Robustness analysis of a maximum correntropy framework for linear regression

Laurent Bako

Robustness analysis of a Maximum Correntropy framework for linear regression

Laurent Bako [a]

[a] *Laboratoire Ampère – Ecole Centrale de Lyon – Université de Lyon, France*

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.

Key words: robust estimation, system identification, maximum correntropy, outliers.

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) [17,14,4,2], the least median of squares [16], the least trimmed squares [17], the class of M-estimators [11]. 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 [17]). More recently the maximum correntropy [18,12,15] 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 [9,7,10].

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 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 [5,6]. 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 rest 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 $x = [x_1 \cdots x_n]^\top \in \mathbb{R}^n$, $\|x\|_p$ 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 $x_t = [u_t \ u_{t-1} \ \cdots \ u_{t-(n-1)}]^\top$, 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 [2]. 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 $|I_\varepsilon^0|$ is "much larger" than that of $|I_\varepsilon^c|$.*

We will formalize latter in the paper what "much larger" can mean. Similarly as in [2], 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

---

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

$\hat{y}_t(\theta) = x_t^\top \theta$. 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 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 [18,12]. 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., [19, 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

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

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 [18] 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

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

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

P1. ℓ is positive-definite: $\ell(a) \geq 0 \ \forall a$ and $\ell(a) = 0$ if and only if $a = 0$.
P2. ℓ is symmetric: $\ell(-a) = \ell(a)$.
P3. ℓ is nondecreasing on $\mathbb{R}_+$: $\ell(a) \leq \ell(b)$ whenever $|a| \leq |b|$.
P4. There exists $\alpha_\ell > 0$ such that
$\ell(a - b) \geq \alpha_\ell \ell(a) - \ell(b) \ \forall (a, b) \in \mathbb{R}^2$.

Property P4 can be interpreted as a relaxed version of the triangle inequality property for ℓ. We can characterize a family of functions ℓ satisfying P1-P4 as follows.

Lemma 3 (Examples of functions obeying P1-P4)
For any real number $p \geq 1$, the function $\ell_p : \mathbb{R} \to \mathbb{R}_+$ defined by $\ell_p(x) = |x|^p$ satisfies the properties P1-P4. In particular, P4 is satisfied with $\alpha_\ell = 1/2^{p-1}$.

PROOF. That ℓ_p satisfies P1-P3 is an obvious fact. As to Property P4, it follows from convexity. In effect the convexity of ℓ_p implies that for all $(a, b) \in \mathbb{R}^2$, $|a + b|^p / 2^p = \ell_p((a + b)/2) \leq 1/2\ell_p(a) + 1/2\ell_p(b)$. Multiplying by 2 gives $1/2^{p-1}\ell_p(a + b) \leq \ell_p(a) + \ell_p(b)$, which by replacing a with $a - b$ can be seen to be equivalent to P4 with $\alpha_\ell = 1/2^{p-1}$. $\square$

The correntropy maximization is an estimation framework where one tries to maximize the correntropy. In the regression problem stated above, we aim to find the parameter vector θ that maximizes [3] $V_{\phi_\ell}(y_t, \hat{y}_t(\theta))$, the correntropy between y_t and $\hat{y}_t(\theta)$ with respect to the kernel ϕ_ℓ. In practice however the distribution [4] p_{x_t, y_t} is generally unknown so that one cannot evaluate the exact correntropy. As a consequence of this difficulty one would be content in practice with maximizing a sample estimate of the correntropy. Assume that we are given a set $Z^N = \{(x_t, y_t)\}_{t=1}^N$ of data points sampled independently from the joint distribution p_{x_t, y_t}. Then in virtue of Assumption 1, an estimate of the correntropy is given by

$$\begin{aligned} \widehat{V}_{\phi_\ell}(y_t, \hat{y}_t(\theta)) &= \frac{1}{N} \sum_{k=1}^N \phi_\ell(y_k, \hat{y}_k(\theta)) \\ &= \frac{1}{N} \sum_{k=1}^N \exp\left[-\gamma \ell(y_k - x_k^\top \theta)\right] \end{aligned} \qquad (4)$$

for all $t \in \mathbb{I} \triangleq \{1, \ldots, N\}$. Hence the maximum correntropy estimator (MCE) studied in this paper is the possibly set-valued map $\Psi_{\text{mce}} : (\mathbb{R}^n \times \mathbb{R})^N \to \mathbb{R}^n$ which maps the data to a parameter space,

$$\Psi_{\text{mce}}(Z^N) = \arg\max_{\theta \in \mathbb{R}^n} \widehat{V}_{\phi_\ell}(y_t, \hat{y}_t(\theta)). \qquad (5)$$

In the form (4)-(5) the MCE can be viewed as a particular instance of the prediction error estimation scheme [13, Chap. 7] with prediction error measured by $\exp\left[-\gamma \ell(y_k - x_k^\top \theta)\right]$. Also, the performance index (4) is reminiscent of the risk-sensitive estimation cost which is used in control, adaptive filtering and parameter estimation [3,8]. But this latter approach, which roughly consists in the minimization of a sum of exponential of positive error terms, is not suitable for handling the effects of impulsive noise such as outliers.

Although the focus of this paper is the analysis of the properties of the estimator (5), let us mention in passing that the underlying optimization problem in (5) is non convex. This implies that solving (5) numerically can be

---

[3] By Assumption 1, $V_{\phi_\ell}(y_t, \hat{y}_t(\theta))$ is indeed constant i.e., independent of t. Hence t refers here to an arbitrary time index.

[4] To be precise, the interest is in $p_{y_t, \hat{y}_t(\theta)}$ but this follows from p_{x_t, y_t}.

challenging. However it can be interpreted iteratively as a weighted least squares problem in the case for example where ϕ_ℓ is taken to be the Gaussian kernel. We will get back to this in Section 4.

3 Robustness properties of the MCE

As discussed in the introduction, an estimator of the form (5) is intuitively thought (and empirically shown) to be endowed with some robustness properties. By this, we mean that it is able to keep behaving reasonably well when a certain fraction of the available data points are affected by noise components v_t of possibly arbitrarily large magnitude. The question of main interest in this paper is to characterize quantitatively up to what extent the estimator defined in (5) can be insensitive to outliers.

3.1 Data informativity

As will be seen, the robustness property is inherited from both the structure of the estimator and the richness of the regression data. We are therefore interested in formalizing as well that richness and how it contributes to the robustness properties of the estimator.

To proceed with the analysis, let us introduce some notations. For convenience we make the following assumption.

Assumption 4 *The regressor sequence $\{x_t\}$ satisfies:* $x_t \neq 0$ *for all $t \in \mathbb{I}$.*

Note that Assumption 4 is without loss of generality. Under this assumption, let us pose

$$r_x = \min_{t \in \mathbb{I}} \|x_t\|_2 > 0, \qquad (6)$$

with $\|\cdot\|_2$ denoting Euclidean norm. Upon dividing the system equation (1) by $\|x_t\|_2$, we can even assume that $r_x = 1$. Let $\alpha \in [0, 1]$ be a real number. For any $\eta \in \mathbb{R}^n$, define the index set

$$\mathscr{I}_\alpha(X, \eta) = \left\{ t \in \mathbb{I} : |x_t^\top \eta| \geq \alpha \|x_t\|_2 \|\eta\|_2 \right\} \qquad (7)$$

with $X = [x_1 \ \cdots \ x_N] \in \mathbb{R}^{n \times N}$ a matrix formed with all the regressors. Finally, let $\rho_\alpha(X)$ be the ratio between the minimum cardinality that $\mathscr{I}_\alpha(X, \eta)$ can attain over all possible values of η, and the number N of columns in X, i.e.,

$$\rho_\alpha(X) = \inf_{\eta \in \mathbb{R}^n} \frac{1}{N} |\mathscr{I}_\alpha(X, \eta)|. \qquad (8)$$

The number $\rho_\alpha(X)$ measures somehow the richness (or informativity/genericity) of the regression data. Intuitively, $\rho_\alpha(X)$ reflects a dense spanning of all directions of the vector space $\mathbb{R}^n$ by the vectors $\{x_t\}$. For a given $\alpha > 0$, it is desired that $\rho_\alpha(X)$ be as large as possible.

We will refer to it as the correlation measure of the matrix X at the level α.

It appears intuitively that $\rho_\alpha(X)$ is a decreasing function of α. Clearly, we get $\rho_\alpha(X) = 0$ for $\alpha = 1$ for finite N while $\rho_\alpha(X) = 1$ for $\alpha = 0$. For a given matrix, it would be interesting to be able to evaluate numerically the quantitative measure $\rho_\alpha(X)$ of richness. Indeed, this value will be required for numerical assessment of the error bound to be derived in Section 3.2. However computing exactly the value of $\rho_\alpha(X)$ is a hard combinatorial problem.

We therefore discuss how to reach estimates of $\rho_\alpha(X)$ at an affordable cost. To this end, let $\tilde{X} = [\tilde{x}_1 \ \cdots \ \tilde{x}_N] \in \mathbb{R}^{n \times N}$ be the matrix obtained from X by normalizing its columns to unit 2-norm, i.e., $\tilde{x}_t = x_t / \|x_t\|_2$ for all t. Then introduce the number

$$\sigma(X) = \min_{\eta \in \mathbb{R}^n} \left\{ \left\| \tilde{X}^\top \eta \right\|_\infty \text{ s.t. } \|\eta\|_2 = 1 \right\} \qquad (9)$$

which is solely a function of the matrix X, hence the notation. Note that the so-defined $\sigma(X)$ lies necessarily in the real interval $[0, 1]$. Moreover, it can be usefully observed that $\sigma(X) \geq \lambda_{\min}^{1/2}(\tilde{X}\tilde{X}^\top)/\sqrt{N}$, with $\lambda_{\min}^{1/2}(\cdot)$ referring to the square root of the minimum eigenvalue. Now for any $t \in \mathbb{I}$ consider the following index set

$$J_{t,\alpha}(X) = \left\{ k \in \mathbb{I} : |\tilde{x}_k^\top \tilde{x}_t| \geq \sqrt{1 - \delta^2} \right\} \qquad (10)$$

where $\delta = \sqrt{1 - \alpha^2} - \sqrt{1 - \sigma(X)^2}$. It is assumed in the definition (10) that $\sigma(X) \geq \alpha$ so that $\delta \geq 0$. For a given t, $J_{t,\alpha}(X)$ collects the indices of the regressors which are the most correlated to $\tilde{x}_t$ in the sense that the cosine of the angle they form with $\tilde{x}_t$ is larger than $\sqrt{1 - \delta^2}$. Finally, let

$$v_\alpha(X) = \frac{1}{N} \inf_{t \in \mathbb{I}} |J_{t,\alpha}(X)| \qquad (11)$$

be the ratio between the minimum cardinality of the finite set $J_{t,\alpha}(X)$ over all t living in $\mathbb{I}$ and the number N of columns in X. Then we can estimate $\rho_\alpha(X)$ as follows.

Proposition 5 *Let $X \in \mathbb{R}^{n \times N}$ be a real matrix. Then, for all $\alpha \in {]0, 1]}$ with $\alpha \leq \sigma(X)$,*

$$v_\alpha(X) \leq \rho_\alpha(X) \leq \min\left(1, \frac{\lambda_{\min}(\tilde{X}\tilde{X}^\top)}{N\alpha^2}\right) \qquad (12)$$

with $\lambda_{\min}(\tilde{X}\tilde{X}^\top)$ denoting the minimum eigenvalue of the matrix $\tilde{X}\tilde{X}^\top$.

The proof of this proposition uses the following lemma.

Lemma 6 *[20, Thm 5.14] Let $x, y, z \in \mathbb{C}^n$ be such that*

4

$x^*x = y^*y = z^*z = 1$ *with* x^* *denoting the conjugate transpose of* x. *Then*

$$\sqrt{1 - |x^*y|^2} \leq \sqrt{1 - |x^*z|^2} + \sqrt{1 - |z^*y|^2}.$$

Equality holds if and only if there exists $\beta \in \mathbb{C}$ *such that either* $z = \beta x$ *or* $z = \beta y$.

PROOF. [Proof of Proposition 5] The upper bound is immediate. To see this, let $\eta_0 \neq 0$ be the eigenvector associated with the smallest eigenvalue of $\tilde{X}\tilde{X}^\top$. Then

$$
\begin{aligned}
\lambda_{\min}(\tilde{X}\tilde{X}^\top) \|\eta_0\|_2^2 &= \sum_{t=1}^{N} (\tilde{x}_t^\top \eta_0)^2 \\
&\geq \sum_{t \in \mathscr{I}_\alpha(X,\eta_0)} (\tilde{x}_t^\top \eta_0)^2 \\
&\geq |\mathscr{I}_\alpha(X,\eta_0)| \, \alpha^2 \, \|\eta_0\|_2^2
\end{aligned}
$$

It follows that $\rho_\alpha(X) = 1/N \inf_{\eta \in \mathbb{R}^n} |\mathscr{I}_\alpha(X,\eta)| \leq 1/N |\mathscr{I}_\alpha(X,\eta_0)| \leq \lambda_{\min}(\tilde{X}\tilde{X}^\top)/(N\alpha^2)$. The upper inequality in (12) follows by additionally taking into consideration the obvious fact that $\rho_\alpha(X) \leq 1$.

We now prove the inequality $v_\alpha(X) \leq \rho_\alpha(X)$. To begin with, note from (9) that for any $\eta \in \mathbb{R}^n$ satisfying $\|\eta\|_2 = 1$, there exists $t(\eta) \in \mathbb{I}$ such that $|\eta^\top \tilde{x}_{t(\eta)}| \geq \sigma(X)$. Consider an index $k \in \mathbb{I}$, such that $|\tilde{x}_k^\top \tilde{x}_{t(\eta)}| \geq h$ for some $h \in [0,1]$. Then observe that $|\tilde{x}_k^\top \eta| \geq \alpha$ is equivalent to

$$\sqrt{1 - (\tilde{x}_k^\top \eta)^2} \leq \sqrt{1 - \alpha^2}.$$

On the other hand, by applying Lemma 6, we can write

$$
\begin{aligned}
\sqrt{1 - (\tilde{x}_k^\top \eta)^2} &\leq \sqrt{1 - (\eta^\top \tilde{x}_{t(\eta)})^2} + \sqrt{1 - (\tilde{x}_k^\top \tilde{x}_{t(\eta)})^2} \\
&\leq \sqrt{1 - \sigma(X)^2} + \sqrt{1 - h^2}.
\end{aligned}
$$

It follows that for $|\tilde{x}_k^\top \eta| \geq \alpha$ to hold, it is sufficient that

$$\sqrt{1 - \sigma(X)^2} + \sqrt{1 - h^2} \leq \sqrt{1 - \alpha^2},$$

which in turn is equivalent to $h \geq \sqrt{1 - \delta^2}$ with $\delta = \sqrt{1 - \alpha^2} - \sqrt{1 - \sigma(X)^2} \in [0,1]$ by the assumption that $\sigma(X) \geq \alpha$. Hence, for $|\tilde{x}_k^\top \eta|$ to be greater than or equal to α, it is enough that $|\tilde{x}_k^\top \tilde{x}_{t(\eta)}| \geq \sqrt{1 - \delta^2}$. This means that for a given t, k being in the index set $J_{t(\eta),\alpha}(X)$ defined in (10) is a sufficient condition for $|\tilde{x}_k^\top \eta| \geq \alpha$ for all $\eta \in \mathbb{R}^n$ such that $\|\eta\|_2 = 1$. Therefore $J_{t(\eta),\alpha}(X) \subset \mathscr{I}_\alpha(X,\eta)$ hence implying that $|J_{t(\eta),\alpha}(X)| \leq |\mathscr{I}_\alpha(X,\eta)|$. Taking now the infimum produces $v_\alpha(X) \leq \inf_{\eta \in \mathbb{R}^n} \frac{1}{N} |J_{t(\eta),\alpha}(X)| \leq$

$$\inf_{\eta \in \mathbb{R}^n} \frac{1}{N} |\mathscr{I}_\alpha(X,\eta)| = \rho_\alpha(X). \quad \square$$

The key benefit of Proposition 5 is that it provides a method for estimating the measure $\rho_\alpha(X)$ defined in (8) at an affordable cost. Note however that while the upper bound in (12) can be computed easily, obtaining the lower bound $v_\alpha(X)$ is still challenging. The reason is that this bound involves the number $\sigma(X)$ in (9) whose numerical evaluation requires solving a nonconvex optimization problem. Nevertheless, it can be approximated through some heuristics, e.g. by solving a sequence of linear programs.

Remark 7 *In comparison to the classical concept of persistence of excitation (PE) in system identification, the richness property requiring that* $\rho_\alpha(X)$ *be large is a stronger property. In finite time, the quantitative persistence of excitation (called specifically sufficiency of excitation in this case) asks for the condition number* $\lambda_{\max}^{1/2}(XX^\top)/\lambda_{\min}^{1/2}(XX^\top)$ *of* $XX^\top$ *to be as close to 1 as possible. The PE condition appears to be a global property of the matrix* X *while the richness condition introduced here is a somewhat local property as it is basically counting the number of vectors* x_t *pointing in any direction of the regression space.*

3.2 Main results

Equipped with the measure of informativity introduced above, we can now state the main result of this paper, which stands as follows.

Theorem 8 *Let* $I_\varepsilon^0 = \{t \in \mathbb{I} : |v_t| \leq \varepsilon\}$ *and* $I_\varepsilon^c = \{t \in \mathbb{I} : |v_t| > \varepsilon\}$ *with* $\{v_t\}$ *denoting the noise sequence in (1). Let* ℓ *be a function obeying P1-P4. Assume that the following condition is satisfied for some* $\alpha \in]0,1]$,

$$\frac{1}{1 + e^{-\gamma\ell(\varepsilon)}} \rho_\alpha(X) + e^{-\gamma\ell(\varepsilon)} \frac{|I_\varepsilon^0|}{N} > 1. \quad (13)$$

Then for any $\theta^\star \in \Psi_{\mathrm{mce}}(Z^N)$ *with* Z^N *being generated by system (1), it holds that*

$$\ell(\alpha r_x \|\theta^\star - \theta^o\|_2) \leq \frac{1}{\gamma\alpha_\ell} \ln(1/\mu_\ell), \quad (14)$$

where

$$\mu_\ell = \frac{1 + e^{-\gamma\ell(\varepsilon)}}{\frac{|I_\varepsilon^0|}{N} + \rho_\alpha(X) - 1} \left[\frac{1}{1 + e^{-\gamma\ell(\varepsilon)}} \rho_\alpha(X) + e^{-\gamma\ell(\varepsilon)} \frac{|I_\varepsilon^0|}{N} - 1 \right] \quad (15)$$

If in addition, ℓ *is strictly increasing on* $\mathbb{R}_+$, *then*

$$\|\theta^\star - \theta^o\|_2 \leq \frac{1}{\alpha r_x} \ell^{-1} \left(\frac{1}{\gamma\alpha_\ell} \ln(1/\mu_\ell) \right). \quad (16)$$

PROOF. Let

$$\theta^\star \in \Psi_{\mathrm{mce}}\left(Z^N\right) = \arg\max_{\theta \in \mathbb{R}^n} \widehat{V}_{\phi_\ell}(y_t, \hat{y}_t(\theta)).$$

Then for any $\theta \in \mathbb{R}^n$, it holds that

$$\sum_{t=1}^N \exp\left[-\gamma\ell(y_t - x_t^\top \theta)\right] \leq \sum_{t=1}^N \exp\left[-\gamma\ell(y_t - x_t^\top \theta^\star)\right]$$

Taking in particular $\theta = \theta^o$ and invoking the system equation (1), it follows that

$$\sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\ell(v_t)\right] + \sum_{t \in I_\varepsilon^c} \exp\left[-\gamma\ell(v_t)\right]$$
$$\leq \sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\ell(v_t - x_t^\top \eta^\star)\right]$$
$$+ \sum_{t \in I_\varepsilon^c} \exp\left[-\gamma\ell(v_t - x_t^\top \eta^\star)\right]$$

where we have posed $\eta^\star = \theta^\star - \theta^o$. This implies that

$$\sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\ell(v_t)\right] \leq \sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\ell(v_t - x_t^\top \eta^\star)\right] + |I_\varepsilon^c|$$

With $|v_t| \leq \varepsilon$ for any $t \in I_\varepsilon^0$, we have $-\gamma\ell(v_t) \geq -\gamma\ell(\varepsilon)$ by the symmetry and nondecreasing properties of ℓ. As a consequence, $\sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\ell(v_t)\right] \geq \sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\ell(\varepsilon)\right]$. On the other hand, by the fourth property of the function ℓ, $\ell(v_t - x_t^\top \eta^\star) \geq \alpha_\ell \ell(x_t^\top \eta^\star) - \ell(v_t) \geq \alpha_\ell \ell(x_t^\top \eta^\star) - \ell(\varepsilon)$ hence implying that $-\gamma\ell(v_t - x_t^\top \eta^\star) \leq -\gamma\alpha_\ell \ell(x_t^\top \eta^\star) + \gamma\ell(\varepsilon)$. Combining these observations allows us to write

$$\sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\ell(\varepsilon)\right] - |I_\varepsilon^c|$$
$$\leq \exp\left[\gamma\ell(\varepsilon)\right] \sum_{t \in I_\varepsilon^0} \exp\left[-\gamma\alpha_\ell \ell(x_t^\top \eta^\star)\right]$$
$$= \exp\left[\gamma\ell(\varepsilon)\right] \sum_{t \in I_\varepsilon^0 \cap \mathscr{I}_\alpha(X,\eta^\star)} \exp\left[-\gamma\alpha_\ell \ell(x_t^\top \eta^\star)\right]$$
$$+ \exp\left[\gamma\ell(\varepsilon)\right] \sum_{t \in I_\varepsilon^0 \cap \mathscr{I}_\alpha^c(X,\eta^\star)} \exp\left[-\gamma\alpha_\ell \ell(x_t^\top \eta^\star)\right]$$

In the last equality we have partitioned the set I_ε^0 into $I_\varepsilon^0 \cap \mathscr{I}_\alpha(X,\eta^\star)$ and $I_\varepsilon^0 \cap \mathscr{I}_\alpha^c(X,\eta^\star)$ with $\mathscr{I}_\alpha^c(X,\eta^\star)$ being the complement of $\mathscr{I}_\alpha(X,\eta^\star)$ in $\mathbb{I}$. Note from (7) that for all $t \in I_\varepsilon^0 \cap \mathscr{I}_\alpha(X,\eta^\star)$, $\ell(x_t^\top \eta^\star) \geq \ell(\alpha r_x \|\eta^\star\|_2)$ so that $-\gamma\alpha_\ell \ell(x_t^\top \eta^\star) \leq -\gamma\alpha_\ell \ell(\alpha r_x \|\eta^\star\|_2)$. Plugging these

observations into the above inequality yields

$$\exp\left[-\gamma\ell(\varepsilon)\right]\left\{|I_\varepsilon^0|\exp\left[-\gamma\ell(\varepsilon)\right] - |I_\varepsilon^c|\right\}$$
$$\leq |I_\varepsilon^0 \cap \mathscr{I}_\alpha(X,\eta^\star)|\exp\left[-\gamma\alpha_\ell \ell(\alpha r_x \|\eta^\star\|_2)\right]$$
$$+ |I_\varepsilon^0 \cap \mathscr{I}_\alpha^c(X,\eta^\star)|.$$

By observing that $|I_\varepsilon^0 \cap \mathscr{I}_\alpha^c(X,\eta^\star)| = |I_\varepsilon^0| - |I_\varepsilon^0 \cap \mathscr{I}_\alpha^0(X,\eta^\star)|$, we can rearrange the above inequality in the form

$$e^{-\gamma\ell(\varepsilon)}\left(|I_\varepsilon^0| e^{-\gamma\ell(\varepsilon)} + |I_\varepsilon^0| - N\right) - |I_\varepsilon^0|$$
$$\leq |I_\varepsilon^0 \cap \mathscr{I}_\alpha(X,\eta^\star)|\left\{\exp\left[-\gamma\alpha_\ell \ell(\alpha r_x \|\eta^\star\|_2)\right] - 1\right\}$$
(17)

Now by exploiting the definition of $\rho_\alpha(X)$, we can observe that

$$|I_\varepsilon^0 \cap \mathscr{I}_\alpha(X,\eta^\star)| = |I_\varepsilon^0| + |\mathscr{I}_\alpha(X,\eta^\star)| - |I_\varepsilon^0 \cup \mathscr{I}_\alpha(X,\eta^\star)|$$
$$\geq |I_\varepsilon^0| + N\rho_\alpha(X) - N.$$

Moreover since

$$N\rho_\alpha(X) + |I_\varepsilon^0| \geq \frac{1}{1 + e^{-\gamma\ell(\varepsilon)}}N\rho_\alpha(X) + e^{-\gamma\ell(\varepsilon)}|I_\varepsilon^0|,$$

the assumption (13) guarantees that $|I_\varepsilon^0| + N\rho_\alpha(X) - N > 0$. Therefore since the term on the right hand side of (17) is negative, it holds that

$$e^{-\gamma\ell(\varepsilon)}\left\{|I_\varepsilon^0| e^{-\gamma\ell(\varepsilon)} + |I_\varepsilon^0| - N\right\} - |I_\varepsilon^0|$$
$$\leq \left(|I_\varepsilon^0| + N\rho_\alpha(X) - N\right)\left\{\exp\left[-\gamma\alpha_\ell \ell(\alpha r_x \|\eta^\star\|_2)\right] - 1\right\}$$

Then direct algebraic calculations lead to

$$\mu_\ell \leq \exp\left[-\gamma\alpha_\ell \ell(\alpha r_x \|\eta^\star\|_2)\right] \leq 1$$

where μ_ℓ is defined as in (15). Indeed, in virtue of the assumption (13), μ_ℓ is positive. Hence we have

$$\ell(\alpha r_x \|\eta^\star\|_2) \leq \frac{1}{\gamma\alpha_\ell} \ln(1/\mu_\ell).$$

Of course, if ℓ is monotonically increasing on $\mathbb{R}_+$ then it is invertible and the error bound in (16) follows. $\square$

A few comments follow from this result. A key assumption of the theorem is condition (13). What it requires is on the one hand, that the proportion of outliers be somehow small and on the other hand, that the regression data X be rich in the sense that $\rho_\alpha(X)$ be large enough for a given nonzero $\alpha \in \,]0,1]$. An important teaching of this condition is that the richer the data matrix X, the larger the number of outliers that can be corrected by the estimator. We can interpret (13) as a sufficient con-

dition for the stability of the estimator since it guarantees a bounded estimation error.

A second comment concerns the amplitude of the error bound given in (16). For the purpose of making this bound small, we need the constant μ_ℓ to be close to one. Again we see that this is favored by a small number of outliers and a rich data set. An interesting special case is when $\ell(\varepsilon) = 0$, which occurs when the data are only affected by some outliers ($\varepsilon = 0$) and no dense noise. In this case the number μ_ℓ defined in (15) reduces to

$$\mu_\ell = 1 - \frac{1 - \frac{|I_\varepsilon^0|}{N}}{\frac{|I_\varepsilon^0|}{N} + \rho_\alpha(X) - 1}$$

which tend to suggest, since $\mu_\ell \neq 1$, that no exact recovery might be achieved once the data are affected by a single outlier unless we consider in (16) the limit case when $\gamma \to +\infty$. A similar observation was made in [6] in a comparable context. We will prove below that the MCE does not possess the exact recovery property, at least in the case when $\ell(a) = a^2$. In contrast, a robust estimator such as the LAD estimator (see, e.g., [1,2]) is able to achieve exact recovery under a relatively significant proportion of nonzero errors.

Proposition 9 *Let Assumption 4 hold and assume that for any $t \in I_\varepsilon^0$, $v_t = 0$. Take the function ℓ in (4) to be the square function, $\ell(a) = a^2$. For all $\varepsilon > 0$, if $|I_\varepsilon^c| \geq 1$ then there exists a sequence $\{v_t\}$ such that $\theta^o \notin \Psi_{\mathrm{mce}}(Z^N)$ when Z^N is generated from (1) under the action of $\{v_t\}$.*

PROOF. We start by observing that with ℓ being the square function, the cost $\hat{V}_{\phi_\ell}(y_t, \hat{y}_t(\theta))$ is differentiable. Therefore, a necessary condition for θ to be in $\Psi_{\mathrm{mce}}(Z^N)$ is that $\nabla \hat{V}_{\phi_\ell}(y_t, \hat{y}_t(\theta)) = 0$, where ∇ refers to the gradient. This, by using the system equation (1) and exploiting the assumption that $v_t = 0$ for $t \in I_\varepsilon^0$, can be translated into

$$\left(\sum_{t=1}^N w_t(\theta) x_t x_t^\top \right) (\theta - \theta^o) = \sum_{t \in I_\varepsilon^c} w_t(\theta) v_t x_t,$$

where $w_t(\theta) = \exp\left(-\gamma\ell(v_t - x_t^\top(\theta - \theta^o))\right)$. Note that the matrix on the left hand side of the equation above is finite regardless of the value of θ. Hence, in the event that $\theta^o \in \Psi_{\mathrm{mce}}(Z^N)$, we would have

$$\sum_{t \in I_\varepsilon^c} w_t(\theta^o) v_t x_t = \sum_{t \in I_\varepsilon^c} \lambda_t x_t = 0$$

with $\lambda_t = \exp(-\gamma\ell(v_t)) v_t$. Clearly, it is possible to find a nonzero sequence $\{v_t\}$ which does not meet this condition. Hence θ^o cannot be in $\Psi_{\mathrm{mce}}(Z^N)$ for an arbitrary $\{v_t : t \in I_\varepsilon^c\}$ no matter how small the cardinality of I_ε^c is. $\square$

We now discuss some special instances of Theorem 8 corresponding to two kernels which are frequently used for estimation. For convenience of the discussion, let us introduce the following notation. Let $\mu : \mathbb{R}_+^3 \to \mathbb{R}_+$

$$\mu(z, \varepsilon, \alpha) = \frac{1 + e^{-z}}{\frac{|I_\varepsilon^0|}{N} + \rho_\alpha(X) - 1} \left[\frac{1}{1 + e^{-z}} \rho_\alpha(X) + e^{-z} \frac{|I_\varepsilon^0|}{N} - 1 \right]$$

whenever $e^{-z} \frac{|I_\varepsilon^0|}{N} + \frac{1}{1+e^{-z}} \rho_\alpha(X) - 1 > 0$ and $\mu(z, \varepsilon, \alpha) = 0$ otherwise.

Remark 10 *The bound* (16) *allows for some degree of freedom in the choice of the parameter α. For a given function ℓ assumed to be invertible on $\mathbb{R}_+$, a better bound can, in principle, be obtained as*

$$\min_{\alpha \in]0,1]} \frac{1}{\alpha r_x} \ell^{-1}\left(\frac{1}{\gamma\alpha_\ell} \ln\left(\frac{1}{\mu(\gamma\ell(\varepsilon), \varepsilon, \alpha)} \right) \right)$$

subject to $\alpha \leq \sigma(X)$ and condition (13). *Although such a minimum might not be easy to compute exactly, one can make the error bound a little tighter by performing for example some grid search. In the same manner one can envision optimizing the parameter γ of the estimator.*

3.3 Laplacian kernel

The Maximum Laplacian Correntropy estimator (MCE-L) corresponds to the case where the function ℓ in (3) is taken to be such that $\ell(a) = |a|$. As a result, the function ϕ_ℓ takes the form

$$\phi_1(y, \hat{y}) = \exp\left(-\gamma_1 |y - \hat{y}|\right). \tag{18}$$

It is straightforward to see that the properties P1-P4 are satisfied by ℓ with $\alpha_\ell = 1$. Theorem 8 can be specialized to this case as follows.

Corollary 11 (Laplacian kernel)
Let I_ε^0 be defined as in Theorem 8. Assume that the following condition is satisfied

$$\frac{1}{1 + e^{-\gamma_1 \varepsilon}} \rho_\alpha(X) + e^{-\gamma_1 \varepsilon} \frac{|I_\varepsilon^0|}{N} > 1 \tag{19}$$

for some $\alpha \in]0,1]$.
Then for any $\theta^\star \in \arg\max_{\theta \in \mathbb{R}^n} \widehat{V}_{\phi_1}(y_t, \hat{y}_t(\theta))$ with ϕ_1 defined as in (18), *it holds that*

$$\|\theta^\star - \theta^o\|_2 \leq \frac{1}{\gamma_1 \alpha r_x} \ln\left(\frac{1}{\mu(\gamma_1 \varepsilon, \varepsilon, \alpha)} \right) \tag{20}$$

3.4 Gaussian kernel

The most used form of correntropy is the one based on the Gaussian kernel which, by omitting the normalizing

factor, can be written in the form

$$\phi_2(y, \hat{y}) = \exp(-\gamma_2(y - \hat{y})^2) \quad (21)$$

with $\gamma_2 > 0$. We will refer to the associated estimator as the maximum Gaussian correntropy estimator (MCE-G). Here, the function ℓ is defined by $\ell(a) = a^2$ and according to Lemma 3, it satisfies the properties P1-P4. In particular, P4 is satisfied with $\alpha_\ell = 1/2$. Moreover ℓ is clearly monotonic on $\mathbb{R}_+$. As a consequence, we get a corollary of Theorem 8 as follows.

Corollary 12 (Gaussian kernel) *Let I_ε^0 be defined as in Theorem 8. Assume that the following condition is satisfied*

$$\frac{1}{1 + e^{-\gamma_2\varepsilon^2}}\rho_\alpha(X) + e^{-\gamma_2\varepsilon^2}\frac{|I_\varepsilon^0|}{N} > 1 \quad (22)$$

for some $\alpha \in \,]0, 1]$.
Then for any $\theta^\star \in \arg\max_{\theta \in \mathbb{R}^n} \widehat{V}_{\phi_2}(y_t, \hat{y}_t(\theta))$, it holds that

$$\|\theta^\star - \theta^o\|_2 \leq \frac{1}{\alpha r_x}\left[\frac{2}{\gamma_2}\ln\left(\frac{1}{\mu(\gamma_2\varepsilon^2, \varepsilon, \alpha)}\right)\right]^{1/2}. \quad (23)$$

3.5 A remark on the error-in-variables scenario

We now consider the situation where only a noisy observation $\bar{x}_t = x_t + w_t$ of the regressor vector x_t in (1) is available for prediction. This scenario is referred to as the robust error-in-variable (EIV) regression problem. Then the predictor output is given by

$$\hat{y}_t(\theta) = \bar{x}_t^\top \theta.$$

Indeed Theorem 8 remains valid for this case. To see this note that the system equation (1) can be rewritten as

$$y_t = \bar{x}_t^\top \theta^o + \bar{v}_t$$

where $\bar{v}_t = v_t - w_t^\top \theta^o$. Then clearly the theorem applies to the EIV scenario with $\{x_t\}$ and $\{v_t\}$, replaced respectively by $\{\bar{x}_t\}$ and $\{\bar{v}_t\}$. One limitation however in this case is that for a given $\varepsilon \geq 0$, the cardinality of the set $I_\varepsilon^0 = \{t \in \mathbb{I} : |\bar{v}_t| \leq \varepsilon\}$ is likely to be much smaller than in the situation where the regressors are noise-free.

4 Numerical experiments

The purpose of this section is to provide a numerical illustration of the richness measure (8) and of the estimation error bound (16). The system example considered for the experiment is of an FIR-type and is given by

$$y_t = 0.5u_t - u_{t-1} + 0.2u_{t-2} + v_t \quad (24)$$

which can be written in the form (1) with $\theta^o = [0.5 \ {-}1 \ 0.2]^\top$ and $x_t = [u_t \ u_{t-1} \ u_{t-2}]^\top$. For the data-generation experiment, assume that $\{u_t\} \overset{iid}{\sim} \mathcal{N}(0, 1)$ i.e., $\{u_t\}$ is sampled independently and identically from a zero-mean Gaussian distribution of unit variance. As for the noise signal $\{v_t\}$, it is defined as $v_t = e_t + f_t$ with $\{e_t\} \overset{iid}{\sim} \mathcal{U}([-\varepsilon, \varepsilon])$ where $\mathcal{U}$ refers to the uniform distribution and $\{f_t\}$ is a sequence of sparse noise with only a few nonzero elements (which are otherwise not constrained in magnitude); the nonzero elements of $\{f_t\}$ are here sampled from $\mathcal{N}(50, 10)$.

4.1 Illustration of estimates

We generate $N = 300$ data pairs (x_t, y_t) and carry out a comparison between three estimators: on the one hand, the maximum Laplacian correntropy estimator (MCE-L) and the maximum Gaussian correntropy estimator (MCE-G) and on the other hand, the Least Absolute Deviation (LAD) estimator (which is also called ℓ_1 estimator). Recall that MCE-G and MCE-L involve non convex optimization. Here they are heuristically implemented as a reweighted iterative least squares estimator and as a reweighted ℓ_1 estimator respectively. The results are represented in Figure 1 in term of average estimation error. What this suggests is that for fixed values of the design parameters γ_1 and γ_2 (see Eqs (18) and (21) for the roles of these parameters), LAD and MCE-L enjoy a similar performance for small amount of noise. But as the noise level increases, LAD shows better stability capabilities than the MCE-L. Note that overall MCE-G tends to perform best in the setting of this experiment as long as the magnitude of the dense noise is reasonable (SNR larger than 2.8 dB). A possible justification for this is that squaring errors that contain outliers as in (21) cancel out their influence more forcefully than just taking their absolute value as in (18).

4.2 Estimation of richness measure $\rho_\alpha(X)$

We provide a graphical representation of how the informativity measure $\rho_\alpha(X)$ may, for a given data matrix $X \in \mathbb{R}^{n \times N}$, evolve with respect to the dimensions N/n of X and the demanded degree α of richness (See Figure 2). The estimated range for $\rho_\alpha(X)$ is based on Eq. (12). Here X is formed from an FIR-type of regressors with an input sampled from a zero-mean and unit variance Gaussian distribution. Our experiments in this specific study tend to suggest that $\rho_\alpha(X)$ is a non decreasing function of the ratio N/n and a decreasing function of α. Moreover, the estimated range (gray regions in Fig. 2) gets wider when n is large.

4.3 Estimates of error bounds

The goal here is twofold: (i) illustrate the variation of the estimation error bounds with respect to the magnitude of the dense noise in the special cases (20) and (23);
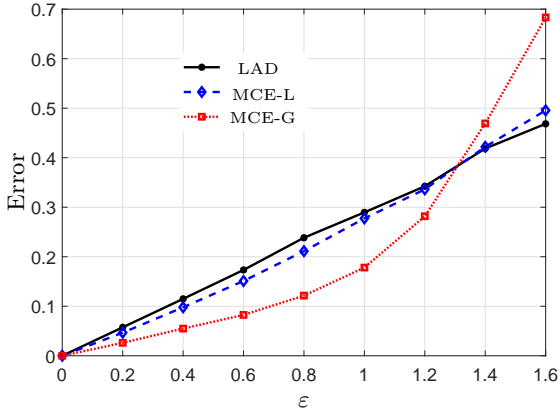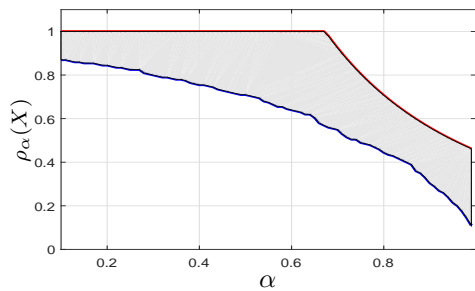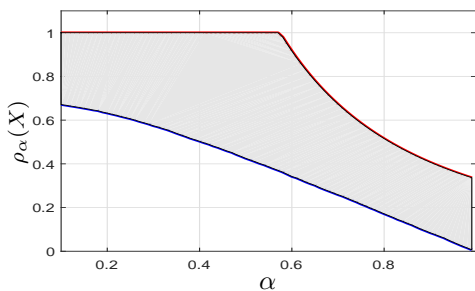
Figure 1. Comparison of the estimators MCE-L, MCE-G and LAD. Evolution of the (average) estimation error in function of the level ε of noise. The range $[0, 1.6]$ of ε corresponds indeed to a range of about $[2, +\infty]$ for the signal-to-noise ratio in this experiment. The results are obtained from a Monte-Carlo simulation of size 1000. For each experiment, the proportion of outliers is maintained fixed and equal to 50%. Design parameters: $\gamma_1 = 0.5$ and $\gamma_2 = 0.25$.

(a) $X \in \mathbb{R}^{2 \times 200}$

(b) $X \in \mathbb{R}^{3 \times 6000}$

Figure 2. Estimates of $\rho_\alpha(X)$ as a function of $\alpha \in [0, 1]$. The true values of $\rho_\alpha(X)$ lie in the region depicted in gray color.

(ii) assess how conservative the derived theoretical error bounds may be with respect to the empirical errors.

Increasing rates of the bounds. If for each level ε of noise, we select the parameter γ such that the product $\gamma\ell(\varepsilon)$ is kept constant, then the error bounds corresponding to both MCE-L and MCE-G have a linear rate of change with respect to ε as depicted in Figure 3. The

increasing rate of the bound corresponding to MCE-L is larger than that of MCE-G for the current setting. Note that the computation of bounds made here is not connected to the experiment of Section 4.1.
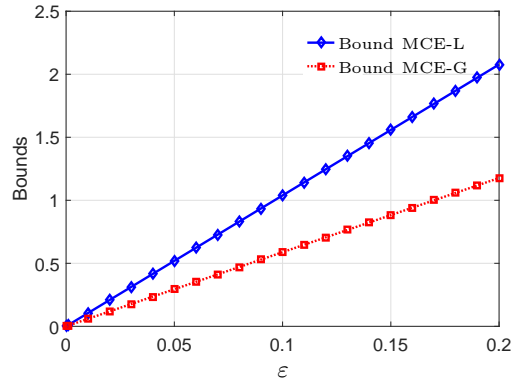

Figure 3. Illustration of increasing rates of the error bounds. Results obtained by assuming that $\gamma\ell(\varepsilon) = 0.2$, $|I_\varepsilon^0|/N = 0.8$, $\rho_\alpha(X) = 0.8$, $\alpha = 0.6$ in both cases. These values are chosen independently of a specific data matrix X but in such a way that (19) and (22) are satisfied.

Comparing theoretical bounds and empirical errors. It might be instructive to see how far away the theoretical error bounds may be from the empirical values. To study this aspect, let us consider a numerical experiment with a similar data-generating process as described in the beginning of Section 4. The dense noise level is set to $\varepsilon = 0.05$ which gives an SNR of about 25 dB and the proportion of outliers is set to 10% (which is small enough to enforce condition (13)). One difficulty in evaluating the theoretical bounds is that this requires evaluating $\rho_\alpha(X)$ which, as already discussed in Section 3.1, is a hard problem. Hence, $\rho_\alpha(X)$ is replaced here with the mean value of the lower and upper estimates displayed in (12). We then let the number N of data vary from 500 to 5000 and plot the empirical errors along with the bounds from (20) and (23) in Figure 4.

It is fair to observe that the theoretical bounds are conservative in the sense that they are generally higher than the true empirical errors. Here the ratio between the bounds and the true errors is about 30. Conservativeness is indeed a common feature for these types of results due to the various inequalities employed for the derivation. Nevertheless, the main interest of Theorem 8 is that it provides a sufficient condition for the robustness of the maximum correntropy estimator, a condition that depends explicitly on the degree of informativity of the regression data and on the proportion of outliers. Moreover, by expressing error bounds which involve explicitly the design parameters, the theorem gives insights into how to tune those parameters with the aim to improve estimation performance.

A further remark one can make is that the general formula for the error bound in (16) has a kind of universal
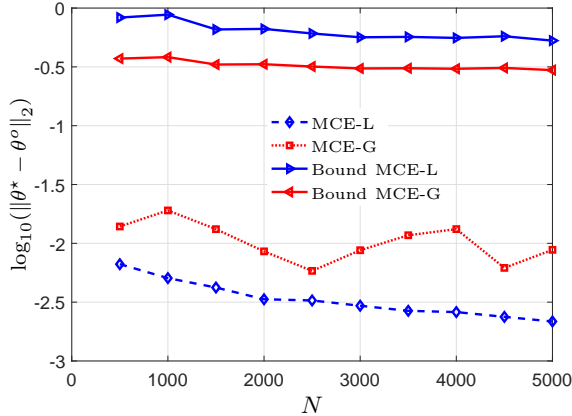
Figure 4. Comparison of theoretical bounds and empirical errors $\|\theta^\star - \theta^o\|_2$. Results obtained with an estimate of $\rho_\alpha(X)$ for $\alpha = 0.6$, $\bar\varepsilon = 0.05$ (i.e., SNR = 25 dB), $\gamma\ell(\varepsilon) = 0.2$, $|I_\varepsilon^0|/N = 0.9$. The results are obtained from a Monte-Carlo simulation of size 1000.

feature in the following sense: since the bound does not involve the magnitude of the true θ^o (for an FIR-type system for example), it is in principle valid regardless of θ^o. Hence the relative error will be as smaller as the norm of the to-be-estimated parameter vector θ^o is larger.

5 Conclusion

In this paper we have proposed an analysis of the robustness properties of a correntropy maximization framework for regression problems. The class of estimators considered is quite general and include the Gaussian and Laplacian kernels as special cases. The contribution of the work consists in (i) deriving an appropriate notion of richness for the regression data; (ii) proving stability of the considered class of estimators under the derived richness condition when the data are subject to dense and sparse noise (outliers). Our main result states that if the regression data are rich enough and if the number of outliers is small in some sense, then the parametric estimation error is bounded. The results come with explicit bounds which, in default of being exactly computable, can be estimated with computable estimates.

Acknowledgements

References

[1] L. Bako. On a class of optimization-based robust estimators. *IEEE Transactions on Automatic Control (To appear)*, 2017.

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

[3] R. K. Boel, M. R. James, and I. R. Petersen. Robustness and risk-sensitive filtering. *IEEE Transactions on Automatic Control*, 47:451–461, 2002.

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

[5] B. Chen, X. Liu, H. Zhao, N. Zheng, and J. Principe. Insights into the robustness of minimum error entropy estimation. *IEEE Transactions on Neural Networks and Learning Systems (To appear)*, 2016.

[6] B. Chen, L.Xing, H. Zhao, B. Xu, and J. C. Principe. Robustness of maximum correntropy estimation against large outliers. `https://arxiv.org/abs/1703.08065`, 2017.

[7] B. Chen, L. Xing, J. Liang, N. Zheng, and J. C. Principe. Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion. *IEEE Signal Processing Letters*, 21:880–884, 2014.

[8] S. Dey and J. B. Moore. Risk-sensitive filtering and smoothing via reference probability methods. *IEEE Transactions on Automatic Control*, 42:1587–1591, 1997.

[9] Y. Feng, X. Huang, L. Shi, Y. Yang, and J. A.K. Suykens. Learning with the maximum correntropy criterion induced losses for regression. *Journal of Machine Learning Research*, 16:993–1034, 2015.

[10] R. He, W.-S. Zheng, and B.-G. Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1561–1576, 2011.

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

[12] W Liu, P. P. Pokharel, and J. C. Principe. Correntropy: Properties and aplications in non-gaussian signal processing. *IEEE Transactions on Signal Processing*, 55(11):5286–5298, 2007.

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

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

[15] J. C. Principe, editor. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives.* New York: Springer Verlag, 2010.

[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] I. Santamaria, P. P. Pokharel, and J. C. Principe. Generalized correlation function: definition, properties, and application to blind equalization. *IEEE Transactions on Signal Processing*, 54:2187–2197, 2006.

[19] B. Schölkopf and A. J. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press, Cambridge, MA, 2002.

[20] F. Zhang. *Matrix Theory: Basic Results and Techniques.* Springer: New York, 1999.