On a Class of Optimization-Based Robust Estimators

Laurent Bako

On a class of optimization-based robust estimators

Laurent Bako

***Abstract*—We consider in this paper the problem of estimating a parameter matrix from observations which are affected by two types of noise components: (i) a sparse noise sequence which, whenever nonzero can have arbitrarily large amplitude (ii) and a dense and bounded noise sequence of "moderate" amount. This is termed a robust regression problem. To tackle it, a quite general optimization-based framework is proposed and analyzed. When only the sparse noise is present, a sufficient bound is derived on the number of nonzero elements in the sparse noise sequence that can be accommodated by the estimator while still returning the true parameter matrix. While almost all the restricted isometry-based bounds from the literature are not verifiable, our bound can be easily computed through solving a convex optimization problem. Moreover, empirical evidence tends to suggest that it is generally tight. If in addition to the sparse noise sequence, the training data are affected by a bounded dense noise, we derive an upper bound on the estimation error.**

I. Introduction

In many engineering fields such as control system design, signal processing, machine learning or statistics, one is frequently confronted with the problem of empirically uncovering a mathematical relationship between a number of signals of interest. The usual method to achieve this goal is to run an experiment during which one measures (a finite number of) samples of the relevant signals and proceed with fitting a certain model structure to the experimental data samples. This process is known as system identification [11], [19]. A issue of critical importance during this process is that the experimental data samples might be contaminated by a measurement noise of relatively high level due for example to intermittent sensor failures or various communication disruptions. To cope with the troublesome effects of the noise, the model estimation must be designed with care.

In this paper we consider the situation where the data are corrupted by two types of noise: a sparse noise sequence which shows up only intermittently in time but can take on arbitrarily large values whenever it is nonzero; and a more standard dense noise component of moderate amount.

II. The robust regression problem

Consider a system described by an equation of the form

$$y_t = A^o x_t + f_t + e_t \tag{1}$$

where $y_t \in \mathbb{R}^m$ and $x_t \in \mathbb{R}^n$ are respectively the output and the regressor vector at time t; $A^o \in \mathbb{R}^{m\times n}$ is an unknown parameter matrix; f_t and e_t are some noise terms which are unobserved.

Problem. Given a finite collection $\{x_t, y_t\}_{t=1}^N$ of measurements obeying the relation (1), the robust regression problem of interest here is the one of finding an estimate of the parameter matrix A^o under the assumptions that $\{e_t\}$ and $\{f_t\}$ are unknown but enjoy the following (informal) properties:

- $\{e_t\}$ is a dense noise sequence with bounded elements accounting for moderate model mismatches or measurement noise.
- $\{f_t\}$ is such that the majority of its elements are equal to zero while the remaining nonzero elements can be of arbitrarily large magnitude. The nonzero elements of that sequence are usually termed gross errors or outliers. They can account for possible intermittent sensor faults. We will refer to $\{f_t\}$ as the sequence of sparse noise.

For the time being, these are just informal descriptions of the characteristics of the sequences $\{f_t\}$ and $\{e_t\}$. They will be made more precise whenever necessary in the sequel for the need of stating more formal results.

Let $Y \in \mathbb{R}^{m\times N}$ and $X \in \mathbb{R}^{n\times N}$ be data matrices formed respectively with N output measurements and regressor vectors. Then it follows from (1) that

$$Y = A^o X + E + F, \tag{2}$$

where $E \in \mathbb{R}^{m\times N}$ and $F \in \mathbb{R}^{m\times N}$ are unknown noise components. The matrices Y and X can be structured or not, depending on whether the system (1) is dynamic or not. For example, when the model (1) is of MIMO FIR type, Y contains a finite collection of output measurements while X is a Hankel matrix containing lagged inputs of the system. In this case Y and X take the form

$$Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix},$$
$$X = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \\ u_0 & u_1 & \cdots & u_{N-1} \\ \vdots & \vdots & \cdots & \vdots \\ u_{1-n_f} & u_{2-n_f} & \cdots & u_{N-n_f} \end{bmatrix}.$$

where $\{u_t\}$ and $\{y_t\}$ stand respectively for the input and output of the system and the maximum lag n_f is called the order of the model. In the sequel, the notations of the type y_t and x_t with subindex $t \in \mathbb{I} \triangleq \{1, \ldots, N\}$ refer to the columns of Y and X respectively.

Relevant prior works. The so formulated regression problem is called a robust regression problem in connection with the fact that the error matrix F assume columns of (possibly) arbitrarily large amplitude. It has applications in e.g., the identification of switched linear systems [1], [15], [14], subspace clustering [2], etc. Existing approaches for solving the robust regression problem can be roughly divided into two groups: methods from the field of robust statistics [17], [12], [9] which have been developed since the early 60s and a class of more recent methods inspired by the compressed sensing paradigm [3], [4], [18], [21], [13]. The first group comprises methods such as the least absolute deviation (LAD) estimator [8], the least median of squares [16], the least trimmed squares [17], the family of M-estimators [9]. The latter group can be viewed essentially as a refreshed look at the so-called least absolute deviation method. There has been however a fundamental shift of philosophy in the analysis. While in the framework of robust statistics, robustness of an estimator is measured in terms of the breakdown point (the asymptotic minimum proportion of points which cause the estimation error induced by an estimator to be unbounded if they were to be arbitrarily corrupted by gross errors), in the compressed-sensing-inspired category of robust methods, the analysis aims generally at characterizing properties of the data that favor exact recovery of the true parameter matrix A^o. In this latter group, the LAD estimator is sometimes regarded as a convex relaxation of a combinatorial sparse optimization problem.

L. Bako is with Laboratoire Ampère – Ecole Centrale de Lyon – Université de Lyon – 36 avenue Guy de Collongue, 69134 Ecully, France – E-mail: laurent.bako@ec-lyon.fr

To the best of our knowledge, only the papers [18] provides an explicit bound on the estimation error induced by the LAD estimator. However that bound does not fully apply to the current setting since the estimators although similar are of different natures. Indeed, the LAD estimator stands only as a special case of the current framework. Moreover the bound in [18] is not easily computable while ours is. The references [4] and [13] provide some bounds for a noise-aware version of the LAD estimator which are based respectively on the Restricted Isometry Property (RIP) and a measure of subspace angles. Unfortunately numerical evaluation of those bounds is a process of exponential complexity, a price that is unaffordable in practice.

A related but different problem from the regression problem considered here is that of sparse signal recovery studied in the field of compressed sensing [5], [7]. This is about finding the sparsest solution to an underdetermined set of linear equations. Various analysis approaches have been devised which rely on the RIP constant, the mutual coherence, the nullspace property, to name but a few. Again, these analysis results either cannot be extended efficiently to the robust regression problem or lead to bounds that are NP-hard to compute [20], [10], [6].

Contributions. In this paper we propose and analyze a class of optimization-based robust estimators. It is shown that the robust properties of the estimators are essentially inherited from a key property of the to-be-optimized performance function (or loss function) called column-wise summability. The proposed framework admits the LAD estimator and its usual variants as special cases. Moreover it applies to both SISO and MIMO systems. When the dense noise component E in (2) is identically equal to zero, we derive bounds on the number of gross errors (nonzero columns of F) that the estimator is able to accommodate while still returning the true parameter matrix A^o. In comparison with the existing literature, the proposed bounds have the important advantage that they are numerically computable through convex optimization. When both E and F are active, exact recovery of the true parameter matrix is no longer possible. In this scenario, we derive upper bounds on the parametric estimation error in function of the amplitude of E and the number of nonzero columns of F. Again, computable but (possibly) looser versions of those bounds are obtainable.

The current paper can be viewed as a generalization of our previous work reported in [3]. While [3] provides an analysis of mostly a single estimator (namely the LAD estimator) relying on nonsmooth optimization theory, we focus here on a much larger class of optimization-based robust estimators by highlighting some key robustness-inducing properties. Moreover, we provide, for the considered class of estimators, stability results which permit the estimation of parametric error bounds.

Outline. The rest of the paper is organized as follows. Section III defines the optimization-based approach to the robust regression problem. Section IV discusses the properties of the proposed estimation framework. Section V provides further comments. Section VI reports some numerical experiments. Lastly, Section VII contains some concluding remarks.

Notations. $\mathbb{R}$ is the set of real numbers; $\mathbb{R}_{\geq 0}$ (respectively $\mathbb{R}_{>0}$) is the set of nonnegative (respectively positive) real numbers ; $\mathbb{R}^N$ is the space of N-tuples (vectors) of real numbers. For any vector $x = [x_1 \quad \cdots \quad x_N]^\top \in \mathbb{R}^N$, the p-norm of x with $p \in \{1, \ldots, \infty\}$ is defined by $\|x\|_p = \left(\sum_{i=1}^N |x_i|^p\right)^{1/p}$. A special case is the limit case $p = \infty$ in which $\|x\|_\infty = \max_{i=1,\ldots,N} |x_i|$. For any matrix $A = [a_1 \quad \cdots \quad a_N]$ with $a_i \in \mathbb{R}^m$, the induced p-norm of A is defined by $\|A\|_p = \sup_{x \in \mathbb{R}^N, \|x\|_p = 1} \|Ax\|_p$.

Cardinality of a finite set. Throughout the paper, whenever $\mathcal{S}$ is a finite set, the notation $|\mathcal{S}|$ will refer to the cardinality of $\mathcal{S}$. However, for a real number x, $|x|$ will denote the absolute value of x.

Submatrices and subvectors. Let $X \in \mathbb{R}^{n \times N}$ and $\mathbb{I} = \{1, \ldots, N\}$ be the index set for the columns of X. If $I \subset \mathbb{I}$, the notation X_I denotes a matrix in $\mathbb{R}^{n \times |I|}$ formed with the columns of X indexed by I. We will use the convention that $X_I = 0 \in \mathbb{R}^n$ when the index set I is empty.

III. A CLASS OF ROBUST ESTIMATORS

Let $\mathcal{D}_N$ be the set of N data points generated by system (1) for any possible values of the noise sequences, i.e.,

$$\mathcal{D}_N = \Big\{(Y, X) \in \mathbb{R}^{m \times N} \times \mathbb{R}^{n \times N} : \exists (E, F) \in \mathcal{G}_N^e \times \mathcal{G}_N^f, \text{ (2) holds}\Big\},$$

with $\mathcal{G}_N^e \subset \mathbb{R}^{m \times N}$ and $\mathcal{G}_N^f \subset \mathbb{R}^{m \times N}$ denoting the set of dense and sparse noise matrices respectively. The estimation problem aims at determining the unknown parameter matrix A^o given a point (Y, X) in $\mathcal{D}_N$. Of course, this quest would not make much sense if the noises E and F were completely arbitrary since in this case, we would have $\mathcal{D}_N = \mathbb{R}^{m \times N} \times \mathbb{R}^{n \times N}$ hence losing any informativity concerning the data-generating system. Therefore some minimum constraints need to be put on E and F as informally described above.

With respect to the estimation problem just stated, an estimator is a set-valued map $\Psi : \mathcal{D}_N \to \mathscr{P}(\mathbb{R}^{m \times n})$, $(Y, X) \mapsto \Psi(Y, X)$ which is defined from the data space $\mathcal{D}_N$ to the power set $\mathscr{P}(\mathbb{R}^{m \times n})$ of the parameter space. For (Y, X) generated by a system of the form (1), one would like to design an estimator achieving, whenever possible, the ideal property that $\Psi(Y, X) = \{A^o\}$. In default of that ideal situation, a more pragmatic goal is to search for a Ψ so that $A^o \in \Psi(Y, X)$ and $\Psi(Y, X)$ is of small size in some sense despite the troublesome effects of the unknown noise components E and F. The design of an optimal estimator requires specifying a performance index (usually called a loss function) which is to be minimized.

In this paper, we study the properties of the estimator of the parameter matrix A^o in (2) defined by

$$\Psi(Y, X) = \underset{A \in \mathbb{R}^{m \times n}}{\arg\min}\, \varphi(Y - AX) \tag{3}$$

where $\varphi : \mathcal{M}(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ is a *convex function* defined on the set $\mathcal{M}(\mathbb{R})$ of (all) real matrices. It is assumed that φ has the following properties:

P1. For all $B, C \in \mathcal{M}(\mathbb{R})$ of compatible dimensions,

$$\varphi([B \quad C]) = \varphi(B) + \varphi(C) \tag{4}$$

with $[B \quad C]$ denoting the matrix formed by concatenating column-wise B and C.

P2. There exists a matrix norm $\ell : \mathcal{M}(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ such that for all $B, C \in \mathcal{M}(\mathbb{R})$, conformable for addition,

$$\varphi(B) \leq \varphi(B - C) + \ell(C) \tag{5}$$

P3. There exists a constant real number $\varepsilon \geq 0$ such that for all $B \in \mathcal{M}(\mathbb{R})$ with n rows and N columns,

$$\ell(B) - |I_\varepsilon^c(B)|\, \varepsilon \leq \varphi(B) \leq \ell(B) \tag{6}$$

where

$$I_\varepsilon^c(B) = \big\{i \in \{1, \ldots, N\} : \ell(b_i) > \varepsilon\big\}$$

and $|I_\varepsilon^c(B)|$ is the cardinality of $I_\varepsilon^c(B)$ and $b_i \in \mathbb{R}^n$ is the ith column of the (n, N)-matrix B.

The property (4) will be called column-wise summability. Since φ is a function defined over the space of real matrices of any dimensions, it is also defined for n-dimensional vectors of real numbers. Hence

according to property (4), if $B = [b_1 \quad \cdots \quad b_N]$ with column vectors $b_i \in \mathbb{R}^n$, then

$$\varphi(B) = \sum_{i=1}^{N} \varphi(b_i).$$

The so-defined function φ is not necessarily a norm. For any $\varepsilon^o \geq 0$ and any vector norm ℓ^o, it can be verified that the function φ defined by

$$\varphi(B) = \sum_{i=1}^{N} \max(0, \ell^o(b_i) - \varepsilon^o) \tag{7}$$

is positive and convex and satisfies properties (4)-(6) but it is not a norm for $\varepsilon^o > 0$ since in this case, $\varphi(B) = 0$ does not imply that $B = 0$. But if $\varepsilon^o = 0$ in (7), then $\varphi = \ell$ by (6) so that φ corresponds to the matrix norm defined by $\varphi(B) = \sum_{i=1}^{N} \ell^o(b_i)$. We note in this latter case that (6) is trivial while (5) reduces to the triangle inequality.

We will show in the sequel that the estimator Ψ in (3) enjoys some impressive robustness properties with respect to the sparse noise matrix F. The term sparse is used here to mean that a relatively large proportion of the column vectors of F are equal to zero. And saying that Ψ is robust with respect to F means that $\Psi(Y, X)$ does not depend on (or is insensitive to) the magnitudes of the nonzero columns of F under the sparsity condition. Therefore those few columns which are nonzero can have arbitrarily large magnitude. As will be shown in the sequel, the robustness properties of Ψ are inherited from the properties P1-P3 of the objective function φ. In the special case where φ is a norm, the properties P2-P3 are automatically satisfied so that P1 becomes the only key property required. As to the convexity of φ, it is intended just for computational reasons as it eases the solving of the optimization problem in (3).

IV. Properties of the robust estimators

A. Exact recoverability

We first study the conditions under which the true parameter matrix A^o in (1) can be exactly recovered. Theorem 1 and Theorem 2 stated next show that if the number of nonzero columns in the matrix $V \triangleq E + F$ is less than a certain threshold, then $\Psi(Y, X) = \{A^o\}$.

Theorem 1 (A necessary and sufficient condition)**.** *Let φ be a function satisfying* (4)-(6) *with $\varepsilon = 0$ and Ψ be defined as in* (3)*. Let d be an integer and assume that* $\mathrm{rank}(X) = n$*. For any $A \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times N}$, let $\mathbb{I}^c(Y - AX) = \{t \in \mathbb{I} : y_t - Ax_t \neq 0\}$. Then the following statements are equivalent.*

(i)

$$\begin{aligned} \forall A \in \mathbb{R}^{m\times n}, \forall Y \in \mathbb{R}^{m\times N},\ |\mathbb{I}^c(Y - AX)| \leq d \\ \Rightarrow \qquad \Psi(Y, X) = \{A\} \end{aligned} \tag{8}$$

(ii)

$$\max_{\substack{I^c\subset\mathbb{I}:\\ |I^c|=d}} \max_{\substack{\Lambda\in\mathbb{R}^{m\times n}\\ \Lambda\neq 0}} \left[\frac{\varphi(\Lambda X_{I^c})}{\varphi(\Lambda X)}\right] < \frac{1}{2} \tag{9}$$

Here and in the following, the notation $\mathbb{I} \triangleq \{1, \ldots, N\}$ is used to denote the index set for the columns of the data matrices.

Proof: We first note that the rank assumption on X is intended to insure that (9) is well-defined since then, with φ being a norm, $\varphi(\Lambda X) \neq 0$ whenever $\Lambda \neq 0$.

(i) $\Rightarrow$ (ii): Assume that (i) holds.

Consider an arbitrary subset I^c of $\mathbb{I}$ such that $|I^c| = d$. Let Λ be any matrix in $\mathbb{R}^{m\times n}$ satisfying $\Lambda \neq 0$. Finally, consider a matrix $Y \in \mathbb{R}^{m\times N}$ defined by $Y_{I^c} = 0$ and $Y_{I^0} = \Lambda X_{I^0}$ where $I^0 = \mathbb{I}\setminus I^c$. Then $\mathbb{I}^c(Y - \Lambda X) \subset I^c$ and so $|\mathbb{I}^c(Y - \Lambda X)| \leq d$. Hence by (i) $\{\Lambda\} = \arg\min_H \varphi(Y - HX)$ which means that $\varphi(Y - \Lambda X) < \varphi(Y - HX)$ for any $H \in \mathbb{R}^{m\times n}$, $H \neq \Lambda$. In particular, by taking $H = 0$ we get $\varphi(Y - \Lambda X) < \varphi(Y)$. It follows from the property (4) that

$$\varphi(Y_{I^c} - \Lambda X_{I^c}) + \varphi(Y_{I^0} - \Lambda X_{I^0}) < \varphi(Y_{I^c}) + \varphi(Y_{I^0}).$$

Using now the relations $Y_{I^c} = 0$ and $Y_{I^0} = \Lambda X_{I^0}$ yields $\varphi(\Lambda X_{I^c}) < \varphi(\Lambda X_{I^0})$ or, equivalently, $\varphi(\Lambda X_{I^c}) < 1/2\varphi(\Lambda X)$. Eq. (9) then follows from the fact that I^c and Λ are arbitrary.

(ii) $\Rightarrow$ (i): To begin with, note that if Eq. (9) holds for some d, then it holds also for any $d_0 \leq d$. As a result, the equality $|I^c| = d$ in (9) can be changed to $|I^c| \leq d$. Assuming (ii), let $A \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times N}$ be matrices satisfying $|\mathbb{I}^c(Y - AX)| \leq d$. Set $I^c = \mathbb{I}^c(Y - AX)$ and $I^0 = \mathbb{I} \setminus I^c$. Then for all $\Lambda \in \mathbb{R}^{m\times n}$ such that $\Lambda \neq 0$,

$$2\varphi(\Lambda X_{I^c}) < \varphi(\Lambda X) = \varphi(\Lambda X_{I^c}) + \varphi(\Lambda X_{I^0}),$$

where the equality is obtained by the property (4) of φ. It follows that

$$\varphi(\Lambda X_{I^c}) < \varphi(Y_{I^0} - (A + \Lambda)X_{I^0}). \tag{10}$$

On the other hand, we know by (5) that

$$\varphi(Y_{I^c} - AX_{I^c}) - \varphi(Y_{I^c} - (A + \Lambda)X_{I^c}) \leq \varphi(\Lambda X_{I^c}).$$

Combining with the inequality (10) yields

$$\varphi(Y - AX) < \varphi(Y - (A + \Lambda)X).$$

Since Λ is an arbitrary nonzero matrix, this inequality says that A is the unique minimizer of $V(H) = \varphi(Y - HX)$. ∎

Consider a data pair (Y, X) generated by (1). By letting

$$\pi^c_\varphi(X) = \max\big\{d : \text{Eq. (9) holds}\big\}, \tag{11}$$

and assuming that $\pi^c_\varphi(X) > 0$ we can see that whenever $|\mathbb{I}^c(Y - A^oX)| \leq \pi^c_\varphi(X)$, A^o can be exactly recovered by computing $\Psi(Y, X)$. Of course this is likely to hold only if the dense noise component E does not exist. So in the situation where $E = 0$, the theorem says that A^o can be uniquely obtained by convex optimization provided that the number of outliers (nonzero columns of F) is less than or equal to $\pi^c_\varphi(X)$. For the condition of exact recoverability to be checkable we must be able to compute $\pi^c_\varphi(X)$. The bad news are that evaluating numerically such a number is likely to be NP-hard in most cases.

In the sequel, we investigate sufficient conditions of exact recovery which are more tractable from a numerical standpoint. For this purpose let us introduce some definitions.

Definition 1. *A matrix $X = [x_1 \quad \cdots \quad x_N] \in \mathbb{R}^{n\times N}$ is said to be self-decomposable if* $\mathrm{rank}(X) = n$ *and for all $k \in \mathbb{I}$, $x_k \in \mathrm{im}(X_{\neq k})$ where $X_{\neq k} \triangleq X_{\mathbb{I}\setminus\{k\}}$ is the matrix obtained from X by removing its k-th column and* $\mathrm{im}(\cdot)$ *refers to range space.*

For a matrix to be self-decomposable it is enough that $X_{\neq k}$ be full row rank for any $k \in \mathbb{I}$. Achieving this condition in practice seems easy provided that the number N of measurements is large enough compared to the dimension n of X.

Definition 2 (self-decomposability amplitude)**.** *Let $X \in \mathbb{R}^{n\times N}$ be a self-decomposable matrix. We call* self-decomposability amplitude *of X, the number $\xi(X)$ defined by*

$$\xi(X) = \max_{k\in\mathbb{I}} \min_{\gamma_k\in\mathbb{R}^{N-1}} \Big\{ \|\gamma_k\|_\infty : x_k = X_{\neq k}\gamma_k \Big\}. \tag{12}$$

The so-defined $\xi(X)$ constitutes a quantitative measure of richness (or genericity) of the regressor matrix X. By richness it is meant here how much, in a global sense, the columns of X are linearly

independent. $\xi(X)$ is expected to be small if the columns of X are somehow strongly linearly independent.

Remark 1. *If for some k the norm of x_k was to be considerably large in comparison to the norm of the other columns of X, then $\xi(X)$ would get large hence reducing recoverability capacity of the considered class of estimators (see also Eq.* (9)*). Such situations can be alleviated by normalizing each column of X, i.e., for example by replacing (y_k, x_k) by $(\tilde{y}_k, \tilde{x}_k) \triangleq (y_k/\|x_k\|, x_k/\|x_k\|)$ under the assumption that $x_k \neq 0$ for all $k \in \mathbb{I}$.*

With the help of the device of self-decomposability amplitude (12), we can state a condition for exact recovery of the parameter matrix A^o by solving the optimization problem in (3). A similar result was proven in [3] for the Least Absolute Deviation (LAD) estimator.

Theorem 2 (A sufficient condition for exact recovery)**.** *Let φ be a function satisfying* (4)-(6) *with $\varepsilon = 0$ and Ψ be defined as in* (3)*. Assume that X is self-decomposable. Then the following statement is true:*

$$\begin{array}{l} \forall A \in \mathbb{R}^{m\times n},\ \forall Y \in \mathbb{R}^{m\times N}, \\ |\mathbb{I}^c(Y - AX)| < T\big(\xi(X)\big) \quad \Rightarrow \quad \Psi(Y,X) = \big\{A\big\}. \end{array} \tag{13}$$

where $T : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is the function defined by $T(\alpha) = \frac{1}{2}\big(1 + \frac{1}{\alpha}\big)$.

Proof: The proof is completely parallel to that of Theorem 11 in [3]. From the assumptions, each x_k, $k \in \mathbb{I}$, can be written as a linear combination of the columns of $X_{\neq k}$. Let $\gamma_k \in \mathbb{R}^{N-1}$ be any vector satisfying $x_k = X_{\neq k}\gamma_k$. It follows that for any $\Lambda \in \mathbb{R}^{m\times n}$,

$$\varphi(\Lambda x_k) = \varphi\big(\sum_{t\in\mathbb{I}\setminus\{k\}} \gamma_{k,t}\Lambda x_t\big)$$

with $\gamma_{k,t}$ denoting the entry of $\gamma_k \in \mathbb{R}^{N-1}$ indexed by t. Under the assumptions of the theorem, φ is a norm. So, it is positive and satisfies the triangle inequality property. As a result we can write

$$\varphi(\Lambda x_k) \leq \sum_{t\neq k} |\gamma_{k,t}|\, \varphi(\Lambda x_t) \leq \|\gamma_k\|_\infty \left(\varphi(\Lambda X) - \varphi(\Lambda x_k)\right)$$

where the rightmost term follows from the property (4) of φ. Since this holds for any γ_k such that $x_k = X_{\neq k}\gamma_k$, it holds also for

$$\gamma_k^\star = \underset{\gamma\in\mathbb{R}^{N-1}}{\arg\min} \Big\{ \|\gamma\|_\infty : x_k = X_{\neq k}\gamma \Big\}.$$

Hence,

$$\varphi(\Lambda x_k) \leq \xi(X)\left(\varphi(\Lambda X) - \varphi(\Lambda x_k)\right) \ \ \forall k \in \mathbb{I}, \forall \Lambda \in \mathbb{R}^{m\times n}. \tag{14}$$

or equivalently,

$$\varphi(\Lambda x_k) \leq \frac{\xi(X)}{1+\xi(X)}\varphi(\Lambda X) \quad \forall k \in \mathbb{I},\ \forall \Lambda \in \mathbb{R}^{m\times n}.$$

Let I^c be any subset of $\mathbb{I}$ and pose $|I^c| = d$. Summing the previous inequality over the set I^c yields

$$\max_{\Lambda\neq 0} \frac{\varphi(\Lambda X_{I^c})}{\varphi(\Lambda X)} \leq \frac{1}{2T\big(\xi(X)\big)}\,|I^c| \tag{15}$$

Note that the term on the right hand side is well-defined since by the self-decomposability assumption, $\operatorname{rank}(X) = n$ which implies that $\varphi(\Lambda X) \neq 0$ whenever $\Lambda \neq 0$. Therefore (9) holds if $|I^c| < T\big(\xi(X)\big)$ and the conclusion follows from Theorem 1. ∎

It is worth noting that the threshold $T(\xi(X))$ on the number of correctable outliers does not depend on φ. Hence this threshold is valid when the estimator is defined from any matrix norm obeying (4).

Remark 2. *The statement of Theorem 2 still holds true if we replace $\xi(X)$ with the φ-dependent number $\delta_\varphi(X)$ defined by*

$$\delta_\varphi(X) = \max_{k\in\mathbb{I}} \sup_{\Lambda\neq 0} \frac{\varphi(\Lambda x_k)}{\varphi(\Lambda X_{\neq k})} \tag{16}$$

when it is assumed that φ is a norm and $\operatorname{rank}(X_{\neq k}) = n$ for all k. Doing so will give a less conservative condition for exact recovery. However $\delta_\varphi(X)$ seems much harder to evaluate numerically than $\xi(X)$.

Remark 3 (A few useful properties of $\xi(X)$)**.**

- *For any nonsingular matrix $R \in \mathbb{R}^{n\times n}$, $\xi(RX) = \xi(X)$. It follows that the number $\xi(X)$ depends only on the subspace spanned by the rows of the regressor matrix X.*
- *For any self-decomposable $X \in \mathbb{R}^{n\times N}$, $\xi(X)$ is lower-bounded in the following sense*

$$\xi(X) \geq \frac{1}{N-1},$$

This follows from the more general observation that

$$\xi(X) \geq \max_{k\in\mathbb{I}} \frac{\|x_k\|}{\sum_{t\neq k}\|x_t\|}$$

for any vector norm $\|\cdot\|$. As a result, $T(\xi(X)$ is upper-bounded as follows

$$T(\xi(X)) \leq \frac{N}{2}.$$

Theorem 2 provides a sufficient condition for exact recovery in the situation where the function φ is a norm. Next, another condition is stated which holds in the general case.

Proposition 1. *Consider a triplet $(\varphi, \ell, \varepsilon)$ satisfying* (4)-(6)*. For $A \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times N}$, pose $I^c = \mathbb{I}^c(Y - AX)$, $I^0 = \mathbb{I}\setminus I^c = \{t \in \mathbb{I} : y_t - Ax_t = 0\}$ and $I^c_\varepsilon(\Lambda X_{I^0}) = \big\{t \in I^0 : \ell(\Lambda x_t) > \varepsilon\big\}$. Then $\Psi(Y,X) = \{A\}$ if*

$$|I^c_\varepsilon(\Lambda X_{I^0})|\,\varepsilon < \ell(\Lambda X_{I^0}) - \ell(\Lambda X_{I^c}) \tag{17}$$

$\forall \Lambda \in \mathbb{R}^{m\times n}, \Lambda \neq 0$.

Proof: $\Psi(Y,X) = \{A\}$ is equivalent to

$$\varphi(Y - AX) < \varphi(Y - (A+\Lambda)X)$$

for any $\Lambda \in \mathbb{R}^{m\times n}$, $\Lambda \neq 0$. Using the definitions of the sets I^0 and I^c and applying property (4) of φ yields the equivalent relation

$$\varphi(Y_{I^c} - AX_{I^c}) - \varphi(Y_{I^c} - (A+\Lambda)X_{I^c}) < \varphi(\Lambda X_{I^0}).$$

By (5), we can note that $\varphi(Y_{I^c} - AX_{I^c}) - \varphi(Y_{I^c} - (A+\Lambda)X_{I^c}) \leq \ell(\Lambda X_{I^c})$. It then follows that

$$\ell(\Lambda X_{I^c}) < \varphi(\Lambda X_{I^0})$$

is a sufficient condition for $\Psi(Y,X) = \{A\}$. Finally, invoking (6) allows us to observe that $\ell(\Lambda X_{I^0}) - |I^c_\varepsilon(\Lambda X_{I^0})|\,\varepsilon \leq \varphi(\Lambda X_{I^0})$ which implies that $\ell(\Lambda X_{I^c}) < \ell(\Lambda X_{I^0}) - |I^c_\varepsilon(\Lambda X_{I^0})|\,\varepsilon$ is a sufficient condition for $\Psi(Y,X) = \{A\}$. We have hence proved the proposition. ∎

B. Uncertainty set induced by dense noise

When both E and F are nonzero in the data-generating system (1), $\Psi(Y,X)$ is likely to be a non-singleton subset of $\mathscr{P}(\mathbb{R}^{m\times n})$ especially if we consider all possible realizations of the unknown components E and F. In this case the desirable properties of the estimator are in default of better (i) that it contains A^o and (ii) that its size with respect to some metric is as small as possible. In this section we are interested in estimating the size of $\Psi(Y,X)$ when both

dense noise E and sparse noise F are active in the data-generating system (1).

A notion of estimator gain. Similarly to the concept of system gain in control [22], one could define the gain of an estimator, that is, a quantitative measure of the sensitivity of the estimator with respect to the perturbations affecting the measurements. Consider a data pair (Y, X) generated by a system of the form (1) with A^o being the parameter matrix sought for. Let us fix the sparse noise matrix F or view it somehow as part of the data-generating system. This consideration proceeds from the fact that Ψ can be insensitive to F (when acting alone) under, for example, the condition derived in Theorem 2. Let E be bounded in the sense that $\ell(E)$ is finite with ℓ being the norm appearing in (6). Then we can define a gain of the estimator with respect to the dense noise component E. More specifically, an (ℓ, q)-gain of the estimator Ψ with respect to the dense noise E may be defined by

$$g_{\ell,q}(Y,X) = \sup_{\substack{A^\star \in \Psi(Y,X) \\ 0<\ell(E)<\infty \\ F \text{ sparse}}} \frac{\|A^\star - A^o\|_q}{\ell(E)}. \tag{18}$$

Here $\|\cdot\|_q$ denotes matrix q-norm. The so-defined number $g_{\ell,q}(Y,X)$ provides an upper bound on the distance from the set $\Psi(Y,X)$ to A^o in function of the amount of dense noise. The following theorem and its corollaries show that if the number of nonzero columns in F is no larger than a certain threshold, then $g_{\ell,q}(Y,X)$ exists and is finite.

Theorem 3. *Let (Y,X) be the data generated by system* (1) *subject to the noise components E and F. Consider a triplet $(\varphi, \ell, \varepsilon)$ satisfying* (4)-(6)*. Let $S^0 \subset \mathbb{I}$ be a set such that $F_{S^0} = 0$ and let $S^c = \mathbb{I} \setminus S^0$. Assume that the matrix X and the partition (S^0, S^c) are such that there exists $\alpha > 0$ such that*

$$\ell(\Lambda X_{S^0}) - \ell(\Lambda X_{S^c}) \geq \alpha \|\Lambda\|_q \ \forall \Lambda \in \mathbb{R}^{m\times n}, \tag{19}$$

with $\|\cdot\|_q$ denoting some matrix q-norm.
Then for any $A^\star \in \Psi(Y,X)$, it holds that

$$\|A^\star - A^o\|_q \leq \frac{1}{\gamma_{\ell,q}(X,S^c)}\big[2\ell(E_{S^0}) + |I_\varepsilon^c|\,\varepsilon\big] \tag{20}$$

with[1] $I_\varepsilon^c = I_\varepsilon^c(Y_{S^0} - A^\star X_{S^0}) = \big\{t \in S^0 : \ell(y_t - A^\star x_t) > \varepsilon\big\}$ *and*

$$\gamma_{\ell,q}(X,S^c) = \inf_{\substack{\Lambda\in\mathbb{R}^{m\times n} \\ \Lambda\neq 0}} \frac{\ell(\Lambda X_{S^0}) - \ell(\Lambda X_{S^c})}{\|\Lambda\|_q} \tag{21}$$

where $\|\cdot\|_q$ refers to matrix q-norm.

Proof: By definition of $\Psi(Y,X)$ in (3),

$$\varphi(Y - A^\star X) \leq \varphi(Y - AX)\ \forall A \in \mathbb{R}^{m\times n}$$

By letting $\Lambda = A - A^o$, $\Lambda^\star = A^\star - A^o$ and applying (2), the last inequality takes the form

$$\varphi(F + E - \Lambda^\star X) \leq \varphi(F + E - \Lambda X)\ \forall \Lambda \in \mathbb{R}^{m\times n}.$$

In particular, for $\Lambda = 0$, we get $\varphi(F + E - \Lambda^\star X) \leq \varphi(F + E)$ which, thanks to property (4) of φ, takes the form

$$\varphi(F_{S^c} + E_{S^c} - \Lambda^\star X_{S^c}) + \varphi(E_{S^0} - \Lambda^\star X_{S^0}) \leq \varphi(F_{S^c} + E_{S^c}) + \varphi(E_{S^0}).$$

Now applying property (5) to the first member of the left hand side and rearranging yields

$$\varphi(E_{S^0} - \Lambda^\star X_{S^0}) - \ell(\Lambda^\star X_{S^c}) \leq \varphi(E_{S^0}).$$

[1]The notation I_ε^c is used for simplicity reasons.

Using (6) gives

$$\ell(E_{S^0} - \Lambda^\star X_{S^0}) - |I_\varepsilon^c|\,\varepsilon - \ell(\Lambda^\star X_{S^c}) \leq \varphi(E_{S^0}) \leq \ell(E_{S^0}).$$

Here we used the fact that $I_\varepsilon^c(E_{S^0} - \Lambda^\star X_{S^0})$ is equal to the set I_ε^c defined in the statement of the theorem.
Applying the triangle inequality property of ℓ, it can be seen that $\ell(\Lambda^\star X_{S^0}) - \ell(E_{S^0}) \leq \ell(E_{S^0} - \Lambda^\star X_{S^0})$. Combining with the previous inequality yields

$$\ell(\Lambda^\star X_{S^0}) - \ell(\Lambda^\star X_{S^c}) \leq 2\ell(E_{S^0}) + |I_\varepsilon^c|\,\varepsilon.$$

Finally, it follows from the definition of $\gamma_{\ell,q}(X,S^c)$ in (21) that

$$\gamma_{\ell,q}(X,S^c)\,\|\Lambda^\star\|_q \leq \big[2\ell(E_{S^0}) + |I_\varepsilon^c|\,\varepsilon\big].$$

The condition (19) guarantees that $\gamma_{\ell,q}(X,S^c)$ is well-defined and is positive. Hence the statement of the theorem is established. ■

Theorem 3 constitutes an interesting stability result in that it provides a finite upper bound on the distance from A^o to the set $\Psi(Y,X)$ as a function of the amplitude of the dense noise matrix E. It applies to any estimator Ψ defined as in (3) with φ a function obeying (4)-(6). In particular, in the situation where φ is a norm (in which case ε can be taken equal to zero in (6)), the inequality in (20) simplifies to

$$\|A^\star - A^o\|_q \leq \frac{2}{\gamma_{\ell,q}(X,S^c)}\ell(E_{S^0}). \tag{22}$$

If φ is defined as in (7) (which, recall, is not a norm) and if the dense noise matrix E is such that $\ell^o(e_t) \leq \varepsilon^o$ for all $t \in \mathbb{I}$, then by taking $\varepsilon = \varepsilon^o$ the set I_ε^c defined in the statement of Theorem 3 corresponds to the empty set so that (22) holds as well in this case. In connection with the concept of estimator gain discussed earlier, one can interpret the factor $2/\gamma_{\ell,q}(X,S^c)$ as an estimate of the gain (of the estimator Ψ) with respect to dense noise.

Lastly, it is interesting to see that when φ is a norm, if $E = 0$ then the result of Theorem 3 implies that $\Psi(Y,X) = \{A^o\}$ provided (19) is true.

V. Discussions on some special cases

For the purpose of illustrating the extent of the results above, let us discuss further the situation where φ reduces to a norm.

A. Scenario when the loss function is a norm

Corollary 1. *Let (Y,X) be the data generated by system* (1) *subject to the noise components E and F. Let S^0 and S^c be defined as in the statement of Theorem 3. Assume that φ is a norm i.e., it satisfies* (4)-(6) *with $\varepsilon = 0$.*
If X is self-decomposable and $|S^c| < T\big(\xi(X)\big)$, then for any $A^\star \in \Psi(Y,X)$,

$$\|A^\star - A^o\|_q \leq \mathscr{B}_{\varphi,q}(|S^0|, X)\varphi(E_{S^0}) \tag{23}$$

where

$$\mathscr{B}_{\varphi,q}(r,X) = \frac{2}{\sigma_{\varphi,q}(X)\Big[1 - \frac{N-r}{T(\xi(X))}\Big]}, \tag{24}$$

$$\sigma_{\varphi,q}(X) = \inf_{\Lambda\neq 0} \frac{\varphi(\Lambda X)}{\|\Lambda\|_q} \tag{25}$$

Proof: The principle of the proof is to show that $\gamma_{\ell,q}(X,S^c)$ is well-defined and then find a positive underestimate of it. Using the property (4) of φ and the fact that $\varphi = \ell$, we can write

$$\frac{\ell(\Lambda X_{S^0}) - \ell(\Lambda X_{S^c})}{\|\Lambda\|_q} = \frac{2\varphi(\Lambda X)}{\|\Lambda\|_q}\left[\frac{1}{2} - \frac{\varphi(\Lambda X_{S^c})}{\varphi(\Lambda X)}\right].$$

On the other hand we know from the proof of Theorem 2 (see Eq. (15)) that

$$\frac{\varphi(\Lambda X_{S^c})}{\varphi(\Lambda X)} \leq \frac{1}{2T(\xi(X))}\,|S^c|$$

so that

$$\left[1-\frac{|S^c|}{T(\xi(X))}\right]\frac{\varphi(\Lambda X)}{\|\Lambda\|_q} \leq \frac{\ell(\Lambda X_{S^0})-\ell(\Lambda X_{S^c})}{\|\Lambda\|_q}$$

Taking now the infimum on both sides of the inequality symbol over all nonzero matrices $\Lambda \in \mathbb{R}^{m\times n}$ yields

$$\sigma_{\varphi,q}(X)\left[1-\frac{|S^c|}{T(\xi(X))}\right] \leq \gamma_{\ell,q}(X,S^c).$$

It follows from the rank condition imposed on X (by the self-decomposability assumption) that $\sigma_{\varphi,q}(X) > 0$. This shows that $\gamma_{\ell,q}(X,S^c)$ is well defined and is strictly positive. Finally, since $\varphi = \ell$, invoking (22) gives the result. ∎

Two important comments can be made at this stage.

- First it is interesting to note that the bound $\mathscr{B}_{\varphi,q}(r,X)$ is an increasing function of $\xi(X)$. Therefore it is all the smaller as $\xi(X)$ is small. That is, the error bound will be small if the data matrix X is rich enough.
- Second, $\mathscr{B}_{\varphi,q}(r,X)$ is a decreasing function of r. This means that the upper bound on the estimation error decreases when the number of gross error columns in F decreases. In the extreme case where $|S^0| = N$ (no gross error), $\mathscr{B}_{\varphi,q}(|S^0|,X)$ in (23) reduces to $2/\sigma_{\varphi,q}(X)$.

Beyond these observations it should be noted that a key assumption of Corollary 1 is that $|S^c| < T\big(\xi(X)\big)$ with S^c being the index set of the nonzero columns in F. Realizing this condition requires on the one hand that the number of nonzero columns in the sparse noise matrix F be small and on the other hand that $\xi(X)$ be small[2] (which means that the data must be generic). Indeed this condition is not necessarily as strong as it might appear to be at first sight. For example, it can be relaxed as follows. Observe that the sum $E+F$ is not uniquely defined from model (2). Taking advantage of this, one can always absorb in E all nonzero columns of F whose magnitude does not exceed a certain level. To see this, let $I = \{t \in S^c : \ell(e_t+f_t) \leq \varepsilon^o\}$ where $\varepsilon^o = \max_{t\in\mathbb{I}}\ell(e_t)$. Then we can define $\tilde{E}$ and $\tilde{F}$ such that $E+F = \tilde{E}+\tilde{F}$ and $\tilde{F}_{S^0\cup I} = 0$ that is, we set $\tilde{e}_t = f_t+e_t$ and $\tilde{f}_t = 0$ for any $t \in I$ and $(\tilde{e}_t,\tilde{f}_t) = (e_t,f_t)$ otherwise. As a consequence, E and F in Corollary 1 can be replaced by $\tilde{E}$ and $\tilde{F}$ respectively so that $|S|$ and $|S^c|$ are replaced by $|S|+|I|$ and $|S^c|-|I|$. The condition of the corollary then becomes $|S^c|-|I| < T\big(\xi(X)\big)$, which is potentially easier to fulfill.

Remark 4 (sum of p-norms)**.** *Evaluating numerically the bound $\mathscr{B}_{\varphi}(r,X)$ might prove to be a hard problem due to the potential difficulty in computing the term $\sigma_{\varphi,q}(X)$ in* (25)*. A particular case of interest is when φ consists of a sum of p-norms of the column vectors, i.e. when it is defined by $\varphi(B) = \sum_{i=1}^N \|b_i\|_p$ for $B = [b_1 \quad \cdots \quad b_N]$. In this case if we take $q = 2$ in* (23) *and* (25)*, it is easy to see that $\lambda_{\min}^{1/2}(XX^\top) \leq \sigma_{\varphi,2}(X)$ with $\lambda_{\min}^{1/2}(\cdot)$ denoting the square root of the minimum eigenvalue. Replacing $\sigma_{\varphi,2}(X)$ with $\lambda_{\min}^{1/2}(XX^\top)$ in* (24) *yields an overestimate of $\mathscr{B}_{\varphi}(r,X)$ which is computable.*

Remark 5. *Corollary 1 still holds true if one replaces $T(\xi(X))$ with $\pi_\varphi^c(X)$ defined in* (11)*. As shown in [18], the number $\pi_\varphi^c(X)$ in* (11) *is computable although at the price of a combinatorial complexity. However if the n-dimension of X is small enough the complexity of the algorithm proposed there can be affordable. Then by using our formula* (24) *and Remark 4 above, it is possible therefore to obtain a smaller bound on the estimation error.*

[2]Recall that T is a decreasing function hence implying that $T(\xi(X))$ is large when $\xi(X)$ is small.

B. Single output case: ℓ_1 norm

In this section, we discuss for an illustrative purpose, the applicability of Theorem 3 to the case of single-output systems. This is an interesting case to highlight since it represents the most classical situation. Consider the single-output system defined by

$$y_t = (\theta^o)^\top x_t + f_t + e_t \tag{26}$$

where y_t, e_t, f_t are scalars and x_t and θ^o are n-dimensional vectors. By letting $Y = [y_1 \quad \cdots \quad y_N] \in \mathbb{R}^{1\times N}$ and defining E and F similarly, we obtain

$$Y = (\theta^o)^\top X + F + E. \tag{27}$$

This last equation corresponds indeed to (2) where the matrix A^o reduces to the row vector $(\theta^o)^\top$. In this case, if we let $\varphi(B) = \sum_{i=1}^N \|b_i\|_2$ then for any $\theta \in \mathbb{R}^n$, the columns of (the row vector) $Y - AX$ are scalars so that

$$\varphi(Y-\theta^\top X) = \sum_{t=1}^N \big\|y_t-\theta^\top x_t\big\|_2 = \sum_{t=1}^N \big|y_t-\theta^\top x_t\big|. \tag{28}$$

As a result, Ψ coincides in this case with the Least Absolute Deviation (LAD) estimator. The following corollary specializes the result of Theorem 3 to the LAD estimator.

Corollary 2. *Let $(Y,X) \in \mathbb{R}^{1\times N}\times\mathbb{R}^{n\times N}$ be generated by model* (26)*. Let $S^c = \big\{t \in \mathbb{I} : f_t \neq 0\big\}$, $S^0 = \mathbb{I}\setminus S^c$. Assume that X is self-decomposable and $|S^c| < T\big(\xi(X)\big)$. Then for any $\theta^\star \in \arg\min_{\theta\in\mathbb{R}^n} \big\|Y-\theta^\top X\big\|_1$,*

$$\|\theta^\star-\theta^o\|_2 \leq \mathscr{B}_{1,2}\big(|S^0|,X\big)\,\|E_{S^0}\|_1$$

where

$$\mathscr{B}_{1,2}(r,X) = \frac{2}{\sigma_{1,2}(X)\Big[1-\dfrac{N-r}{T(\xi(X))}\Big]},$$

$$\sigma_{1,2}(X) = \inf_{\eta\neq 0}\frac{\big\|X^\top\eta\big\|_1}{\|\eta\|_2}.$$

Again here the bound $\mathscr{B}_{1,2}(r,X)$ can be numerically overestimated by following the idea of Remark 4.

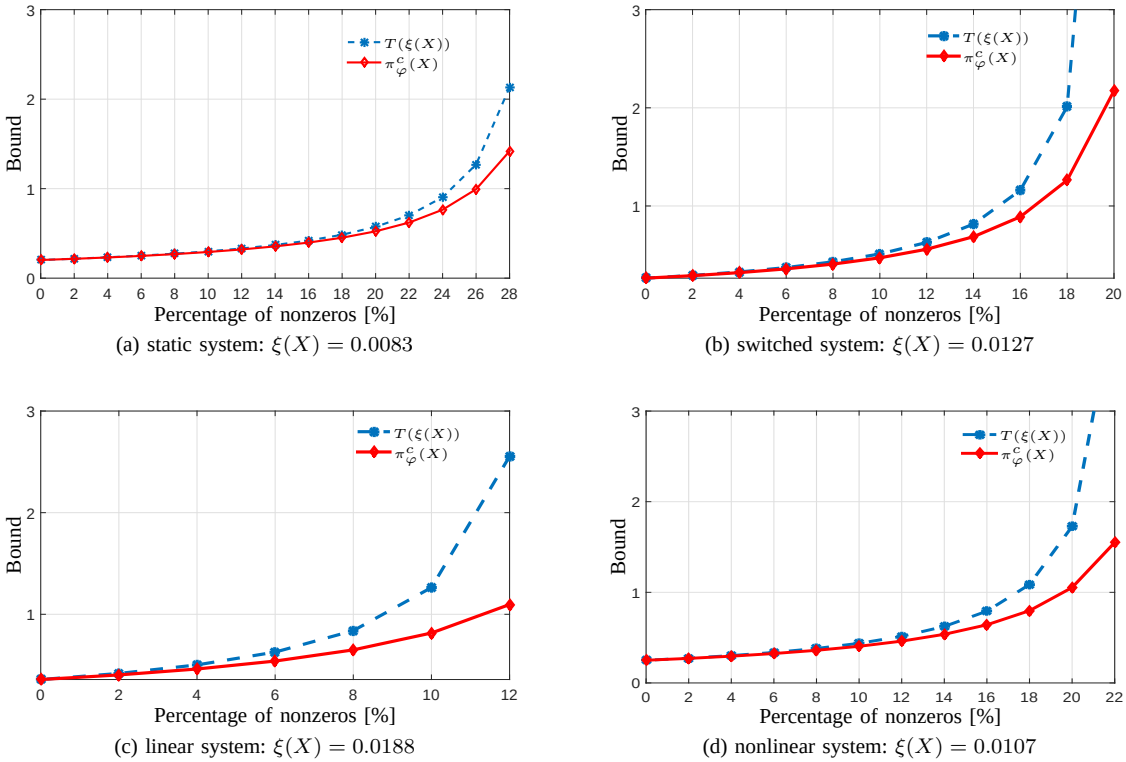
VI. Numerical illustrations

The performance of the estimator Ψ has been extensively tested in some existing papers in the special case of the LAD (see e.g., [3]) . We therefore concentrate here on evaluating numerically an estimate of the gain of the estimator based on Corollary 1 and Remark 4. The estimation is carried out for the case where φ consists in the sum of 2-norms and $q = 2$. Four different cases are studied:

(a) Static data: $X \in \mathbb{R}^{2\times 200}$ is sampled from a *Gaussian distribution* $\mathcal{N}(0,I_2)$ with zero-mean and identity-covariance.

(b) Dynamic data generated by a *switched linear system*: $X \in \mathbb{R}^{2\times 200}$ is formed with the regressors (y_{t-1},u_{t-1}) generated by a switched linear system composed of 3 subsystems of order 1. This is a switched ARX system defined by $y_t = a_{\sigma(t)}y_{t-1}+b_{\sigma(t)}u_{t-1}$ with the switching signal $\sigma(t) \in \{1,2,3\}$ generated from a uniform distribution and input u_t being a white noise with Gaussian distribution; $(a_1,b_1) = (-0.40,-0.15)$, $(a_2,b_2) = (1.55,-2.10)$ and $(a_3,b_3) = (1,-0.65)$.

(c) Dynamic data generated by a *linear ARX system* defined by $y_t = a_1y_{t-1}+b_1u_{t-1}$ with the (a_1,b_1) defined above in case (b).



(a) static system: $\xi(X) = 0.0083$

(b) switched system: $\xi(X) = 0.0127$

(c) linear system: $\xi(X) = 0.0188$

(d) nonlinear system: $\xi(X) = 0.0107$

Fig. 1: An overestimate of $\mathscr{B}_\varphi$ using respectively $\pi^c_\varphi(X)$ and $T(\xi(X))$ for a data matrix $X \in \mathbb{R}^{2\times 200}$: (a) static data sampled from a Gaussian distribution; (b) data generated by a switched system; (c) data generated by a linear dynamic system ; (d) data generated by a dynamic nonlinear system. In each case, the x-axis is limited to the range of nonzero gross errors proportions which statisfy the stability condition $|S^c|/N < T\big(\xi(X)\big)/N$ (see e.g., Corollary 1).

(d) Dynamic data generated by a *nonlinear NARX system* defined by $y_t = (y_{t-1} + 2.5)/(1 + y_{t-1}^2) + u_{t-1}$.

Following Remark 1, the columns of all data matrices X have been normalized to unit 2-norm before being processed.

Figure 1 plots the obtained estimate of the estimator gain against the proportion of correctable outliers. As remarked in Section V, the gain estimate increases as the proportion of outliers gets larger. But the growth rate of the gain estimate depends on the genericity of the data matrix X. The more generic the columns of X are, the smaller the growth rate of the estimation error is when regarded as a function of the proportion of outliers. The experiment confirms also the intuition according to which static data tend to be more generic than data generated by a dynamic system. Among the three cases of dynamic systems, the linear system appears to be the one generating the least generic data.

VII. Conclusions

In this paper we have discussed a somewhat general framework for designing a robust estimator. Given the training data, the estimator is defined as the minimizing set of a certain performance index applying to the data. We have shown that if the performance function possesses some key properties, then the so-defined estimator will inherit robustness properties. Considering a data set generated by a linear model subject to both sparse and dense noises, we showed that the estimator is insensitive to the sparse noise when this latter is acting alone and provided that the number of its nonzero components is no larger than a certain (computable) threshold. Conditions are proposed for the exact recovery of the true parameter matrix when only the sparse noise is active. When both types of noises affect the measurements we propose computable bounds on the parametric estimation error. By assuming stochasticity of the dense noise sequence, the obtained bounds are probably improvable by exploiting appropriately the statistics of the dense noise. This is a matter than can be investigated in future research.

Acknowledgement

The author is grateful to the Associate Editor and the anonymous reviewers for constructive feedback.

References

[1] L. Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47:668–677, 2011.

[2] L. Bako. Subspace clustering through parametric representation and sparse optimization. *IEEE Signal Processing Letters*, 21:356–360, 2014.

[3] L. Bako and H. Ohlsson. Analysis of a nonsmooth optimization approach to robust estimation. *Automatica*, 66:132–145, 2016.

[4] E. Candès and P. A. Randall. Highly robust error correction by convex programming. *IEEE Transactions on Information Theory*, 54:2829–2840, 2006.

[5] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Society*, 25:21–30, 2008.

[6] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52:6–18, 2006.

[7] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Birkhäuser, 2013.

[8] P. J. Huber. The place of l_1-norm in robust estimation. *Computational Statistics and Data Analysis*, 5:255–262, 1987.

[9] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. A. John Wiley & Sons, Inc. Publication (2nd Ed), 2009.

[10] A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Mathematical Programming*, 127:57–88, 2011.

[11] L. Ljung. *System Identification: Theory for the user (2nd Ed.)*. PTR Prentice Hall., Upper Saddle River, USA, 1999.

[12] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, Inc., 2006.

[13] K. Mitra, A. Veeraraghavan, and R. Chellappa. Analysis of sparse regularization based robust regression approaches. *IEEE Transactions on Signal Processing*, 61:1249–1257, 2013.

[14] N. Ozay and M. Sznaier. Hybrid system identification with faulty measurements and its application to activity analysis. In *IEEE Conference on Decision and Control and European Control Conference, Orlando, FL, USA*, 2011.

[15] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. A sparsification approach to set membership identification of a class of affine hybrid systems. *IEEE Transactions on Automatic Control*, 57:634–648, 2012.

[16] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[17] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., 2005.

[18] Y. Sharon, J. Wright, and Y. Ma. Minimum sum of distances estimator: Robustness and stability. In *American Control Conference, St Louis, MO, USA*, 2009.

[19] T. Soderstrom and P. Stoica. *System identification*. Prentice Hall, Upper Saddle River, USA, 1989.

[20] A. M. Tillmann and M. E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60:1248–1259, 2014.

[21] W. Xu, E.-W. Bai, and M. Cho. System identification in the presence of outliers and random noises: A compressed sensing approach. *Automatica*, 50:2905–2911, 2014.

[22] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26:301–320, 1981.