Automatica 87 (2018) 218–225

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Robustness analysis of a maximum correntropy framework for linear regression☆

Laurent Bako

Laboratoire Ampère – Ecole Centrale de Lyon – Université de Lyon, France

ARTICLE INFO

Article history:
Received 14 March 2017
Received in revised form 30 July 2017
Accepted 31 August 2017

Keywords:
Robust estimation
System identification
Maximum correntropy
Outliers

ABSTRACT

In this paper we formulate a solution of the robust linear regression problem in a general framework of correntropy maximization. Our formulation yields a unified class of estimators which includes the Gaussian and Laplacian kernel-based correntropy estimators as special cases. An analysis of the robustness properties is then provided. The analysis includes a quantitative characterization of the informativity degree of the regression which is appropriate for studying the stability of the estimator. Using this tool, a sufficient condition is expressed under which the parametric estimation error is shown to be bounded. Explicit expression of the bound is given and discussion on its numerical computation is supplied. For illustration purpose, two special cases are numerically studied.

1. Introduction

Given a set of empirical observations generated by a system along with a class of parameterized candidate models, a parameter estimator is a function which maps the available data to the parameter space associated with the model class. A very desirable property for an estimator is that of robustness which characterizes a relative insensitivity of the estimator to deviations of the observed data from the assumed model. More specifically, this property is central in situations where the data are prone to non Gaussian noise or disturbances of possibly arbitrarily large amplitude (often called outliers). The quest for robust estimators has led to the development of many estimators such as the Least Absolute Deviation (LAD) (Bako & Ohlsson, 2016; Candès & Randall, 2006; Maronna, Martin, & Yohai, 2006; Rousseeuw & Leroy, 2005), the least median of squares (Rousseeuw, 1984), the least trimmed squares (Rousseeuw & Leroy, 2005), the class of M-estimators (Huber & Ronchetti, 2009). Evaluating formally to what extent a given estimator is robust requires setting a quantitative measure of robustness. Incidentally such a measure can serve as comparison criterion between different robust estimators. Generally, the robustness is assessed in term of the maximum proportion of outliers in the total data set that the estimator can handle while remaining stable (see for example the concept of breakdown point (Rousseeuw & Leroy, 2005)). More recently the maximum correntropy (Liu, Pokharel, & Principe, 2007; Principe, 2010; Santamaria, Pokharel, & Principe, 2006) has emerged as an information-theoretic estimation framework which induces some robustness properties with respect to outliers. Although maximum correntropy estimation is closely related to M-estimation, its discovery has broadened the horizon of possibilities for designing robust identification schemes. As a matter of fact, it has been successfully applied to a variety of estimation problems such as linear/nonlinear regression, filtering, face recognition in computer vision (Chen, Xing, Liang, Zheng, & Principe, 2014; Feng, Huang, Shi, Yang, & Suykens, 2015; He, Zheng, & Hu, 2011).

Contribution. Although the maximum correntropy based estimators have been gaining an increasing success, the formal analysis of its robustness properties is still a largely open research question. In this paper we propose such an analysis for a class of maximum correntropy based estimators applying to linear regression problems. More precisely, the contribution of the current paper is articulated around the following three questions:

- To what extent the maximum correntropy estimation framework is robust to outliers? By robustness, it is meant here a certain insensitivity of the estimator to large errors of possibly arbitrarily large magnitude. To address this question, we derive parametric estimation error bounds induced by the estimator in function of both the degree of richness of the regression data and on the fraction of outliers. In summary, we show that if the regression data enjoy some richness properties and if the number of outliers is reasonably small, then the parametric estimation error remains stable. Indeed the proportion of outliers that the estimator is capable to correct depends on how rich the regressor matrix

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

E-mail address: laurent.bako@ec-lyon.fr.

https://doi.org/10.1016/j.automatica.2017.09.006
0005-1098/

is. Moreover, the estimation error appears to be a decreasing function of the richness measure.
- How does richness of the training data set influence the robustness of the estimator and how to characterize it? We provide an appropriate characterization of the richness in terms of the cardinality of the regressor vectors which are strongly correlated to any vector of the regression space. As such however, this quantitative measure of richness is not computable at an affordable price. To alleviate this difficulty the paper proposes some estimates of this measure thus allowing for the approximation of the parametric estimation error bounds.
- Does the maximum correntropy estimator (MCE) possess the exact recovery property? We show that unlike the LAD estimator, the MCE is not able to return exactly the true parameter vector once the measurement is affected by a single arbitrary nonzero error. The proof is given for the Gaussian kernel based estimator.

We note that an analysis of robustness of the maximum correntropy has been presented recently in Chen, Liu, Zhao, Zheng, and Principe (in press) and Chen, Xing, Zhao, Xu, and Principe (2017). However the analysis there is limited to the Gaussian kernel based correntropy and to a single parameter estimation problem. Moreover these works do not make clear how the properties of the data contribute to the robustness of the estimator.

Outline. The remainder of this paper is organized as follows. Section 2 presents the robust regression problem and define the class of maximum correntropy estimators whose properties are to be studied in the paper. It also introduces the general setting of the paper. The main analysis results are developed in Section 3. In Section 4 we run numerical experiments to illustrate the richness measure and the evolution of the derived error bounds with respect to the amount of noise. Finally, Section 5 contains concluding remarks concerning this work.

Notations. $\mathbb{R}$ is the set of real numbers; $\mathbb{R}_+$ is the set of real nonnegative numbers; $\mathbb{N}$ is the set of natural integers; $\mathbb{C}$ denotes the set of complex numbers. N will denote the number of data points and $\mathbb{I} = \{1, \ldots, N\}$ the associated index set. For any finite set $\mathcal{S}$, $|\mathcal{S}|$ refers to the cardinality of $\mathcal{S}$. However, whenever x is a real (respectively complex) number, $|x|$ will refer to the absolute value (respectively modulus) of x. For $x = [x_1 \quad \cdots \quad x_n]^\top \in \mathbb{R}^n$, $\|x\|_p$ will denote the p-norm of x defined by $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$, for $p \in \{1, 2\}$, $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$. The exponential of a real number z will be denoted $\exp(z)$ or e^z according to visual convenience; $\ln(z)$ is the natural logarithm function. For a square and positive semi-definite matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote, respectively, the minimal and maximal eigenvalues of A.

2. Robust regression problem

2.1. The data-generating system

Let $\{x_t\}_{t\in\mathbb{N}}$ and $\{y_t\}_{t\in\mathbb{N}}$ be some stochastic processes taking values, respectively, in $\mathbb{R}^n$ and $\mathbb{R}$. They are assumed to be related by an equation of the form

$$y_t = x_t^\top \theta^o + v_t, \tag{1}$$

where $\{v_t\}_{t\in\mathbb{N}}$ represent an unobserved error sequence; $\theta^o \in \mathbb{R}^n$ is an unknown parameter vector. Eq. (1) may describe a static (memoryless) system or a dynamic one. In the latter case, we will conveniently assume that the so-called regressor (or explanatory vector) x_t has the following structure $x_t = [u_t \quad u_{t-1} \quad \cdots \quad u_{t-(n-1)}]^\top$, i.e., (1) is an FIR-type (Finite Impulse Response) system, with u_t then denoting its input signal at time t.

Assumption 1. The joint stochastic process $\{(x_t, v_t)\}_{t\in\mathbb{N}}$ is independently and identically distributed.

While this assumption can hold naturally for a static system, it might not be satisfied in some practical situations. For example, if (1) is a dynamic system (for instance, of FIR-type), this assumption is not satisfied.[1] But as will be seen, its only role is to highlight the correntropic origin of the estimation framework considered in this paper.

Assumption 2. The noise sequence $\{v_t\}$ satisfies the following: there is $\varepsilon \geq 0$ such that if we define the index sets $I_\varepsilon^0 = \{t : |v_t| \leq \varepsilon\}$ and $I_\varepsilon^c = \{t : |v_t| > \varepsilon\}$, then the cardinality of $\left|I_\varepsilon^0\right|$ is "much larger" than that of $\left|I_\varepsilon^c\right|$.

We will formalize latter in the paper what "much larger" can mean. Similarly as in Bako and Ohlsson (2016), we can assume that v_t is of the form $v_t = f_t + e_t$ where $\{f_t\}$ is a sparse noise sequence in the sense that only a few elements of it are different from zero. However its nonzero elements are allowed to take on arbitrarily large values (called in this case, outliers). As to $\{e_t\}$, it is assumed to be a bounded and dense (i.e., not necessarily sparse) noise sequence of rather moderate amplitude.

Problem. Given a finite collection $Z^N = \{(x_t, y_t)\}_{t=1}^N$ of measurements obeying the system equation (1), the robust regression problem of interest here is the one of finding a reliable estimate of the parameter vector θ^o despite the effect of arbitrarily large errors.

Let θ denote a candidate parameter vector (PV) which we would like, ideally, to coincide with the true PV θ^o. Given x_t and θ, the prediction we can make of y_t is $\hat{y}_t(\theta) = x_t^\top \theta$. It is then the goal of the estimation method to select θ such that y_t and $\hat{y}_t(\theta)$ are close in some sense for any t. Closeness will be measured in term of the so-called maximum correntropy between the measured output y_t and the predicted value $\hat{y}_t(\theta)$.

2.2. Maximum correntropy estimation

The correntropy is an information-theoretic measure of similarity between two arbitrary random variables (Liu et al., 2007; Santamaria et al., 2006). More specifically, consider two random variables Y and $\hat{Y}$ defined on the same probability space, and taking values in $\mathbb{R}$. Let $\phi_\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a positive-definite kernel function on $\mathbb{R}$ (see e.g., Schölkopf and Smola (2002, Chap. 2, p. 30) for a definition). The correntropy $V_{\phi_\ell}(Y, \hat{Y})$ between Y and $\hat{Y}$ with respect to a kernel function ϕ_ℓ, is defined by

$$V_{\phi_\ell}(Y, \hat{Y}) = \mathbb{E}_{Y,\hat{Y}}\left[\phi_\ell(Y, \hat{Y})\right],$$

where $\mathbb{E}_{Y,\hat{Y}}[\cdot]$ refers to the expected value with respect to the joint distribution of $(Y, \hat{Y})$. In a more explicit form, we have

$$V_{\phi_\ell}(Y, \hat{Y}) = \int_{\mathbb{R}}\int_{\mathbb{R}} \phi_\ell(y, \hat{y}) p_{Y,\hat{Y}}(y, \hat{y}) dy d\hat{y} \tag{2}$$

with $p_{Y,\hat{Y}}$ being the joint probability density function of $(Y, \hat{Y})$. The correntropy constitutes a similarity measure between Y and $\hat{Y}$ through the kernel ϕ_ℓ. Although the original definition of correntropy in Santamaria et al. (2006) fixes ϕ_ℓ to be the Gaussian kernel, it is indeed possible to extend it to any positive definite kernel function.

We consider in this paper a kernel function of the form

$$\phi_\ell(y, \hat{y}) = \exp(-\gamma \ell(y - \hat{y})), \tag{3}$$

where $\gamma > 0$ is a user-specified parameter and $\ell : \mathbb{R} \to \mathbb{R}_+$ is a function which satisfies the following properties:

[1] Indeed this assumption can be relaxed to an appropriate notion of stationarity and ergodicity for the joint process $\{(x_t, v_t)\}$.