Automatica 66 (2016) 132–145

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Analysis of a nonsmooth optimization approach to robust estimation[☆]

Laurent Bako[a,1], Henrik Ohlsson[b,c]

[a] Laboratoire Ampère, Ecole Centrale de Lyon, Université de Lyon, France
[b] Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden
[c] Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 8 February 2014
Received in revised form
30 September 2015
Accepted 23 December 2015

Keywords:
Robust estimation
Outliers
System identification
Nonsmooth optimization

ABSTRACT

In this paper, we consider the problem of identifying a linear map from measurements which are subject to intermittent and arbitrarily large errors. This is a fundamental problem in many estimation-related applications such as fault detection, state estimation in lossy networks, hybrid system identification, robust estimation, etc. The problem is hard because it exhibits some intrinsic combinatorial features. Therefore, obtaining an effective solution necessitates relaxations that are both solvable at a reasonable cost and effective in the sense that they can return the true parameter vector. The current paper discusses a nonsmooth convex optimization approach and provides a new analysis of its behavior. In particular, it is shown that under appropriate conditions on the data, an exact estimate can be recovered from data corrupted by a large (even infinite) number of gross errors.

1. Introduction

1.1. Problem and motivations

We consider a linear measurement model of the form

$$y_t = x_t^\top \theta^\circ + f_t + e_t \tag{1}$$

where $y_t \in \mathbb{R}$ is the measured signal, $x_t \in \mathbb{R}^n$ the regression vector, $\{e_t\}$ a sequence of *zero-mean and bounded* errors (e.g., measurement noise, model mismatch, uncertainties, etc.) and $\{f_t\}$ a sequence of *intermittent and arbitrarily large* errors. Assume that we observe the sequences $\{x_t\}_{t=1}^N$ and $\{y_t\}_{t=1}^N$ and would like to compute the parameter vector θ° from these observations. We are interested in doing so without knowing any of the sequences $\{f_t\}$ and $\{e_t\}$. We do however make the following assumptions:

- $\{e_t\}$ is a bounded sequence.
- $\{f_t\}$ is a sequence containing zeros and intermittent gross errors with (possibly) arbitrarily large magnitudes.

This is an important estimation problem arising in many situations such as fault detection (Chen, Bako, & Lecoeuche, 2011; Ozay & Sznaier, 2011), hybrid system identification (Garulli, Paoletti, & Vicino, 2012), subspace clustering (Bako, 2014; Vidal, 2010), error correction in communication networks (Candès & Randall, 2006). The case when $\{f_t\}$ is zero and $\{e_t\}$ is a Gaussian process has been well-studied in linear system identification theory (see, e.g., the text books Ljung, 1999; Soderstrom & Stoica, 1989). A less studied, but very relevant scenario in the system identification community, is when the additional perturbation $\{f_t\}$ in (1) is nonzero and contains intermittent and arbitrarily large errors (Candès & Randall, 2006; Mitra, Veeraraghavan, & Chellappa, 2013; Sharon, Wright, & Ma, 2009; Xu, Bai, & Cho, 2014). It is worth noticing the difference with the problem studied in the field of compressive sensing (Candès & Randall, 2006; Candès & Wakin, 2008; Donoho, 2006). In compressive sensing, the sought parameter vector is assumed sparse and the measurement noise $\{e_t\}$, often Gaussian or bounded. Here, no assumptions are made concerning sparsity of θ°. We will, in this contribution, study essentially the case when the data is noise-free (*i.e.*, $e_t = 0$ for all t) and $\{f_t\}$ is a sequence with intermittent gross errors. We will derive conditions for perfect recovery and point to effective algorithms for computing θ°. In the second part of the paper, the model assumption is relaxed to allow both e_t and f_t to be simultaneously nonzero. Note that this might be a more realistic scenario since most applications have measurement noise.

For illustrative purposes, let us discuss briefly some applications where a model of the form (1) is of interest.

[☆] The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Söderström.

E-mail addresses: laurent.bako@ec-lyon.fr (L. Bako), ohlsson@berkeley.edu (H. Ohlsson).

[1] Tel.: +33 472 186 452.

http://dx.doi.org/10.1016/j.automatica.2015.12.024
0005-1098/© 2016 Elsevier Ltd. All rights reserved.

Switched linear system identification. A discrete-time Multi-Input Single-Output (MISO) Switched Linear System (SLS) can be written in the form

$$y_t = x_t^\top \theta^o_{\sigma_t} + e_t, \tag{2}$$

where $x_t \in \mathbb{R}^n$ is the regressor at time $t \in \mathbb{Z}_+$ defined by

$$x_t = \begin{bmatrix} y_{t-1} & \cdots & y_{t-n_a} & u_t^\top & u_{t-1}^\top & \cdots & u_{t-n_b}^\top \end{bmatrix}^\top, \tag{3}$$

where $u_t \in \mathbb{R}^{n_u}$ and $y_t \in \mathbb{R}$ denote respectively the input and the output of the system. The integers n_a and n_b in (3) are the maximum output and input lags (also called the orders of the system). $\sigma_t \in \{1, \ldots, s\}$ is the discrete mode (or discrete state) indexing the active subsystem at time t; it is in general assumed *unobserved*. $\theta^o_{\sigma_t} \in \mathbb{R}^n$, $n = n_a + n_b n_u$, is the parameter vector (PV) associated with the mode σ_t. For $\theta^\circ \in \{\theta_1{}^\circ, \ldots, \theta_s^o\}$, the Switched Auto-Regressive eXogenous (SARX) model (2) can be written in the form (1), with unknown f_t of the following structure $f_t = x_t^\top(\theta^o_{\sigma_t} - \theta^o)$. For a background on hybrid system identification, we refer to the references (Bako, 2011; Garulli et al., 2012; Maruta & Sugie, 2011; Ohlsson & Ljung, 2013; Ozay, Sznaier, Lagoa, & Camps, 2012; Paoletti, Juloski, Ferrari-Trecate, & Vidal, 2007; Vidal, Soatto, Ma, & Sastry, 2003).

Identification from faulty data. A model of the form (1) also arises when one has to identify a linear dynamic system which is subject to intermittent sensor faults. This is the case in general when the data are transmitted over a communication network (Candès & Randall, 2006; Ozay & Sznaier, 2011). Model (1) is suitable for such situations and the sequence $\{f_t\}$ then models occasional data packets losses or potential outliers. More precisely, a dynamic MISO system with process faults can be directly written in the form (1). In the case of sensor faults, the faulty model might be defined by

$$\begin{cases} \bar{y}_t = \bar{x}_t^\top \theta^\circ + e_t \\ y_t = \bar{y}_t + w_t \end{cases}$$

where $y_t \in \mathbb{R}$ is the *observed output* which is affected by the fault w_t (assumed to be nonzero only occasionally) ; $\bar{x}_t$ is defined as in (3) from the *known input* u_t and the *unobserved output* $\bar{y}_t$. We can rewrite the faulty model exactly in the form (1) with $f_t = w_t - \begin{bmatrix} w_{t-1} & \cdots & w_{t-n_a} \end{bmatrix} \theta^\circ$. Sparsity of $\{w_t\}$ induces sparsity of $\{f_t\}$ but in a lesser extent.

State estimation in the presence of intermittent errors. Considering a MISO dynamic system with state dynamics described by $z_{t+1} = Az_t + Bu_t$ and observation equation $\tilde{y}_t = C^\top z_t + f_t$, (A, B, C) being known matrices of appropriate dimensions, and $\{f_t\}$ a sparse sequence of possibly very large errors, the finite horizon state estimation problem reduces to the estimation of the initial state $z_0 = \theta$. We get a model of the form (1) by setting $y_t = \tilde{y}_t - C^\top \Delta_t \bar{u}_t$ and $x_t = (A^t)^\top C$, with $\Delta_t = \begin{bmatrix} A^{t-1}B & \cdots & AB & B \end{bmatrix}$, $\bar{u}_t = \begin{bmatrix} u_0^\top & \cdots & u_{t-1}^\top \end{bmatrix}^\top$. Examples of relevant works are those reported in Bako and Lecoeuche (2013), Fawzi, Tabuada, and Diggavi (2014). In this latter application, it can however be noted that the dataset $\{x_t\}$ may not be generic enough.[2]

Connection to robust statistics. Indeed, the problem of identifying the parameters from model (1) under the announced assumptions can be viewed as a robust regression problem where the nonzero elements in the sequence $\{f_t\}$ are termed outliers. As such, it has received a lot of attention in the robust statistics literature (see, e.g., Huber & Ronchetti, 2009; Maronna, Martin, & Yohai, 2006; Rousseeuw & Leroy, 2005 for an overview). Examples of methods to tackle the robust estimation problem include the least absolute deviation (Huber, 1987), the least median of squares (Rousseeuw, 1984), the least trimmed squares (Rousseeuw & Leroy, 2005), the M-estimator (Huber & Ronchetti, 2009), etc. Most of these estimators come with an analysis in terms of the breakdown point (Hampel, 1971; Seber & Lee, 2003), a measure of the (asymptotic) minimum proportion of points which cause an estimator to be unbounded if they were to be arbitrarily corrupted by gross errors. The current report focuses on the analysis of a nonsmooth convex optimization approach which includes the least absolute deviation method as a particular case corresponding to the situation when the output in (1) is a scalar. The analysis approach taken in the current paper is different in the following sense.

- In robust statistics the quality of an estimator is measured by its breakdown point. The higher the breakdown point, the better. The available analysis is therefore directed to determining a sort of absolute robustness: how many outliers (expressed in proportion of the total number of samples) cause the estimator to become unbounded.
- Here, the question of robust performance of the estimator is posed differently. We are interested in estimating the maximum number of outliers that a nonsmooth-optimization-based estimator can accommodate while still returning the exact value one would obtain in the absence of any outlier. This is more related to the traditional view developed in compressive sensing.

Contributions of this paper. One promising method for estimating model (1) is by nonsmooth convex optimization as suggested in Candès and Randall (2006), Sharon et al. (2009), Bako (2011), Mitra et al. (2013) and Xu et al. (2014). More precisely, inspired by the recent theory of compressed sensing (Candès & Randall, 2006; Candès & Wakin, 2008; Donoho, 2006), the idea is to minimize a nonsmooth (and non differentiable) sum-of-norms objective function involving the fitting errors. Under noise-free assumption, such a cost function has the nice property that it is able to provide the true parameter vector in the presence of arbitrarily large errors $\{f_t\}$ provided that the number of nonzero errors is small in some sense. Of course, when the data are corrupted simultaneously by the noise $\{e_t\}$ and the gross errors $\{f_t\}$, the recovery cannot be exact any more. It is however expected (as Proposition 17 and simulations tend to suggest) that the estimate will still be close to the true one.

The current paper intends to present a new analysis of the nonsmooth optimization approach and provide some elements for further understanding its behavior. The line of analysis goes from a full characterization of the nonsmooth optimization based estimator (both for SISO and MIMO systems) to the study of robustness to outliers including in the presence of dense noise. With respect to relevant works (Bako, 2011; Candès & Randall, 2006; Mitra et al., 2013; Sharon et al., 2009; Xu et al., 2014), we derive new bounds on the number of outliers (in the least favorable situations) that the estimator is capable to accommodate. It is emphasized that a quite broad spectrum of such bounds can be derived based on the basic characterization of the nonsmooth identifier. Note however that evaluating numerically the tightest of these bounds is a high computational process while less tight bounds have a more affordable complexity. Some of the bounds developed in this contribution meet both relative tightness requirement and computability in polynomial time (see the bound based on $\xi(X)$ in Theorem 11). Finally, the paper show how the results derived in the first part for ℓ_1-norm estimator when

[2] In this paper, the term genericity for a dataset characterizes a notion of linear independence. For example, a set of $N > n$ data points in general linear position in $\mathbb{R}^n$ is more generic than a set of data points contained in one subspace. We will introduce different quantitative measures of data genericity in the sequel (see Definition 2 and Theorem 11).