mathbf{g}}^T \mathbf{P}_p \right), \end{cases} \quad (6)$$

where $\bar{\mathbf{g}} = \frac{\mathbf{g}}{\sqrt{\mathbf{g}^T \mathbf{P}_p \mathbf{g}}}$ and $\alpha = \frac{h}{\sqrt{\mathbf{g}^T \mathbf{P}_p \mathbf{g}}}$.

The derivation of this update rule is given in [5]. It has found main application for bounding convex sets as in the membership set method [7] as commonly used in system identification. It also plays an important historic role in finding polynomial time solver for solving Linear Programming (LP) problems [5]. In the following, we will give a rule which decides if a hyperplane intersects with an ellipsoid or not.

2.2. Intersection test

Lemma 1 *Given $h > 0, \mathbf{x}_p \in \mathbb{R}^n, \mathbf{g} \in \mathbb{R}^n, \mathbf{P}_p \in \mathbb{R}^{n \times n}$. Define*

$$\alpha = \frac{h}{\sqrt{\mathbf{g}^T \mathbf{P}_p \mathbf{g}}}.$$

If $|\alpha| > 1$, then the intersection of the hyperplane $L_h(\mathbf{x}_p, \mathbf{g})$ with the ellipsoid $E(\mathbf{x}_p, \mathbf{P}_p)$ is empty.

Proof 1 *From a geometric viewpoint, this lemma follows by the following reasoning. Since*

$$\{\mathbf{z} : \mathbf{g}^T(\mathbf{z} - \mathbf{x}_p) + h = 0\} \\ \cap \{\mathbf{z} : (\mathbf{z} - \mathbf{x}_p)^T \mathbf{P}_p^{-1} (\mathbf{z} - \mathbf{x}_p) \leq 1\} = \varnothing, \quad (7)$$

holds if and only if

$$\{\mathbf{z} : \mathbf{g}^T \mathbf{P}_p^{\frac{1}{2}} \mathbf{z} + h = 0\} \cap \{\mathbf{z} : \mathbf{z}^T \mathbf{z} \leq 1\} = \varnothing. \quad (8)$$

Notice that the distance from 0 to the hyperplane given as $\{\mathbf{z} : \mathbf{g}^T \mathbf{P}_p^{\frac{1}{2}} \mathbf{z} + h = 0\}$ is equal to

$$\frac{|h|}{\sqrt{\mathbf{g}^T \mathbf{P}_p \mathbf{g}}}.$$

Hence it follows that if $|\alpha| > 1$, the intersection will be empty. This concludes the proof.

In the following, we will characterize how much of the volume will be shrunken by the update.

2.3. Shrinkage of the Volume

Lemma 2 *Define*

$$\alpha = \frac{h}{\sqrt{\mathbf{g}^T \mathbf{P}_p \mathbf{g}}}.$$

If $0 \le \alpha \le 1$, then one has that after the ellipsoid update as depicted in eq. (6), the volumes are shrunken as:

$$\frac{\mathrm{vol}\,(E(\mathbf{x}_p, \mathbf{P}_p))}{\mathrm{vol}\,(E(\mathbf{x}_u, \mathbf{P}_u))} = \frac{n^n}{(1+n)(n^2-1)^{\frac{n-1}{2}}}(1-\alpha)(1-\alpha^2)^{\frac{n-1}{2}}.$$

Proof 2 *We have that*

$$\begin{aligned}
\frac{\mathrm{vol}^2\,(E(\mathbf{x}_p, \mathbf{P}_p))}{\mathrm{vol}^2\,(E(\mathbf{x}_u, \mathbf{P}_u))} &= \frac{|\mathbf{P}_u|}{|\mathbf{P}_p|} \qquad (9)\\
&= \frac{\left|\frac{n^2(1-\alpha^2)}{n^2-1}\left(\mathbf{P}_p - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)}\mathbf{P}_p \bar{g} \bar{g}^T \mathbf{P}_p\right)\right|}{|\mathbf{P}_p|}\\
&= \frac{|\frac{n^2(1-\alpha^2)}{n^2-1}\mathbf{P}_p^{\frac{1}{2}}\left(I - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)}\mathbf{P}_p^{\frac{1}{2}}\bar{\mathbf{g}}\bar{\mathbf{g}}^T\mathbf{P}_p^{\frac{1}{2}}\right)\mathbf{P}_p^{\frac{1}{2}}|}{|\mathbf{P}_p|}\\
&= \left(\frac{n^2(1-\alpha^2)}{n^2-1}\right)^n \left|I - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)}\mathbf{P}_p^{\frac{1}{2}}\bar{\mathbf{g}}\bar{\mathbf{g}}^T\mathbf{P}_p^{\frac{1}{2}}\right|\\
&= \left(\frac{n^2(1-\alpha^2)}{n^2-1}\right)^n \left|1 - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)}\bar{\mathbf{g}}^T\mathbf{P}_p\bar{\mathbf{g}}\right|\\
&= \left(\frac{n^2(1-\alpha^2)}{n^2-1}\right)^n \left|1 - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)}\right|\\
&= \frac{n^{2n}}{(1+n)^2(n^2-1)^{n-1}}(1-\alpha)^2(1-\alpha^2)^{n-1},
\end{aligned}$$

as desired.

Remark 1 *From both lemmas, we see that α plays a remarkable role. This factor not only let us decide whether the hyperplane will intersect with the ellipsoid or not, but also can help to characterize how much the volume of the updated ellipsoid will be shrunken. Especially, from Lemma 2, we can see that the larger α is, the more the volume of the updated ellipsoid will shrink.*

3. ALGORITHM

Using the same notations as in [1, 2], we define $\lambda_{\max} = \max_i |\mathbf{x}^T \mathbf{b}_i|$. The vector $\mathbf{b}_*$ is defined so as to satisfy $\lambda_{\max} = \mathbf{x}^T \mathbf{b}_*$. It can be verified that $\mathbf{x}/\lambda_{\max}$ is a feasible solution to the dual (3). In order to avoid the trivial case, we assume that $\lambda < \lambda_{\max}$ as in [1, 2]. Define the region $R_1 \subset \mathbb{R}^n$ as

$$R_1 = \{\theta : \mathbf{b}_*^T \theta \le 1\} \bigcap \left\{\theta : \|\theta - \mathbf{x}/\lambda\|_2 \le \sqrt{1/\lambda - 1/\lambda_{\max}}\right\}.$$

We can see that R_1 is a a region where $\bar{\theta}$ will locate in. This region has been referred to as a 'dome' in [2] (an intersection of a halfspace and a ball). As discussed before, if for $i \in \{1, \ldots, m\}$ neither of the hyperplanes $L(\mathbf{b}_i)$ or $L(-\mathbf{b}_i)$ intersects with R_1, then $\bar{\mathbf{w}}_i$ has to equal zero. In the references, the authors bound R_1 with different balls (different center and radius), which led them to convenient yet effective test as the 'SAFE/ST1','ST2','ST3' test [1, 6] or the 'dome' test[2]. The 'dome' test is considered to be the most effective one in the sense of its effectiveness (the number of irrelevant atoms screened out) and low computation cost. Hence, in the experiment part we will mainly compare the proposed screening test with the 'dome' test. As stated briefly in the previous part, the proposed test will consist of two stages. The formal and precise descriptions are given as follows.

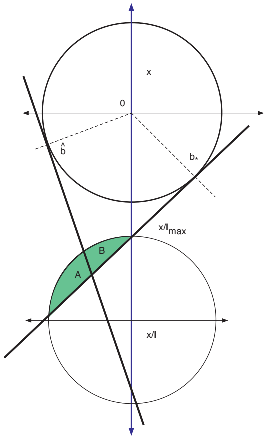


Fig. 1. *A schematic explanation of the idea behind the 2-stage, ellipsoid-based screening test when $n = 2$. The unit circle S^{n-1} at the top indicates the unit sphere including the vectors $\{\mathbf{b}_i, -\mathbf{b}_i\}_{i=1}^m$ and $\mathbf{x}$. The circle at the bottom indicates the set of vectors θ with distance to $\frac{\mathbf{x}}{\lambda}$ equal to $\|\frac{\mathbf{x}}{\lambda} - \frac{\mathbf{x}}{\lambda_{\max}}\|_2$. Hence, the optimum $\bar{\theta}$ of problem (3) will be inside this circle. The solid lines indicate the hyperplanes $L(\mathbf{b}_*)$ and $L(\hat{\mathbf{b}})$ as explained in Subsections 3.1 and 3.2. The first stage of the test computes an ellipsoid estimation of the dome $R_1 = A \bigcup B$ which contains $\bar{\theta}$. Then a first round of ellipsoid based screening is applied, and many of the irrelevant dictionary atoms will be screened away. As a byproduct, those calculations give the halfspace $H(\hat{\mathbf{b}})$ which shrinks the volume of the ellipsoid estimation the most. So, the potential region of $\bar{\theta}$ will be shrunken from R_1 to region B. In the second stage of the test, screening is applied to the remaining dictionary atoms using the updated ellipsoid.*

3.1. Stage 1

1. Compute the minimum volume ellipsoid containing R_1. This calculation is a direct consequence of the update rule described in section 2, in which $\mathbf{x}_p = \frac{\mathbf{x}}{\lambda}$, $\mathbf{g} = \mathbf{b}_*$, $h = \frac{\lambda_{\max}}{\lambda} - 1$, and $\mathbf{P}_p = (\frac{1}{\lambda} - \frac{1}{\lambda_{\max}})\mathbf{I}_n$. Denote the updated ellipsoid as

   $$E_1(\mathbf{x}_1, \mathbf{P}_1) = \{\mathbf{z} : (\mathbf{z} - \mathbf{x}_1)^T \mathbf{P}_1^{-1}(\mathbf{z} - \mathbf{x}_1) \le 1\},$$

   where

   $$\begin{cases} \mathbf{x}_1 = \mathbf{x}_p - \frac{1+\alpha n}{n+1}\mathbf{P}_p\bar{\mathbf{g}} \\ \mathbf{P}_1 = \frac{n^2(1-\alpha^2)}{n^2-1}\left(\mathbf{P}_p - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)}\mathbf{P}_p\bar{\mathbf{g}}\bar{\mathbf{g}}^T\mathbf{P}_p\right), \end{cases} \qquad (10)$$

   in which $\bar{\mathbf{g}} = \frac{\mathbf{g}}{\sqrt{\mathbf{g}^T\mathbf{P}_p\mathbf{g}}}$ and $\alpha = \frac{h}{\sqrt{\mathbf{g}^T\mathbf{P}_p\mathbf{g}}}$.

2. Test for any $i = 1, \ldots, m$ whether $L(\mathbf{b}_i)$ or $L(-\mathbf{b}_i)$ intersect with $E_1(\mathbf{x}_1, \mathbf{P}_1)$ or not. If both do not intersect, then set

$\bar{\mathbf{w}}_i = 0$. Formally, calculate

$$\alpha_i^+ = \frac{\mathbf{b}_i^T \mathbf{x_1} - 1}{\sqrt{\mathbf{b}_i^T P_1 \mathbf{b}_i}},$$

and

$$\alpha_i^- = \frac{-\mathbf{b}_i^T \mathbf{x_1} - 1}{\sqrt{\mathbf{b}_i^T P_1 \mathbf{b}_i}}.$$

If $|\alpha_i^+| > 1$ and $|\alpha_i^-| > 1$ hold together, or equivalently if

$$\sqrt{\mathbf{b}_i^T \mathbf{P}_1 \mathbf{b}_i} < \min\{|\mathbf{b}_i^T \mathbf{x_1} + 1|, |\mathbf{b}_i^T \mathbf{x_1} - 1|\}, \quad (11)$$

then set $\bar{\mathbf{w}}_i = 0$.

Remark 2 *Eq. (11) is a direct application of Lemma 1 in Section 2.*

Stage 1 gives an ellipsoid approximation to the 'dome' area R_1. As we have seen in the previous section, the proposed screening test is also convenient to compute as described above. An interesting fact is that, while we are doing the screening, if eq. (11) does not hold, one has the fact that the corresponding halfspace intersects with ellipsoid $E_1(\mathbf{x}_1, \mathbf{P}_1)$. Since α_i^- and α_i^- have been calculated in hand, we see from Lemma 2 that α_i^-, α_i^+ actually also indicate the volume shrinkage of the ellipsoid approximation of the intersection. The lager they are, the more the volume will shrink. This motivates the next stage which causes no extra significant computational overhead.

3.2. Stage 2

1. Choose the the maximum value $\hat{\alpha}$ from $\{\alpha_i^+, \alpha_i^-\}_{i=1}^m$ which satisfies $0 < \hat{\alpha} < 1$. Denote the corresponding halfspace as $H(\hat{\mathbf{b}})$ and the hyperplane as $L(\hat{\mathbf{b}})$. In the ellipsoid update rules as in eq. (6), let $\mathbf{g} = \hat{\mathbf{b}}$ and $h = \hat{\mathbf{b}}^T \mathbf{x}_1 - 1$ in order to compute the updated ellipsoid

   $$E_2(\mathbf{x}_2, \mathbf{P}_2) = \{\mathbf{z} : (\mathbf{z} - \mathbf{x}_2)^T \mathbf{P}_2^{-1} (\mathbf{z} - \mathbf{x}_2) \leq 1\},$$

   where

   $$\begin{cases} \mathbf{x}_2 = \mathbf{x}_1 - \frac{1+\alpha n}{n+1} \mathbf{P}_1 \bar{\mathbf{g}} \\ \mathbf{P}_2 = \frac{n^2(1-\alpha^2)}{n^2-1} \left(\mathbf{P}_1 - \frac{2(1+\alpha n)}{(n+1)(\alpha+1)} \mathbf{P}_1 \bar{\mathbf{g}} \bar{\mathbf{g}}^T \mathbf{P}_1 \right) \end{cases}$$

   and in which $\bar{\mathbf{g}} = \frac{\mathbf{g}}{\sqrt{\mathbf{g}^T \mathbf{P}_1 \mathbf{g}}}$ and $\alpha = \frac{h}{\sqrt{\mathbf{g}^T \mathbf{P}_1 \mathbf{g}}}$.

2. For $i = 1, \ldots, m$, if $\bar{w}_i$ is not screened away yet during the first stage, test wether

   $$\sqrt{\mathbf{b}_i^T \mathbf{P}_2 \mathbf{b}_i} < \min\{|\mathbf{b}_i^T \mathbf{x_2} + 1|, |\mathbf{b}_i^T \mathbf{x_2} - 1|\}. \quad (12)$$

   If this holds, then set $\bar{\mathbf{w}}_i = 0$.

4. ILLUSTRATIVE EXAMPLES

This part describes an example which indicates the efficacy of the proposed screening test, and compares result with earlier proposed screening tests. In our example, we choose the dictionary atoms $\{\mathbf{b}_i\}_{i=1}^m$ sampled randomly from a unit normally distributed random variable, and then normalize each of them to norm one. we generate the normalized vector $\mathbf{x}$ in the same way. In this example, we let $n = 10$ and $m = 200$. Estimation problems of this size are typical in applications of BPDN, while the relative low-dimensional nature will already indicate the benefit of the proposed technique. The displayed figures are obtained by averaging out results over 50 randomizations of the experiment. In Fig. 2, results of different screening tests are given:

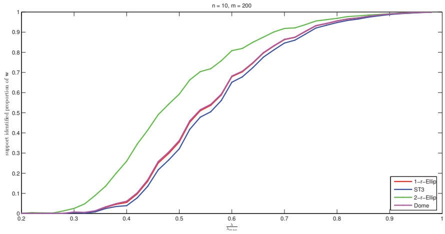


Fig. 2. *Performance of the different screening test methods, including the 'Dome test', the 'ST3 test', the '1-stage ellipsoid test', '2-stage ellipsoid test'. The x-axis represents $\frac{\lambda}{\lambda_{\max}}$, the y-axis represents the proportion of the number of the screened out zero elements in $\bar{\mathbf{w}}$. This result illustrates a significant benefit of the proposed 2-stage screening test for appropriate range of $\lambda/\lambda_{\max}$.*

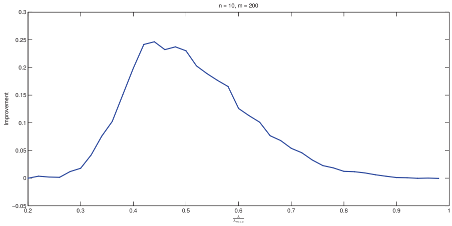


Fig. 3. *Improvement of the performance of the '2-stage ellipsoid test' and the 'Dome test'. The x-axis represents $\frac{\lambda}{\lambda_{\max}}$, the y-axis represents the ratio of proportions of the number of the screened out zero elements in $\bar{\mathbf{w}}$. Larger values indicate less remaining dictionary elements after the '2-stage, ellipsoid test'. This plot indicates that the present test can have a significant gain in terms of number of screened out dictionary elements for an appropriate range of $\frac{\lambda}{\lambda_{\max}}$.*

1. the ST3 method [1], in Fig. 2, with the tag 'ST3';
2. the 1-stage ellipsoid method as derived in Subsection 3.1 (only performing the first stage), in Fig. 2, with the tag '1-r-Ellip';
3. the Dome test method [2], in Fig. 2, with the tag 'Dome';
4. the 2-stage ellipsoid method as proposed in Section 3 (including both stages), in Fig. 2, with the tag '2-r-Ellip'.

From these figures, we observe the following:

1. When the ratio $\frac{\lambda}{\lambda_{\max}}$ is relatively small (in this example less than 0.3), then all the screening test methods are relatively ineffective. But in other words, this phenomena is reasonable. If $\lambda_{\max}$ is fixed, when $\frac{\lambda}{\lambda_{\max}}$ is small (which means that λ is small), then the dome R_1 will become very large, and all hyperplanes $\{L(\mathbf{b}_i), L(-\mathbf{b}_i)\}_i$ could be expected to intersect the dome with more chance. In the extremal case that $\lambda \to 0$, any screening test would do poor, meaning that such tests cannot be applied straightforwardly to the noiseless Basis Pursuit

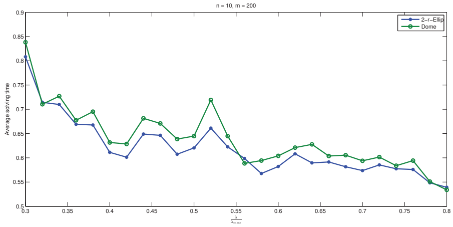


Fig. 4. *Comparison of the average time cost when 'screening with the Dome test, and solving the reduced BPDN', and 'screening with the proposed 2-stage, ellipsoid test, and solving the reduced BPDN', The x-axis represents $\frac{\lambda}{\lambda_{\max}}$, the y-axis represents the time needed to solve a corresponding complete problem. This result indicates that the improved screening capacity of the 2-stage ellipsoid screening test does not result in computational overheads, and may well lead to computational speedups.*

(BP) case.

2. When $\frac{\lambda}{\lambda_{\max}}$ is between 0.35 to 0.8, we see that the 'Dome test' and the '1-stage ellipsoid test' perform quite similar (their performance curve nearly overlapping) to each other, but do slightly better than 'ST3 test'. However, the 2-stage ellipsoid test outperforms those as more irrelevant variables are screened away, the quantitative improvement can be seen from Fig. 3. This phenomena means that, in this case, it's better to use the 2-stage ellipsoid method to do screening. Here, we need to notice that the time cost for the '2-stage-ellipsoid test' is also quite low. Fig. 4, displays the time cost for solving the same BPDN problem with 'Dome test' for screening and '2-stage ellipsoid test' for screening. After screening, we solve the reduced dimension BPDN problem with 'cvx' [5], using the internal Sedumi solver. We can see the time-saving of adopting the '2-stage-ellipsoid test' method for screening.
3. When $\frac{\lambda}{\lambda_{\max}}$ is larger than 0.8, it appears that all the methods give very similar performance. This is due to the fact that in this case the dome area is relatively small, and most of the hyperplanes $\{L(\mathbf{b}_i), L(-\mathbf{b}_i)\}_i$ will not intersect this area (which means that most coefficients of the solution are zero for such λ).

Again, note that by construction the screening tests are conservative, that is, they cannot screen variables away which would be nonzero in the final solution. Or, no performance can be lost, the screening stage can only be beneficial since the resulting optimization problem has smaller dimensionality.

5. CONCLUSION

This note presented an improvement of a screening test method for the BPDN problem. The motivation is based on an ellipsoid approximation of the potential region for $\bar{\theta}$, while the involved quantities are found to be useful in the second stage. This second stage leads to improved screening capabilities, only requring quantities which were computed in the first stage anyway. Through simulations it is found that such screening test is most effective when the ratio $\frac{\lambda}{\lambda_{\max}}$ is moderate. The comparative experiment shows that, the proposed '2-stage ellipsoid' method results in both effectiveness (more irrelevant dictionary atoms are screened away) and efficiency (the time cost for solving the whole optimization problem is reduced) improvement over the state-of-art screening test method.

However, the following questions remain open: (1) Starting with a feasible point $\frac{\mathbf{x}}{\lambda_{\max}}$, the present approach uses an initial potential region for $\bar{\theta}$ which is the 'dome' region R_1. Can we find a better starting feasible point in order to make the initial 'dome' region more accurate? (2) Can we find a way to generalize the method (including the 'SAFE/ST1', 'ST2', 'ST3', 'Dome test', and the proposed method) to the the case where $1/\lambda \rightarrow \infty$ as in Basis Pursuit (BP)?

6. REFERENCES

[1] Z. J. Xiang, H. Xu, P. J. Ramadge, Learning sparse representations of high dimensional data on large scale dictionaries, NIPS 2011.

[2] Z. J. Xiang, P. J. Ramadge, Fast lasso screening tests based on correlations, ICASSP 2012.

[3] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, 2010.

[4] M.A. Davenport, M.F. Duarte, Y.C. Eldar, and G. Kutyniok. Introduction to Compressed Sensing: Theory and Applications, Cambridge University Press, 2011.

[5] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[6] L.E. Ghaoui, V. Viallon, T. Rabbani, Safe feature elimination in sparse supervised learning, Arxiv preprint arXiv:1009.3515, 2010.

[7] R. Kosut, M. Lau, S. Boyd, Set-membership identification of systems with parametric and nonparametric uncertainty, IEEE Transactions on Automatic Control, 37(7):929-941, July 1992.

[8] D. Ge, X. Jiang, Y. Ye. A Note on Complexity of Lp Minimization, Mathematical Programming, Volume 129, Number 2, 285-299, 2010.

Paper IV

On the Randomized Kaczmarz Algorithm

Liang Dai, Mojtaba Soltanalian, Kristiaan Pelckmans
Department of Information Technology, Uppsala University, Sweden.

***Abstract*—The Randomized Kaczmarz Algorithm is a randomized method which aims at solving a consistent system of over determined linear equations. This note discusses how to find an optimized randomization scheme for this algorithm, which is related to the question raised by [2]. Illustrative experiments are conducted to support the findings.**

***Index Terms*—Randomized Kaczmarz Algorithm, Convex Optimization, Linear System Solver**

I. Problem Statement

In this note, we discuss the Kaczmarz Algorithm (KA)[4], in particular the Randomized Kaczmarz Algorithm (RKA) [1], to find the unknown vector $\mathbf{x} \in \mathbb{R}^n$ of the following set of *consistent* linear equations:

$$A\mathbf{x} = \mathbf{b}, \tag{1}$$

where matrix $A \in \mathbb{R}^{m \times n}, m \geq n$, is of full column rank, and $\mathbf{b} \in \mathbb{R}^m$. Since [4], the KA has been applied to different fields and many new developments are reported. For instance, in [6], the author study the RKA when applied to the case of the linear systems are inconsistent. In [5], RKA is applied to the Computer Tomography. In [7], the authors present a method to accelerate the convergence of the RKA with the application of the Johnson-Lindenstrauss Lemma. In [8], the authors analyze the almost sure convergence of the RKA when proper stochastic properties of matrix A are introduced. In [9], the authors presented a practically more efficient approach to solve the linear systems by projecting to different blocks of rows of A, and a randomization technique is applied to find a good partition of the rows.

The KA can be described as follows. Let us define the hyperplane H_i as:

$$H_i = \{\mathbf{x} | \mathbf{a}_i^T \mathbf{x} = b_i\},$$

where the i-th row of A is denoted as $\mathbf{a}_i^T$ and the i-th element of $\mathbf{b}$ is denoted as b_i. Geometrically, the solution of (1) can be thought as the intersection of all hyperplanes $\{H_i\}_{i=1}^m$, and the KA seeks to find the solution by successively projecting to the hyperplanes from an initial approximation $\mathbf{x}_0$. The process is mathematically written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{b_i - \mathbf{a}_i^T \mathbf{x}_k}{\|\mathbf{a}_i\|_2^2} \mathbf{a}_i, \tag{2}$$

where $i = mod(k, m) + 1$. Here we use the Matlab convention $mod(\cdot, \cdot)$ to denote the *modulus after the division* operation. Fig. 1 illustrates the algorithm in a low dimensional case.

The key difference between the RKA and the KA is that RKA chooses the rows following a specified probability distribution. More precisely, the probability for selecting $\mathbf{a}_i^T$ is given as $\frac{\|\mathbf{a}_i\|_2^2}{\|A\|_F^2}$. Note that this probability is proportional to the row norms.

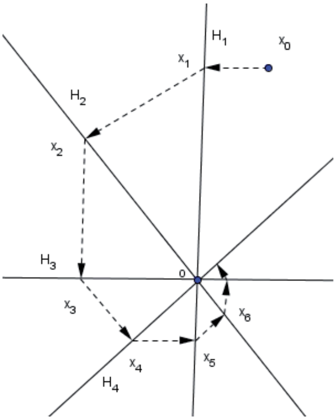


Fig. 1. A geometrical interpretation of the algorithm. Here, $m = 4$ and $n = 2$, and the solution $\mathbf{x}$ to $A\mathbf{x} = \mathbf{b}$ is represented by the point o. We can see that by this sequence of projections, $\mathbf{x}_k$ converges to the solution.

Although the KA is simple to state, its rate of convergence is still not completely explored. While for the RKA, with the predescribed choice of the probability distribution, the following convergence result is set up in [1]:

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}\|_2^2) \leq (1 - \kappa(A)^{-2})^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2, \tag{3}$$

in which $\kappa(A) = \|A\|_F \|A^\dagger\|_2$, and with $\mathbb{E}$ concerning the random choices of rows in the RKA.

However, it is argued in [2] that *'Assigning probabilities corresponding to the row norms is in general certainly not optimal'*. In the follows, we will try to find an optimized probability distribution for selecting the rows from A, so that a better performance can be obtained. The distribution vector is derived by minimizing an upper bound to the convergence rate which can be obtained by solving a convex optimization problem.

This note is organized as follows. The next section discusses the main results; In section 3, we discuss how to approximately solve the arising Semi-Definite-Programming (SDP) problem with smaller computational cost; In section 4, illustrative experiments will be conducted to verify the findings; Finally, we draw some conclusions in section 5.

II. Optimized RKA

In the following, for convenience of discussion, we will introduce a new matrix $B \in \mathbb{R}^{m \times n}$. Let $\mathbf{b}_i^T$ denote the i-th row of B, which is defined as

$$\mathbf{b}_i = \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}, \forall i = 1, \cdots, m, \tag{4}$$

i.e. every row of the matrix B is a normalized version of the corresponding row of matrix A.

Let $\mathbf{p} \in \mathbb{R}^m$ be a probability distribution vector (i.e. $\mathbf{p} \geq 0$, $\mathbf{1}^T\mathbf{p} = 1$) for selecting the rows in the RKA method and let p_i denote the ith element of $\mathbf{p}$.

Assume that currently we have $\mathbf{x}_{k-1}$, and based on $\mathbf{x}_{k-1}$, the next approximation $\mathbf{x}_k$ is given by (2), in which the index i is chosen randomly according to $\mathbf{p}$. By the property of the projection operation, we have that

$$\|\mathbf{x}_k - \mathbf{x}\|_2^2 = \|\mathbf{x}_{k-1} - \mathbf{x}\|_2^2 \sin^2(\alpha_i), \tag{5}$$

in which α_i denotes the angle between $\mathbf{x}_{k-1} - \mathbf{x}$ and the selected $\mathbf{b}_i$, i.e. the normal direction of the chosen hyperplane.

Based on the previous formula, we have that

$$\mathbb{E}_{\cdot|\mathbf{x}_{k-1}}(\|\mathbf{x}_k - \mathbf{x}\|_2^2) = \|\mathbf{x}_{k-1} - \mathbf{x}\|_2^2 \sum_{i=1}^{m} p_i \sin^2(\alpha_i), \tag{6}$$

in which $\mathbb{E}_{\cdot|\mathbf{x}_{k-1}}$ denotes the expectation operator conditioned on $\mathbf{x}_{k-1}$. It follows that:

$$\sum_{i=1}^{m} p_i \sin^2(\alpha_i) \leq \sup_{\mathbf{y}\in\mathbb{R}^n, \mathbf{y}\neq\mathbf{0}} \sum_{i=1}^{m} p_i \sin^2(\beta_i) \triangleq \Omega_1, \tag{7}$$

and

$$\sum_{i=1}^{m} p_i \sin^2(\alpha_i) \geq \inf_{\mathbf{y}\in\mathbb{R}^n, \mathbf{y}\neq\mathbf{0}} \sum_{i=1}^{m} p_i \sin^2(\beta_i) \triangleq \Omega_2, \tag{8}$$

in which β_i denotes the angle between $\mathbf{y}$ and $\mathbf{b}_i$.

Based on the relations in (6), (7) and (8), we have that

$$\mathbb{E}_{\cdot|\mathbf{x}_{k-1}}(\|\mathbf{x}_k - \mathbf{x}\|_2^2) \leq \Omega_1 \|\mathbf{x}_{k-1} - \mathbf{x}\|_2^2, \tag{9}$$

and

$$\mathbb{E}_{\cdot|\mathbf{x}_{k-1}}(\|\mathbf{x}_k - \mathbf{x}\|_2^2) \geq \Omega_2 \|\mathbf{x}_{k-1} - \mathbf{x}\|_2^2. \tag{10}$$

By iterating the relations given in eq. (9) and eq. (10), the following results follow.

Theorem 1: We have that

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}\|_2^2) \leq \Omega_1^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2, \tag{11}$$

and

$$\mathbb{E}(\|\mathbf{x}_k - \mathbf{x}\|_2^2) \geq \Omega_2^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2, \tag{12}$$

in which the expectations are taken with respect to all the random choices of the rows up to time k.

Remark 1: Note that $\Omega_1 < 1$ can be guaranteed if $\mathbf{p}$ is a strictly positive vector. This can be proven by a contradiction argument as follows. If $\Omega_1 = 1$, and since $\sin^2(\beta_i) \leq 1$ for any i and $\sum_{i=1}^m p_i = 1$, we have that $\sin^2(\beta_i) = 1$, i.e. $\cos(\beta_i) = 0$ holds for all i. Considering that $rank(A) = n$, i.e. $rank(B) = n$, hence $\mathbf{x}_k - \mathbf{x}$ can not be orthogonal to the vectors $\{\mathbf{b}_i\}_{i=1}^m$, and the result follows. Based on this observation, we can see that exponential convergence in expectation can be obtained by a wide range of probability distribution vectors. This finding extends the result in [1], which only guarantees the exponential convergence for a given specific choice of the probability distribution vector. ■

According to Theorem 1, in order to get a better performance, we need to find a probability distribution vector, such that Ω_1 can be made as small as possible. When the optimized Ω_1 is obtained, we can also have a lower bound to the convergence speed of the RKA based on Ω_2. In the following, we will first derive a closed form for Ω_1 and Ω_2, and then introduce a convex optimization problem to calculate the probability distribution vector $\hat{\mathbf{p}}$ which minimizes Ω_1.

Notice that

$$\sum_{i=1}^{m} p_i \sin^2(\beta_i) = 1 - \sum_{i=1}^{m} p_i \cos^2(\beta_i),$$

so in order to minimize Ω_1, equivalently, we can maximize the following

$$\inf_{\mathbf{y}\in\mathbb{R}^n, \mathbf{y}\neq\mathbf{0}} \sum_{i=1}^{m} p_i \cos^2(\beta_i).$$

If we restrict $\|\mathbf{y}\|_2 = 1$, then we have that

$$\cos^2(\beta_i) = \mathbf{y}^T \mathbf{b}_i \mathbf{b}_i^T \mathbf{y}.$$

Therefore

$$\sum_{i=1}^{m} p_i \cos^2(\beta_i) = \sum_{i=1}^{m} p_i \mathbf{y}^T \mathbf{b}_i \mathbf{b}_i^T \mathbf{y},$$

where the right hand side equals

$$\mathbf{y}^T B^T \operatorname{diag}(\mathbf{p}) B \mathbf{y}.$$

Notice that

$$\min_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|_2=1} \mathbf{y}^T B^T \operatorname{diag}(\mathbf{p}) B \mathbf{y} = \sigma_n(B^T \operatorname{diag}(\mathbf{p}) B),$$

in which $\sigma_n(\cdot)$ denotes the smallest singular value of the matrix. The previous discussions can be summarized as:

Theorem 2:

$$\Omega_1 = 1 - \sigma_n(B^T \operatorname{diag}(\mathbf{p}) B). \tag{13}$$

Similarly, we have that:

Corollary 1:

$$\Omega_2 = 1 - \sigma_1(B^T \operatorname{diag}(\mathbf{p}) B), \tag{14}$$

in which $\sigma_1(\cdot)$ denotes the maximal singular value of the matrix.

Notice that minimizing Ω_1 is equivalent to maximizing $\sigma_n(B^T \operatorname{diag}(\mathbf{p}) B)$, then we can solve the following problem instead:

$$\begin{aligned} \max_{\mathbf{p}\in\mathbb{R}^m} \quad & \sigma_n(B^T \operatorname{diag}(\mathbf{p}) B) \\ s.t. \quad & \mathbf{1}^T\mathbf{p} = 1; \\ & p_i \geq 0,\ i = 1, \ldots, m. \end{aligned} \tag{15}$$

This problem can be rewritten as the following SDP problem, in which $\hat{t}$ denotes the optimized σ_n and $\hat{\mathbf{p}}$ denotes the corresponding probability distribution vector:

$$
\begin{aligned}
(\hat{\mathbf{p}}, \hat{t}) = & \operatorname*{arg\,max}_{\mathbf{p}\in\mathbb{R}^m, t\in\mathbf{R}} \; t \qquad (16)\\
s.t. \quad & \mathbf{1}^T\mathbf{p} = 1;\\
& p_i \geq 0,\; i = 1,\ldots,m;\\
& B^T \operatorname{diag}(\mathbf{p}) B - tI_n \succeq 0.
\end{aligned}
$$

After solving the optimization problem of (16), $\hat{\mathbf{p}}$ is applied to the RKA to select the rows. Such a scheme will be abbreviated as ORKA in the following.

Remark 2: There exist cases such that $\Omega_1 = \Omega_2$, i.e. there exists a vector $\mathbf{p}$, such that

$$\sigma_1(B^T \operatorname{diag}(\mathbf{p})B) = \sigma_n(B^T \operatorname{diag}(\mathbf{p})B),$$

i.e. $B^T \operatorname{diag}(\mathbf{p})B = \frac{1}{n}I_n$. In such cases, $\Omega_1 = \Omega_2 = 1 - \frac{1}{n}$, and the optimized probability distribution obtained by solving eq. (16) is the same as suggested in [1]. It can be verified that when the columns of A are orthogonal and of equal norm, then such property will hold. ■

Remark 3: The optimization problem (16) can also be formulated as

$$
\begin{aligned}
\hat{\mathbf{q}} = & \operatorname*{arg\,min}_{\mathbf{q}\in\mathbb{R}^m} \; \mathbf{1}^T\mathbf{q} \qquad (17)\\
s.t. \; & B^T \operatorname{diag}(\mathbf{q})B - I_n \succeq 0;\\
& q_i \geq 0,\; i = 1,\ldots,m.
\end{aligned}
$$

in the sense that $\hat{t} = \frac{1}{\mathbf{1}^T\hat{\mathbf{q}}}$ and $\hat{\mathbf{p}} = \hat{t}\hat{\mathbf{q}}$.

Since $\mathbf{q}$ in (17) is nonnegative, one has that $\mathbf{1}^T\mathbf{q} = \|\mathbf{q}\|_1$. It is known that the l_1 norm minimization problem is likely to return sparse solutions[11], which gives that $\hat{\mathbf{q}}$ is likely to be sparse. In the experiment section, we will also illustrate this phenomena. ■

Next, we discuss the relation between the ORKA and the RKA. It is obvious that the projection operations in (2) depend only on the corresponding normal vectors of the hyperplanes $\{H_i\}_{i=1}^m$, so we can optimize $\kappa(A) = \|A\|_F\|A^\dagger\|_2$ subject to the norms of the rows of matrix A. The optimization problem is given as

$$\min_{\{\|\mathbf{a}_i\|_2\}_{i=1}^m} \quad \kappa(A) = \|A\|_F\|A^\dagger\|_2.$$

Define $\mathbf{q} \in \mathbb{R}^m$, in which $q_i = \|\mathbf{a}_i\|_2^2$ for $i = 1\cdots m$. Then the previous optimization problem can be written as

$$\min_{\mathbf{q}} \quad \frac{\sqrt{\mathbf{1}^T\mathbf{q}}}{\sigma_n(A)}.$$

Set $\mathbf{1}^T\mathbf{q} = 1$ and notice the fact that $A^TA = B^T \operatorname{diag}(\mathbf{q})B$, then we can rewrite the previous problem as follows

$$
\begin{aligned}
(\hat{\mathbf{q}}, \hat{\sigma}_n) = & \operatorname*{arg\,max}_{\mathbf{q}\in\mathbb{R}^m, \sigma_n(A)\in\mathbf{R}} \; \sigma_n^2(A) \qquad (18)\\
s.t. \quad & \mathbf{1}^T\mathbf{q} = 1;\\
& q_i \geq 0,\; i = 1,\ldots,m;\\
& B^T \operatorname{diag}(\mathbf{q})B - \sigma_n^2(A)I_n \succeq 0.
\end{aligned}
$$

It can be observed that this optimization is equivalent to the problem given by (16).

We conclude this observation in the following theorem.

Theorem 3: The ORKA can do at least as good as the RKA, in the sense that if we optimize $\kappa(A)$ over the norms of rows of A, we obtain the same probability distribution vector as the one obtained by the ORKA.

III. Further Discussions

Note that although the formulation in (16) is convex, it is still time consuming to solve this SDP optimization problem. In this section, we will discuss two possibilities to solve it approximately , which can alleviate some of the computational cost. One approximation of (16) is obtained by relaxing the constraint $B^T \operatorname{diag}(\mathbf{p})B - tI_n \succeq 0$ by the following linear constraints:

$$\mathbf{b}_i^T \operatorname{diag}(\mathbf{p})\mathbf{b}_i \geq t; \forall i = 1,\ldots,m. \qquad (19)$$

It is due to the fact that, for two positive semidefinite matrices $P_1, P_2 \in \mathbb{R}^{n\times n}$, if $P_1 \succeq P_2$, then $P_1(i,i) \geq P_2(i,i)$ holds for $i = 1,\cdots,n$. Such relaxation reduces the SDP problem into a Linear Programming (LP) problem, which is computationally easier to solve.

In order to get a better relaxation, we introduce another approximation method which relates to the research of *Optimal Input Design* [10]. Notice that $tr(B^T \operatorname{diag}(\mathbf{p})B) = 1$, i.e. the summation of all the singular values of $B^T \operatorname{diag}(\mathbf{p})B$ is fixed, then maximizing $\sigma_n(B^T \operatorname{diag}(\mathbf{p})B)$ means that we want all the singular values of $B^T \operatorname{diag}(\mathbf{p})B$ to be close. This leads us to consider maximizing the product of the singular values of $B^T \operatorname{diag}(\mathbf{p})B$, or maximizing the determinant of $B^T \operatorname{diag}(\mathbf{p})B$. As the log function is monotonically increasing, we can optimize the following

$$\max_{\mathbf{p}\in\mathbb{R}^m} \log|B^T \operatorname{diag}(\mathbf{p})B|, \qquad (20)$$

in which $|\cdot|$ denotes the matrix determinant. Optimizing this quantity subject to the same constraints of (15) boils down to solve the so-called *D-Optimal Design* problem. One simple iterative algorithm to solve such problem has been suggested in [12], which is given as

$$
\begin{aligned}
p_i^0 &= \frac{\|\mathbf{a}_i\|^2}{\|A\|_F^2};\; i = 1,\ldots,m;\\
p_i^{t+1} &= p_i^t \frac{\mathbf{b}_i^T(B^T \operatorname{diag}(\mathbf{p}^t)B)^{-1}\mathbf{b}_i}{n};\; i = 1,\ldots,m. \qquad (21)
\end{aligned}
$$

Here, $\mathbf{p}^t$ denotes the estimation at time t, and p_i^t denotes its i-th element. It has been proven in [13] that for this algorithm, $\log|B^T \operatorname{diag}(\mathbf{p}^t)B|$ decreases monotonically w.r.t. t. We will make use of such property to approximately solve (15) when the objective function is replace by (20). More discussions will be given in next section.

IV. Experiments

In this section, we will conduct experiments to illustrate the efficacy of the presented methods. The setup of our experiment is given as follows. The matrix A is first generated by *randn(m,n)* in Matlab with $m = 200$ and $n = 20$, after that, each row is normalized, and then scaled with a random number which is uniformly distributed in $[0, 1]$. The reason for generating A as such is that in the first stage, the generated rows of A will have different directions which are uniformly distributed on the sphere S^{n-1}[14]; and in the second stage, different rows of A with be assigned with different norms, which is directly related to the probability distribution vector chosen in [1]. $\mathbf{x}$ is generated by *randn(n,1)*, and $\mathbf{b}$ is generated as $\mathbf{b} = A\mathbf{x}$. We will compare the Mean Square Error (MSE) along the projection path obtained by all these methods, the first is the one suggested in [1] (abbreviated as *RKA*), the second is the one obtained by the SDP optimization given by (16) (abbreviated as *ORKA*) and the third is the one obtained by the LP approximations given by (19) (abbreviated as *LPORKA*), the last is the one obtained by the iterative method to solve the D-Optimal Design criteria (abbreviated as *ITEORKA*). We iterate (21) for 10 times in this experiment. For each method, we run the experiment 2000 times to get the averaged performance. The CVX toolbox[1] is used to solve the SDP and LP optimization problems. From the experiment, we can observe that the time for solving the LP problem in LPORKA is close to the time needed for the 10 iterations of (21), and the time needed for solving (16) in ORKA is approximately 7 times as them.

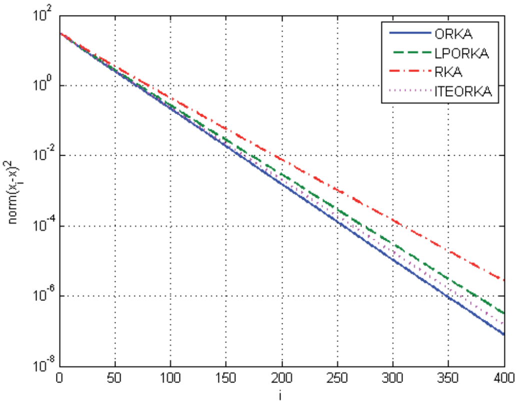


Fig. 2. The curves demonstrate the MSE for different methods. We can see that the ORKA improves the convergence speed the most; the LPORKA method and the ITERKA method also improve the convergence speed, and the ITEORKA method improves more than the LPORKA method.

V. Conclusion

This note discusses the possibility and methodology to find a probability distribution vector for selecting the rows of A to result in a better convergence speed of the Randomize Kaczmarz Algorithm. The lower bound and upper bound for the convergence speed is derived first. Then an optimized probability distribution vector is obtained by minimizing the upper bound, which turns to be given by solving a convex optimization problem. Properties of the approach are also discussed along the note.

[1] http://cvxr.com/

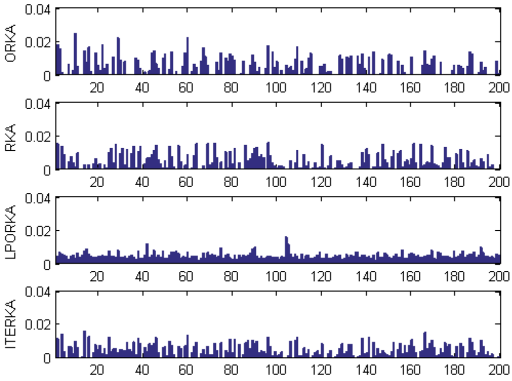


Fig. 3. An illustration of the probability distribution vectors obtained by different methods. Note that there are 68 zero elements of the probability distribution vector obtained by the ORKA method, which is 34% sparsity of the total length.

References

[1] T. Strohmer, R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, Journal of Fourier Analysis and Applications, 15(2), 262-278, 2009.

[2] Y. Censor, G.T. Herman, and M. Jiang, A note on the behavior of the randomized Kaczmarz algorithm of Strohmer and Vershynin, Journal of Fourier Analysis and Applications, 15(4), 431-436, 2009.

[3] T. Strohmer, R. Vershynin, Comments on the randomized Kaczmarz method, Journal of Fourier Analysis and Applications, 15(4), 437-440, 2009.

[4] S. Kaczmarz, Angenaherte Auflosung von Systemen linearer Gleichungen, Bulletin International de l'Acad*é*mie Polonaise des Sciences et des Lettres, 35, 355-357, 1937.

[5] F. Natterer, The Mathematics of Computerized Tomography, Wiley, New York, 1986.

[6] D. Needell, Randomized Kaczmarz solver for noisy linear systems, BIT Numerical Mathematics, 50(2), 395-403, 2010.

[7] Y. Eldar, D. Needell, Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss lemma, Numerical Algorithms, 58(2), 163-177, 2011.

[8] X. Chen, A. Powell, Almost sure convergence for the Kaczmarz algorithm with random measurements, Journal of Fourier Analysis and Applications, 18(6), 1195-1214, 2012.

[9] D. Needell and J. A. Tropp, Paved with Good Intentions: Analysis of a Randomized Block Kaczmarz Method, Linear Algebra and its Applications, 441, 199-221, 2014.

[10] V. V. Fedorov, Theory of Optimal Experiments, Academic Press, 1971.

[11] D. L. Donoho, Compressed Sensing, IEEE Transactions on Information Theory, 52(4), 1289-1306, 2006

[12] S. D. Silvey, D. M. Titterington and B. Torsney, An algorithm for optimal designs on a finite design space, Commun. Stat. Theory Methods, 14, 1379-1389, 1978.

[13] Y. Yu, Monotonic convergence of a general algorithm for computing optimal designs, The Annals of Statistics, 38(3), 1593-1606, 2010.

[14] G. Marsaglia, Choosing a Point from the Surface of a Sphere, The Annals of Mathematical Statistics, 43(2), 645-646, 1972.

Recent licentiate theses from the Department of Information Technology

2014-002 Johannes Nygren: *Output Feedback Control - Some Methods and Applications*

2014-001 Daniel Jansson: *Mathematical Modeling of the Human Smooth Pursuit System*

2013-007 Hjalmar Wennerström: *Meteorological Impact and Transmission Errors in Outdoor Wireless Sensor Networks*

2013-006 Kristoffer Virta: *Difference Methods with Boundary and Interface Treatment for Wave Equations*

2013-005 Emil Kieri: *Numerical Quantum Dynamics*

2013-004 Johannes Åman Pohjola: *Bells and Whistles: Advanced Language Features in Psi-Calculi*

2013-003 Daniel Elfverson: *On Discontinuous Galerkin Multiscale Methods*

2013-002 Marcus Holm: *Scientific Computing on Hybrid Architectures*

2013-001 Olov Rosén: *Parallelization of Stochastic Estimation Algorithms on Multicore Computational Platforms*

2012-009 Andreas Sembrant: *Efficient Techniques for Detecting and Exploiting Runtime Phases*

2012-008 Palle Raabjerg: *Extending Psi-calculi and their Formal Proofs*

2012-007 Margarida Martins da Silva: *System Identification and Control for General Anesthesia based on Parsimonious Wiener Models*

UPPSALA
UNIVERSITET

Department of Information Technology, Uppsala University, Sweden