Automatica 50 (2014) 2905–2911

Contents lists available at ScienceDirect

Automatica

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

Brief paper

System identification in the presence of outliers and random noises: A compressed sensing approach ☆

Weiyu Xu [a], Er-Wei Bai [a,b,1], Myung Cho [a]

[a] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, United States

[b] School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, UK

ARTICLE INFO

Article history:
Received 14 May 2013
Received in revised form
21 April 2014
Accepted 2 July 2014
Available online 28 October 2014

Keywords:
System identification
Least absolute deviation
ℓ_1 minimization
Toeplitz matrix
Compressed sensing

ABSTRACT

In this paper, we consider robust system identification of FIR systems when both sparse outliers and random noises are present. We reduce this problem of system identification to a sparse error correcting problem using a Toeplitz structured real-numbered coding matrix and prove the performance guarantee. Thresholds on the percentage of correctable errors for Toeplitz structured matrices are established. When both outliers and observation noise are present, we have shown that the estimation error goes to 0 asymptotically as long as the probability density function for observation noise is not "vanishing" around origin. No probabilistic assumptions are imposed on the outliers.

1. Introduction

In a linear system identification setting, an unknown system parameter vector $\mathbf{x} \in R^m$ is often observed through a Toeplitz matrix $H \in R^{n\times m}$ $(n \geq m)$, namely

$$\mathbf{y} = H\mathbf{x} = \begin{bmatrix} u_1 & u_0 & \cdots & u_{-m+2} \\ u_2 & u_1 & \cdots & u_{-m+3} \\ \vdots & \vdots & \ddots & \vdots \\ u_n & u_{n-1} & \cdots & u_{-m+n+1} \end{bmatrix} \mathbf{x}, \tag{1.1}$$

where H is a Toeplitz matrix with u_i, $-m+2 \leq i \leq n$, being the system input sequence and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ the system output. In this paper, we denote this system input sequence by a row vector $h = (u_{-m+2}, u_{-m+1}, \ldots, u_n)$.

In this paper, we consider system identification under finite impulse response (FIR) models. Though applicable to control applications where FIR models are used, it is not as complete as IIR models from a control point of view. However, identification in the presence of outliers and random noises is not limited to control applications and in fact routinely applied to many other areas, e.g., signal processing and communication (Rauhut, 2009; Sanandaji, Vincent, Poolla, & Wakin, 2012). Note in those areas, the systems are dominantly and overwhelmingly FIR models and thus the results derived in this paper can be readily applied.

If there is no interference or noise in the observation $\mathbf{y}$, one can then simply recover $\mathbf{x}$ from matrix inversion. However, in applications, the observations $\mathbf{y}$ are corrupted by noises and a few elements can be exposed to large-magnitude gross errors or outliers. Mathematically, when both additive observation noise and outliers are present, the observation $\mathbf{y}$ can be written as

$$\mathbf{y} = H\mathbf{x} + \mathrm{e} + \mathbf{w}, \tag{1.2}$$

where e is a sparse outlier vector with $k \ll n$ non-zero elements, and $\mathbf{w}$ is a measurement noise vector with each element usually being assumed to be i.i.d. random variables. We further assume that m is fixed and n can increase, which is often the case in system identifications (Ljung, 1987).

If only random measurement errors are present, the least-square solutions generally provide an asymptotically good estimate. However, the least-square estimate breaks down in the presence of outliers. Thus, it is necessary to protect the estimates from both random noise and outliers. Research along this direction has attracted a significant amount of attention, for example,

☆ This work was supported in part by NSF CNS-1329657, the Simons Foundation and DoE DE-FG52-09NA29364. The material in this paper was partially presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), May 26–31, 2013, Vancouver, British Columbia, Canada. This paper was recommended for publication in revised form by Associate Editor Alessandro Chiuso under the direction of Editor Torsten Söderström.

E-mail addresses: weiyu-xu@uiowa.edu (W. Xu), er-wei-bai@uiowa.edu (E.-W. Bai), myung-cho@uiowa.edu (M. Cho).

[1] Tel.: +1 3193355949; fax: +1 3193356028.

http://dx.doi.org/10.1016/j.automatica.2014.10.017
0005-1098/

Bai, Cho, Tempo, and Ye (2002), Ljung (1987), Rousseeuw and Leroy (1987) and Söderström and Stoica (1989). An effective way is to visually inspect the residual plot and change the obviously erroneous measurements "by hand" to an appropriately interpolated values (Ljung, 1987). The approach does not however always work. The need for human intervention, which prevents automatic/adaptive implementation, is an additional shortcoming of the visual inspection of the residual plot as a means to account for the presence of outliers. Another approach to deal with the outliers was the idea of few violated constraints (Bai et al., 2002) in the setting of the bounded error parameter estimation. The other two popular methods in the statistical literature to deal with the outliers are the least median squares and the least trimmed squares (Huber, 1981; Rousseeuw & Leroy, 1987). Instead of minimizing the sum of the residual squares, the least median squares yield the smallest value for the median of squared residuals computed from the entire data set and the least trimmed squares try to minimize the sum of squared residues over a subset of the given data. Both have shown robustness against the outliers (Huber, 1981; Rousseeuw & Leroy, 1987). The problem is their computational complexity. Both algorithms are nonlinear and in fact combinatory in nature. This limits their practical applications if n and/or m are not small or even modest. The most popular way to deal with the outliers in the statistical literature is the least absolute deviation estimate (ℓ_1 minimization) which has been extensively studied (Candès & Randall, 2008; Candès & Tao, 2005; Chen, Bai, & Zao, 1990; Xu, Wang, & Tang, 2011). Instead of searching for all the $\binom{n}{k}$ possibilities for the locations of outliers, Candès and Randall (2008), Candès and Tao (2005) and Chen et al. (1990) proposed to minimize the least absolute deviation,

$$\hat{\mathbf{x}} = \arg\min_{\xi} \|\mathbf{y} - H\xi\|_1, \tag{1.3}$$

where $\hat{\mathbf{x}}$ is an estimate of $\mathbf{x}$. Under the assumption that the error $\mathrm{e} + \mathbf{w}$ is an i.i.d. random sequence with a common density which has median zero and is continuous and positive in the neighborhood of zero, the difference between the unknown $\mathbf{x}$ and its estimate is asymptotically Gaussian of zero mean (Chen et al., 1990). The problem is that the assumption of i.i.d. of median zero on the unknown outliers is very restrictive and seldomly satisfied in reality. We study the least absolute deviation estimator or ℓ_1 minimization from the compressed sensing point of view and show that i.i.d. of median zero on the outliers are unnecessary. In fact only the number of outliers relative to the total number of data length plays a role.

Recovering signals from outliers or errors have been studied (Candès & Randall, 2008; Candès & Tao, 2005; Xu & Hassibi, 2011; Xu et al., 2011). In their setting, each element of the $(n-m) \times n$ matrix A such that $AH = 0$, is assumed to be i.i.d. random variables following a certain distribution, for example, Gaussian distribution or Bernoulli distribution. These types of matrices have been shown to obey certain conditions such as restricted isometry conditions (Candès & Randall, 2008) so that (1.3) can correctly recover $\mathbf{x}$ when there are only outliers present; and can recover $\mathbf{x}$ approximately when both outliers and measurement noise exist. However, in the system identification problem, H has a natural Toeplitz structure and the elements of H are not independent but correlated. The natural question is whether (1.3) also provides performance guarantee for recovering $\mathbf{x}$ with a Toeplitz matrix. We provide a positive answer in this paper. The main contribution of this paper is the establishment of the performance guarantee of Toeplitz structured matrices in parameter estimation in the presence of both outliers and random noises.

With the development of compressed sensing theory in recent years, the role of ℓ_1 regularization has been studied in system identification (Chen, Gu, & Hero, 2009; Kopsinis, Slavakis, & Theodoridis, 2011; Rauhut, 2009; Romberg, 2009; Sanandaji et al., 2012). In these works, system parameters are often assumed to be sparse, and then ℓ_1 regularization can be used to reduce the number of needed samples for system identification. This paper instead considers system identification under sparse outliers without requiring the system state to be sparse. We would like to point out that there is a well known duality between compressed sensing (Donoho, 2006; Donoho & Tanner, 2005) and sparse error detection (Candès & Randall, 2008; Candès & Tao, 2005): the null space of sensing matrices in compressed sensing corresponds to the tall matrix H in sparse error corrections. Toeplitz and circulant matrices have been studied in compressed sensing in several papers (Rauhut, 2009; Romberg, 2009). In these papers, it has been shown that Toeplitz matrices are good for recovering sparse vectors from undersampled measurements. In contrast, in our model, the signal itself is *not* sparse and the linear system involved is overdetermined rather underdetermined. Also, the null space of a Toeplitz matrix does not necessarily correspond to another Toeplitz matrix. Thus, the problem studied in this paper is essentially different from those studied by Rauhut (2009) and Romberg (2009).

The rest of this paper is organized as follows. In Section 2, we derive the convergence results when both outliers and random noises are present. In Section 3, we derive the worst-case performance bounds on the number of outliers in ℓ_1 minimization when only outliers are present. In Section 4, we extend our results to non-Gaussian inputs in system identification. In Section 5, we provide numerical results and Section 6 concludes the paper by discussing extensions and future directions.

2. Average-case performance bounds: with both outliers and observation noises

We consider the case when both outliers and random observation errors are present and show that, under mild conditions, the identification error $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ goes to 0, where $\hat{\mathbf{x}}$ is the solution to (1.3).

Theorem 2.1. *Let m be a fixed positive integer and H be an $n \times m$ Toeplitz matrix ($m < n$) in (1.1) with each element u_i, $-m+2 \leq i \leq n$, being i.i.d. $N(0, 1)$ Gaussian random variables. Suppose*

$$\mathbf{y} = H\mathbf{x} + \mathrm{e} + \mathbf{w},$$

where e is a sparse vector with $k \leq \beta n$ non-zero elements ($\beta < 1$ is a constant) and $\mathbf{w}$ is the observation noise vector. For any constant $t > 0$, we assume that, with probability 1 as $n \to \infty$, at least $\alpha(t)n$ (where $\alpha(t) > 0$ is a constant depending on t) elements in $\mathbf{w} + \mathrm{e}$ are no bigger than t in amplitude.

Then $\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \to 0$ in probability as $n \to \infty$, where $\hat{\mathbf{x}}$ is the solution to (1.3).

We remark that, in Theorem 2.1, the condition on the unknown outlier vector is merely $\beta < 1$, and the condition on the random noise $\mathbf{w}$ is weaker than the usual condition of having i.i.d. elements with median 0 (Chen et al., 1990). In fact, if $\mathbf{w}$ is independent from e, and the elements of $\mathbf{w}$ are i.i.d. random variables following a distribution which is not "vanishing" in an arbitrarily small region around 0 (namely the cumulative distribution function $F(t)$ satisfies that $F(t) - F(-t) > 0$ for any $t > 0$). Note that the probability density function $f(t)$ is allowed to be 0, however, the conditions in Theorem 2.1 will be satisfied. To see that, first observe that $(1-\beta)n$ elements of the outlier vector are zero. If elements of $\mathbf{w}$ are i.i.d. following a probability density function $f(s)$ that is not "vanishing" around $s = 0$, with probability converging to one as $n \to \infty$, at least $[F(t) - F(-t)](1-\beta)(1-\epsilon)n = \alpha(t)n$ elements of the vector $e + \mathbf{w}$ are no bigger than t, where $\epsilon > 0$