# Entropy-Based Estimators in the Presence of Multicollinearity and Outliers

Enrico Ciavolino[1] and Giovanni Indiveri[2]

[1] University of Salento, Department of Social Science, Palazzo Parlangeli,
Via Stampacchia n.1, 73100, Lecce, Italy
enrico.ciavolino@unisalento.it

[2] University of Salento, Dipartimento Ingegneria Innovazione,
via Monteroni, 73100 Lecce, Italy
giovanni.indiveri@unisalento.it

**Abstract.** The concept and the mathematical properties of entropy play an important role in statistics, cybernetics, and information sciences. Indeed, many algorithms and statistical data processing tools, with a wide range of targets and scopes, have been designed based on entropy. The paper describes two estimators inspired by the concept of entropy that allow to robustly cope with multicollinearity, in one case, and outliers, in the other. The Generalized Maximum Entropy (GME) estimator optimizes the Shannon's entropy function subject to consistency and normality constraints. In regression applications GME allows, for example, to estimate model coefficients in the presence of multicollinearity. The Least Entropy-Like (LEL) estimator is a novel prediction error model coefficient identification algorithm that minimizes a nonlinear cost function of the fitting residuals. As the cost function that is minimized shares the same mathematical properties of entropy, it allows to compute an estimate of the model coefficients corresponding to a positively skewed distribution of the residuals. The resulting estimator exhibits higher robustness to outliers with respect to standard, as ordinary least squares (OLS) model coefficient approaches. Both the GME and LEL estimation methods are applied to a common case study to illustrate their respective properties.

## 1 Introduction

When talking to Claude Shannon about what name to use for the measure of uncertainty (or information) that he had introduced [18], John von Neumann is quoted for having suggested [22]: *You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.*

Indeed, entropy is a rather general concept: since it was first acknowledged that it could be used well beyond thermodynamics and statistical mechanics [12], the number of different areas where it has been successfully exploited has grown dramatically.

This paper describes two specific data processing algorithms inspired by the mathematical definition of Shannon's (and Gibb's) entropy. Interestingly, neither of the two

A.G.S. Ventre et al. (Eds.): *Multicriteria & Multiagent Decision Making*, STUDFUZZ 305, pp. 65–76.
DOI: 10.1007/978-3-642-35635-3_5

algorithms are strictly related to probability or classical statistical signal processing theory, but rather they are built exploiting the mathematical properties of entropy. The first of the two algorithms (LEL - Least Entropy-Like estimator) is a model coefficient estimation filter designed to yield robustness to outliers with respect to Ordinary Least Squares (OLS) or similar approaches. The need for robust parameter identification solutions is very strong in the area of control systems and signal processing [2][3][8][16]. The second algorithm is known as GME — Generalized Maximum Entropy [9] and is particularly useful when dealing with multicollinearity. This may occur, by example, in parameter estimation problems with fewer data than parameters or in the presence of rank deficient regression matrices when dealing with linear in the parameters models.

The rationale of the paper is to illustrate two examples of estimation methods designed by exploiting the properties of entropy. In particular, the estimators illustrated in this paper exhibit a noticeable robustness to multicollinearity and outliers. Giving a general overview of entropy-related methods for signal processing or even only parameter estimation goes well beyond the scope of this paper. The discussion will be limited to the LEL and GME approaches. Standard estimators, as OLS, become numerically ill conditioned as the condition number of the regression matrix grows. The OLS solution, in particular, is not defined in the presence of a regression matrix with infinite condition number. On the contrary, the GME approach is not ill conditioned for rank deficient regression matrices and it can robustly tackle the case of multicollinearity (large, but finite, regression matrix condition number). As for the LEL method, this was introduced in [11] and it consists in a model coefficient estimator based on the minimization of an entropy-like cost function. The cost function to be minimized in the LEL method is a nonlinear prediction error function exploiting the mathematical properties of entropy. In particular, this method exhibits an enhanced robustness to outliers as compared with OLS due to the very structure of the cost function to be minimized. Moreover, the proposed solution is computationally much less demanding with respect to alternative outlier robust approaches as the Least Median of Squares [17]. Indeed, thanks to these properties it has also been successfully employed in computer vision applications with stringent computational time requirements [7].

The paper is organized as follows: Sections 2 and 3 focus on briefly summarizing the LEL and GME algorithms, respectively. In Sect.4, we report validation results of the two methods as applied to a classical data set while concluding remarks are reported in Sect. 5.

## 2 Entropy-Like Estimator

Consider the model

$$y_i = f(x_{i1}, x_{i2}, \dots, x_{im}, \boldsymbol{\theta}_r) + \varepsilon_i \; : \; i = 1, 2, \dots, n. \tag{1}$$

where $\boldsymbol{\theta}_r \in \mathbb{R}^{m\times 1}$ is the unknown parameter vector, $y_i \in \mathbb{R}$ is the response variable, $x_{i1}, x_{i2}, \dots, x_{im}$ are the explanatory variables and $\varepsilon_i$ the error term. Index $i$ runs on the number of observations $n$ that is assumed to be strictly larger than $m$ (notice that this might not be the case for the GME method described in the next sections). The error term $\varepsilon_i$ is assumed to be a random variable with zero mean. Denoting with

$Z_n = \{(y_i, x_{i1}, x_{i2}, \ldots, x_{im}) : i = 1, 2, \ldots, n\}$ the set of the available observations, a regression estimator $T$ is an algorithm associating to $Z_n$ an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_r$, namely $T(Z_n) = \hat{\boldsymbol{\theta}}$. Prediction error estimators $T$ are designed based on the properties of the regression residuals

$$r_i := y_i - \hat{y}_i \tag{2}$$

being $\hat{y}_i$ the predicted responses $\hat{y}_i = f(x_{i1}, x_{i2}, \ldots, x_{im}, \hat{\boldsymbol{\theta}})$. The most popular prediction error estimators are the Least Squares (LS) and weighted LS (WLS) estimators defined respectively as

$$\hat{\boldsymbol{\theta}}_{LS} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} r_i^2 \tag{3}$$

$$\hat{\boldsymbol{\theta}}_{WLS} = \arg\min_{\boldsymbol{\theta}} \left(\mathbf{r}^T \boldsymbol{\Gamma}\, \mathbf{r}\right) \tag{4}$$

being $\mathbf{r} \in \mathbb{R}^{n\times 1}$ the residual vector $\mathbf{r} = (r_1, r_2, \ldots, r_n)^T$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times n}$ a symmetric positive definite (or eventually semidefinite) matrix of weights. Many other estimators have been proposed in the literature as, by example, M-estimators [10] or Least Median of Squares (LMS) estimators [17] that are defined through the minimization of a properly defined cost function. M-estimators, as LS or WLS estimators, share a common structure related to the additive nature of the corresponding cost function to be minimized: in particular such estimators can be all modeled as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(r_i) \tag{5}$$

for some scalar function $\rho : \mathbb{R} \longrightarrow [0, +\infty)$ of the residuals that depend from $\boldsymbol{\theta}$. For the LS estimator $\rho$ is $\rho(r_i) = r_i^2$ while for M-estimators there are many possible different choices [10]. The design of the $\rho$ function for M estimators is usually performed aiming at achieving robustness to outliers resulting in cost functions that, in general, do not admit a closed form solution to the minimization problem as for the LS case.

The LMS estimator [17], instead, is defined through a cost function that differs in nature from the structure reported in Eq. (5); namely the LMS estimator results in

$$\hat{\boldsymbol{\theta}}_{LMS} = \arg\min_{\boldsymbol{\theta}} \operatorname{med}_i \left\{r_i^2\right\} \tag{6}$$

where $\operatorname{med}_i \left\{r_i^2\right\}$ is the median of the squared residuals. This estimator has been shown to exhibit very strong robustness to outliers [17] having, by example, the maximum possible breakdown point (50%). Nevertheless, the LMS estimator cannot be computed in closed form [19] [20] and is thus of limited applicability especially in real-time scenarios. For a qualitative understanding of the robustness of the LMS estimator notice that in case of linear regression problems of the form

$$y_i = \theta_1 x_i + \theta_2, \tag{7}$$

the LMS line corresponds to the center of the stripe in the $x,y$ plane containing half plus one of the data points. Intuitively, it appears that minimizing a cost function as the one in Eq. (6) achieves higher robustness to outliers then cost functions as in (5), because the median of the squared residuals gives a "global" measure of the scatter of the residuals. M or LS cost functions, to the contrary, are not directly related to the distribution of the resulting residuals. Indeed, the by now classical example of F. J. Anscombe [1] explicitly shows how the same LS parameter estimates and residuals can be obtained for very different data sets. Other outlier robust model coefficient estimation techniques as Random Sample and Consensus (RANSAC, [8]) can be casted among the so-called "voting" techniques where a sufficiently large consensus set of data points in agreement with a specific value of the model coefficients is sought for. These methods may be indeed effective, but tend to be computationally demanding.

The proposed Least Entropy-Like (LEL) estimator is designed with the twofold objective of obtaining an estimator that directly relates to the distribution of residuals (in order to achieve high robustness to outliers) while also being quickly computable from a numerical view point. The LEL estimator was first presented in [11]: given the residual $r_i$ in Eq. (2), define:

$$D = \sum_{j=1}^{n} r_j^2, \tag{8}$$

namely the LS estimation cost. Then define the *relative squared residuals* $q_i$ as

$$\text{if } D \neq 0 \implies q_i := \frac{r_i^2}{\sum_{j=1}^n r_j^2} \quad : \quad q_i \in [0,1] \ \text{ and } \ \sum_{i=1}^{n} q_i = 1, \tag{9}$$

and finally

$$H = \begin{cases} 0 & \text{if } D = 0 \\ -\frac{1}{\log n} \sum_{i=1}^n q_i \log q_i & \text{otherwise.} \end{cases} \tag{10}$$

The function $H$ in equation (10) enjoys all the mathematical properties of a normalized *entropy* [6] associated to the sequence of "probability" - like $q_i \ : \ i = 1,2,\ldots,n$. In particular:

$$H \in [0,1] \tag{11}$$

$$H = 0 \quad \text{if and only if} \quad \begin{cases} r_i = 0 \ \ \forall \ \ i \in [1,n] \\ \text{or} \\ \exists \ ! \ \ i^* : r_{i^*} \neq 0 \text{ and } r_i = 0 \ \ \forall \ \ i \neq i^* \end{cases} \tag{12}$$

$$H = 1 \quad \text{if and only if} \quad r_i^2 = r_j^2 \neq 0 \ \ \forall \ \ i,j \in [1,n]. \tag{13}$$

Indeed, the above is formally equivalent to the Entropy of Information as introduced by Shannon in the 1948 [18] in analogy with the concept that was already known in thermodynamic and mechanical statistics, where Clausius and Boltzamann gave the first functional expression of the entropy, as a measure of the degree of disorder in a thermodynamic system.

Shannon, in particular, defined the entropy of information as a propriety associated to any probability distribution, while, the so-called *experimental entropy* used in thermodynamic is a property of real physic measurements.

Letting $X$ be a random variable with possible outcome $x_i$ ($i$=1,...,$n$), its mass probability $p_i$ such that $\sum_{i=1}^{n} p_i = 1$ identifies a global uncertainty measure [18] through the function:

$$H(P) = -\sum_{i=1}^{n} p_i \ln p_i \tag{14}$$

exhibiting the following properties:

- *H(P)* is concave.
- It is equal to zero (perfect certainty) when one of the probabilities is exactly 1.
- It reaches a maximum for uniform probabilities (complete ignorance): $p_1 = p_2 = \ldots = p_n = 1/n$.
- The entropy *H(P)* is a function of the probability distribution and not a function of the actual values taken by the random variable.

With respect to the above properties of entropy, the only possible difference of the entropy-like cost function defined in Eq. (10) is related to the possible singularity $D = 0$. Notice that this would correspond to a perfect LS fit that is quite unlikely. Indeed, there is no practical limitation as prior to computing $H$ in (10) one can always check if the LS fit is perfect. In such case, there is of course no need to compute any other estimate of the parameters. Also notice that for null values of $q_i$ the terms $0\log 0 = \log 0^0$ in equation (10) are zero.

In words, it can be stated that when the relative squared residuals $q_i$ are properly defined (i.e. $D \neq 0$), the $H$ function is a measure of their spread. When they are not properly defined, it is simply because the residuals are all identically null which corresponds to a null value of $H$ exactly as in the case when all the residuals are zero except one. In Physics, the (Gibbs) entropy of a system admitting $n$ discrete states with probabilities $p_1, p_2, \ldots, p_n$ is computed as $-\sum_{i=1}^{n} p_i \log p_i$. It is a very well-known fact that such function is a very sensitive measure of the distribution of the probabilities. Configurations with only a fraction of highly probable states have a much lower entropy of configurations where most states are approximately equally probable. Motivated by this fact, the function $H$ is defined with the aim of computing a robust estimate of the model parameter vector $\theta$. In particular, given that the entropy-like function $H$ as defined by Eq. (8) depends on $\boldsymbol{\theta}$ through the residuals $r_i$ (equation (2)), the following estimator is proposed:

$$\hat{\boldsymbol{\theta}}_{LEL} := \arg\min_{\boldsymbol{\theta}} H \tag{15}$$

where LEL stands for Least Entropy-Like. Such name was chosen with the twofold objective ($i$) of underlining that the $H$ function is not properly an entropy and ($ii$) of avoiding confusion with the Minimum Entropy estimation approach described, by example, in [21][23]. The idea behind the $\hat{\boldsymbol{\theta}}_{LEL}$ estimator defined in (15) is that such estimate will correspond either to making all the residuals null, or to making the relative squared residuals as little equally distributed as possible according to the entropy-like function $H$, the available data and the model structure. Notice that due to the normalization of the relative squared residuals $q_i$ in (9), forcing them to be “as little equally distributed as possible” means that “most” residual $r_i$ will need to be “small” (with respect to the normalization constant, namely the Least Squares cost $D$) and “a few” of the residuals

$r_i$ will need to be "large". Data points corresponding to these "large" residuals are candidate outliers. Stated differently, the key to robustness with respect to outliers is related to the fact that the devised penalty function does not directly measure the (weighted) mean square error (that as known tends to level out or "low pass" residuals), but rather the distribution of the relative squared errors. In particular, the devised LEL method tends to enforce a positive skewness to the distribution of the squared relative errors according to a metric give by the entropy-like function $H$ in Eq. (10).

Notice that, in general, there is no guarantee for the $H$ function to have a unique minima with respect to $\boldsymbol{\theta}$. Indeed, the entropy-like penalty function $H$ is highly nonlinear and may have many local minima. The minimization of $H$ needs to be carried out numerically paying attention to the initialization of $\boldsymbol{\theta}$: indeed the proposed estimator should be regarded as a local in nature. The gradient and Hessian matrix of the LEL cost functions can be analytically computed in closed form to aid numerical minimization routines. In case of models that are linear in the parameters as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{16}$$

being $\mathbf{y} \in \mathbb{R}^{n\times 1}$ the $n-$dimensional measurement vector, $\mathbf{X} \in \mathbb{R}^{n\times m}$ the regression matrix, $\boldsymbol{\theta} \in \mathbb{R}^{m\times 1}$ the parameter vector and $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ the measurement noise vector, it can be shown that the gradient of the $H$ cost is always well defined and the elements of the Hessian matrix result eventually ill posed (i.e. infinite) if and only if a residual $r_{i^*} = 0$ for some $i^*$. This is a highly unlikely situation in practice and even if it should occur an approximation of the Hessian can be computed through a regularization technique based on replacing $r_{i^*} = 0$ with $r_{i^*} = \delta$ for a sufficiently small $\delta$. The gradient and Hessian values of $H$ for the linear in the parameter model (16) have been explicitly computed, but are not here reported for the sake of brevity. Their closed form expressions are used to numerically compute the (local) minimum of $H$ in the case studies described in the paper. For a deeper discussion about the properties of the LEL estimator refer to [11].

For a qualitative and intuitive understanding of the proposed method, the LEL estimates of the of Anscombe data sets are reported in Fig. (1). Anscombe's data sets (or Anscombe's quartet) are four artificial data sets proposed in 1973 by Francis Anscombe [1] to illustrate the importance of graphs and plots in the interpretation of statistical analysis. Each data set is made of 11 $(x,y)$ points in a plane: the plot of the four data sets immediately and intuitively reveals the different structure of the four sets. Yet, if a line $y = \theta_1 x + \theta_0$ is fitted to the data through OLS, the same model coefficients $\hat{\theta}_{LS_1} = 0.5$ and $\hat{\theta}_{LS_0} = 3$ are found for all four data sets. Moreover, the $y$ residual sum of squares of the four data sets is the same revealing the difficulty in analyzing the fitting results in the presence of outliers or of a mismatching model. The LEL estimate $\hat{\theta}_{LEL_1}$ and $\hat{\theta}_{LEL_0}$ of the line $y = \theta_1 x + \theta_0$ coefficients of Anscombe's data sets are expected to generate residuals that are smaller than the OLS residuals for the majority of the points. As illustrated in Fig. (1) and Fig. (2) this is indeed the case for the first three data sets. In the third case where 10 out of the 11 data points fit the model, the LEL method perfectly succeeds in fitting the 10 inliers, whereas the OLS estimated is biased. The fourth is a singular case, as the "real" $\theta_1$ for these points should be infinite and both OLS and LEL, of course, fail. The results plotted in in Fig. (1) and Fig. (2) were obtained by numerically minimizing the LEL cost function starting from the OLS solution.

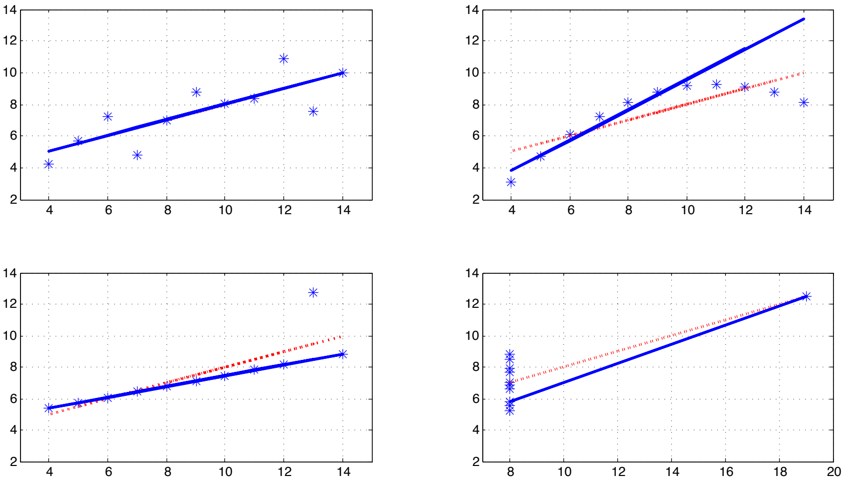

**Fig. 1.** OLS (red dotted lines) and LEL estimates (blue solid line) for Anscombe's Quarted [1] data sets (blue $*$ points, starting from the top left plot in clockwise direction: sets I, II, III and IV). Notice that in the lower left case the LEL estimator yields the perfect solution, namely the line of the 10 inliers. Also notice that in the bottom right case the regression model is singular hence both the LEL and OLS methods fail.

# 3 Generalized Maximum Entropy

Building on the concept of information entropy introduced by Claude Shannon [18], the Maximum Entropy Principle (MEP) was firstly proposed by Edwin T. Jaynes [12] [13] defining an objective method for estimating probability distributions in case of limited data. A generalization of the MEP is given to the contribution of Amos Golan *et al.* [9], that proposed an alternative method for parameters estimation called Generalized Maximum Entropy (GME), as an extension of the MEP.

## 3.1 Shannon's Entropy Measure and the Maximum Entropy Principle

Edwin Jaynes, building on the Shannon's Entropy function in Eq. (14), proposed the *Maximum Entropy Principle* (MEP) [12], [13] to estimate the probability distributions in presence of constraints generated from the data, and given in the form of expectations. Under MEP, the probability distribution is chosen among those distributions consistent with known information (the constraints), that maximizes the entropy. The MEP can be used to solve pure inverse problems defined as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{p} \tag{17}$$

where $\mathbf{y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times m}$, and $\mathbf{p} \in \mathbb{R}^{m\times 1}$ even for $n < m$. To recover the unknown probability $\mathbf{p}$ vector, the MEP suggests to maximize the *H(P)* function (14) subjected

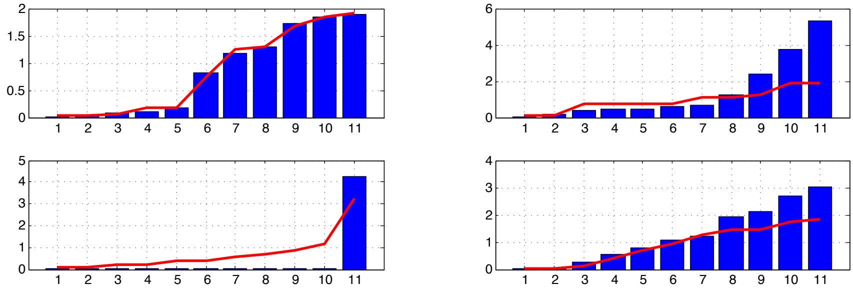

**Fig. 2.** Sorted absolute value of the residuals for the LEL estimate (bar plot in blue) and the OLS estimate (solid red line) for Anscombe's quartet in figure (1). Cases II and III, in particular, reveal the enhanced outlier robustness of the LEL approach as compared to OLS.

to data consistency and normalization constraints. The data consistency constraint is defined by the Eq. (17). The normalization constraints imply that: $\mathbf{p}^T\mathbf{1} = 1$ being $\mathbf{1}$ an $m$-dimensional column vector having all components equal to one.

Using Lagranges method we can carry out the analytical solution to the entropy maximization problem as follows:

$$L = -\mathbf{p}^T \ln \mathbf{p} + \boldsymbol{\lambda}^T(\mathbf{y} - \mathbf{X}\mathbf{p}) + \mu(1 - \mathbf{p}^T\mathbf{1}) \tag{18}$$

and the corresponding first-order conditions are:

$$\partial L/\partial \mathbf{p} = -\ln \mathbf{p} - 1 - \mathbf{X}^T\boldsymbol{\lambda} - \mu = 0$$

$$\partial L/\partial \boldsymbol{\lambda} = \mathbf{y} - \mathbf{X}\mathbf{p} = 0$$

$$\partial L/\partial \mu = 1 - \mathbf{p}^T\mathbf{1} = 0$$

The solution of the above conditions will lead to the following estimated value of $\mathbf{p}$:

$$\hat{\mathbf{p}} = \exp(-\mathbf{X}^T\hat{\boldsymbol{\lambda}}) / \sum_j \exp(-\mathbf{X}^T\hat{\boldsymbol{\lambda}}) \tag{19}$$

where $\Omega(\hat{\boldsymbol{\lambda}}) = \sum_j \exp(-\mathbf{X}^T\hat{\boldsymbol{\lambda}})$ is the normalization factor, known also as the partition function, that transforms the relative probabilities into absolute probabilities. The solution (19) can be applied to solve ill-posed problems, as the classical example of the Jaynes's dice experiment: in this case the unknowns are represented by the six unknown probabilities of the dice faces, the term $\mathbf{X}$ in Eq. (17) is determined by the dice model (i.e. $\mathbf{X} = (1,2,3,4,5,6)$) and the a priori knowledge is represented by the average of the outcome, namely 3.5 in case of fair dice. This is an ill-posed problem with one observation and six unknowns, which can be solved by maximizing the constrained *H(P)* function (14).

In case the observed moments are noisy, for instance coming from a sampling experiment, the consistency constraint became stochastic and model (17) can be modified by adding an error term:

$$\mathbf{y} = \mathbf{X}\mathbf{p} + \boldsymbol{\varepsilon} \tag{20}$$

The idea behind Eq. (20) is that the term $\boldsymbol{\varepsilon}$ (in general not equal to zero as in equation (17)) allows to model stochastic moments in $\mathbf{y}$. Consequently, the samples moments are allowed (but not forced) to be different from the underlying population moments, a flexibility that seems natural for finite data sets. Further details are illustrated in the following section adressing the regression model.

### 3.2 GME Regression Model

The GME estimator is consistent and asymptotically normal under some regularity conditions. The idea underlying the GME estimator consists in viewing the parameters and the error vectors as convex linear combinations of some known discrete support values and unknown proportions to be interpreted as probabilities. Considering a regression model with $n$ observations and $m$ variables:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{21}$$

in order to use the MEP method to estimate the regression parameters, the coefficients and the error terms are re-parameterized as a convex combination of expected values of discrete random variables.

Given a parameter $\boldsymbol{\theta}_j$, it is always possible to write it [15] as a convex combination of a support variable as: $\boldsymbol{\theta}_j = \mathbf{z}_j^T \mathbf{p}_j$, where $\mathbf{z}_j^T = [z_{j1}, ..., 0, ..., z_{jM}]$ defines the lower and upper bounds of the $j^{th}$ parameter, with $M$ usually [9] in the interval $2 \leq M \leq 7$. The vector $\mathbf{p}_j^T = [p_{j1}, ..., p_{jM}]$ contains positive probabilities that sum to one.

Similarly, each error term is treated as a discrete random variable: $\boldsymbol{\varepsilon}_i = \mathbf{v}_i^T \mathbf{w}_i$, where $\mathbf{v}_i^T = [v_{i1}, ..., 0, ..., v_{iN}]$ defines the error bound, with $M$ usually [9] in the interval $2 \leq N \leq 7$. The vector $\mathbf{w}_i^T = [w_{i1}, ..., w_{iN}]$ contains positive probabilities that sum to one.

The GME method, therefore, estimates the regression coefficients and the error terms, by recovering the probability distribution of a discrete random variables set. The model (21) can be rewritten as follows:

$$\mathbf{y} = \mathbf{X}\left(\mathbf{I}_{m\times m} \bigotimes \mathbf{z}^T\right)\mathbf{p} + \left(\mathbf{I}_{n\times n} \bigotimes \mathbf{v}^T\right)\mathbf{w} \tag{22}$$

where $\mathbf{I}_{m\times m}$ and $\mathbf{I}_{n\times n}$ are the identity matrices for the parameters and the error terms and the symbol $\bigotimes$ is the Kronecker product.

The supervectors $\mathbf{p}$ and $\mathbf{w}$ contain respectively $m$ and $n$ probabilities vectors, related at each support variable $\mathbf{z}_j$ and $\mathbf{v}_i$. The aim is to estimate the probabilities vectors $\mathbf{p}_j\{j = 1, \ldots, m\}$ and $\mathbf{w}_i\{i = 1, \ldots, n\}$, associated respectively to the $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ parameters. The estimation is made by the maximization of the Shannon's entropy function:

$$H(\mathbf{p}, \mathbf{w}) = -\mathbf{p}^T \ln(\mathbf{p}) - \mathbf{w}^T \ln(\mathbf{w}) \tag{23}$$

subjected to the *consistency constraints*, that represent a part of the regression model (22), and *normalization constraints*, that means, the element of the probabilities vectors $\mathbf{p}$ and $\mathbf{w}$, have to satisfy respectively the conditions of containing positive probabilities that sum to one.

The optimization problem is obtained via definition of the Lagrangian function, which can be easily solved in the same fashion we reported for the MEP case. The GME has advantages to address some circumstances, as for instance ill-behaved data or no distributional error assumptions ([4], [5], [9]).

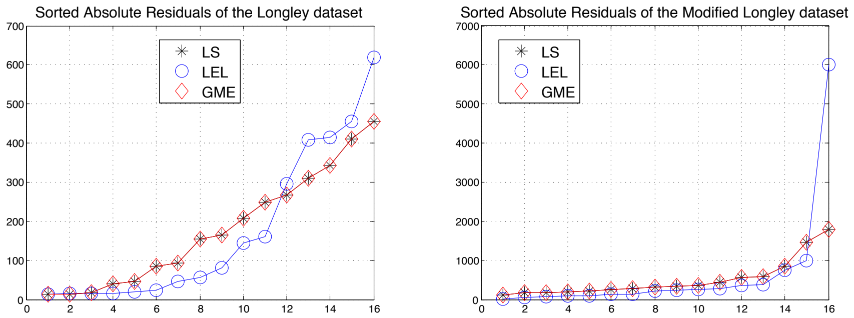

**Fig. 3.** Sorted Absolute Residuals of the Longley dataset (left) and of the Modified (outlier case) Longley dataset (right). Refer to the text for details.

## 4 Validation Study: The Longley Data Set Case

In order to compare the proposed entropy, based methods, we have considered the Longley Data Set [14] available through the Internet on the NIST - National Institute of Standards and Technology's website. According to NIST the Linear Least Squares Regression problem on this data set has a "Higher Level of Difficulty". Moreover it is reported that *this classic dataset of labor statistics was one of the first used to test the accuracy of least squares computations. The response variable (y) is the Total Derived Employment and the predictor variables are GNP Implicit Price Deflator with Year 1954 = 100 (x1), Gross National Product (x2), Unemployment (x3), Size of Armed Forces (x4), Non-Institutional Population Age 14 & Over (x5), and Year (x6).*

The difficulty in processing with OLS this data set is basically related to the large condition number ($\kappa = 4.8593\text{E}+09$) of the associated regression matrix. This is a typical situation (close to perfect multicollinearity) where the GME approach is particularly useful. The sorted absolute values of the residuals obtained by the GME, LEL, and OLS estimation approaches are depicted in the left plot of Fig. 3. Notice that the GME and OLS solutions are in perfect agreement (in spite of the fact that the OLS estimate is computed by inverting a matrix that is very close to being singular). To the contrary, the LEL estimate yields a different solution. In particular, the LEL Score (i.e. the percentage of LEL residuals in absolute value being smaller than the OLS residuals in absolute value) is of 56.25%. In order to illustrate the effectiveness of the LEL approach to compute robust estimates in the presence of outliers, the Longley data set has been modified by replacing the last $y$ (explanatory variable) value with another $y$ value, in particular imposing $y(16) = y(12)$. All the other data values are left identical. This new data set will be referred to as the "Modified Longley Dataset". The sorted absolute values of the residuals obtained by the GME, LEL and OLS estimation approaches on the Modified Longley Data set are depicted in the right plot of Fig. 3: remarkably, the LEL Score in this case results in 93.75% confirming the robustness of the LEL solution to outliers. Moreover, notice that the largest residual (in absolute value) in this case

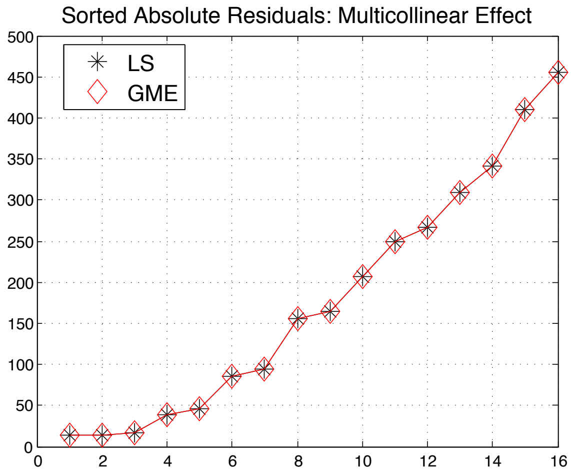

**Fig. 4.** Multicollinear Effect. Sorted Absolute Residuals of the GME and OLS approaches on a multicollinear modified data set and on the original data set respectively.

corresponds precisely to $y(16)$. Also notice that, once again, the GME and OLS solutions for the Modified Longley Data set are in perfect agreement.

At last, in order to confirm the ability of the GME approach to cope with multicollinearity, the regression matrix of the (original) Longley data set is modified by repeating one of its columns. This yields a singular regression matrix on which both the OLS and LEL algorithms cannot be applied. To the contrary, the GME solution can be computed without numerical issues and the obtained residuals are perfectly equivalent to the OLS residuals obtained with the original (non modified) data set. This is confirmed in the plot of Fig. 4.

## 5 Conclusion and Discussion

The objective of this paper was to describe two instances of entropy-based estimators illustrating their properties and their performances as compared with more standard methods as the OLS. The described methods consist in the Least Entropy-like LEL and the Generalized Maximum Entropy estimators. The first is particularly useful for model coefficient estimation in the presence of outliers, whereas the second is robust to multicollinearity. Moreover, the LEL solution, although local in nature and hence potentially sensitive to its initialization, is computationally much less demanding than alternative outlier robust approaches as LMS [17] or RANSAC [8]. In order to illustrate the specific characteristics of the two approaches, both methods have been applied to the Longley Data Set [14]. The results confirm the expected outlier robustness properties of LEL and the multicollinearity robustness of GME. The integration of the potentialities of both estimators is one future objective of our research plan.

## References

1. Anscombe, F.: Graphs in statistical analysis. The American Statistician 27(1), 17–21 (1973)
2. Bai, E.W.: An optimization based robust identication algorithm in the presence of outliers. Journal of Global Optimization 23(3), 195–211 (2002)
3. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
4. Ciavolino, E., Al-Nasser, A.: Comparing generalised maximum entropy and partial least squares methods for structural equation models. Journal of Nonparametric Statistics 21(8), 1017–1036 (2009)
5. Ciavolino, E., Dahlgaard, J.: Simultaneous equation model based on the generalized maximum entropy for studying the effect of management factors on enterprise performance. Journal of Applied Statistics 36(7), 801–815 (2009)
6. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, Inc. (2001), http://dx.doi.org/10.1002/0471200611
7. Distante, C., Indiveri, G.: RANSAC-LEL: An optimized version with least entropy like estimators. In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 1425–1428. IEEE (2011), http://dx.doi.org/10.1109/ICIP.2011.6115709
8. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 381–395 (1981), http://dx.doi.org/10.1145/358669.358692
9. Golan, A., Judge, G.G., Miller, D.: Maximum Entropy Econometrics: Robust Estimation with Limited Data. John Wiley & Sons Inc. (1996)
10. Huber, P., Ronchetti, E.: Robust Statistics. Wiley, Hoboken (2009)
11. Indiveri, G.: An entropy-like estimator for robust parameter identification. Entropy 11(4), 560–585 (2009)
12. Jaynes, E.: Information theory and statistical mechanics. Physical Review 106(4), 620–630 (1957)
13. Jaynes, E.: Prior probabilities. IEEE Transactions on Systems Science and Cybernetics 4(3), 227–241 (1968)
14. Longley, J.: An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association 62(319), 819–841 (1967)
15. Paris, Q.: Multicollinearity and maximum entropy estimators. Economics Bulletin 3(11), 1–9 (2001)
16. Poljak, B., Tsypkin, J.: Robust identification. Automatica 16(1), 53–63 (1980)
17. Rousseeuw, P., Leroy, A.: Robust regression and outlier detection. Wiley, New York (2003)
18. Shannon, C.E.: A mathematical theory of communications. Bell System Technical Journal 27(7), 379–423 (1948)
19. Steele, J., Steiger, W.: Algorithms and complexity for least median of squares regression. Discrete Applied Mathematics 14(1), 93–100 (1986)
20. Stromberg, A.: Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. SIAM Journal on Scientific Computing 14(6), 1289–1299 (1993)
21. Ta, M., DeBrunner, V.: Minimum entropy estimation as a near maximum-likelihood method and its application in system identification with non-gaussian noise. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 2, pp. ii–545. IEEE (2004)
22. Tribus, M., McIrvine, E.C.: Energy and information. Scientic American 225(3), 179–188 (1971)
23. Wolsztynski, E., Thierry, E., Pronzato, L.: Minimum-entropy estimation in semi-parametric models. Signal Processing 85(5), 937–949 (2005)