Normalized Entropy Aggregation for Inhomogeneous Large-Scale Data[⋆]

Maria da Conceição Costa and Pedro Macedo

Department of Mathematics and CIDMA – Center for Research and Development in Mathematics and Applications, University of Aveiro, 3810-193, Aveiro, Portugal
{lopescosta,pmacedo}@ua.pt
http://www.ua.pt

Abstract. *It was already in the fifties of the last century that the relationship between information theory, statistics, and maximum entropy was established, following the works of Kullback, Leibler, Lindley and Jaynes. However, the applications were restricted to very specific domains and it was not until recently that the convergence between information processing, data analysis and inference demanded the foundation of a new scientific area, commonly referred to as Info-Metrics [1, 2]. As huge amount of information and large-scale data have become available, the term "big data" has been used to refer to the many kinds of challenges presented in its analysis: many observations, many variables (or both), limited computational resources, different time regimes or multiple sources. In this work, we consider one particular aspect of big data analysis which is the presence of inhomogeneities, compromising the use of the classical framework in regression modelling. A new approach is proposed, based on the introduction of the concepts of info-metrics to the analysis of inhomogeneous large-scale data. The framework of information-theoretic estimation methods is presented, along with some information measures. In particular, the normalized entropy is tested in aggregation procedures and some simulation results are presented.*

Keywords: Big Data, Info-Metrics, Maximum Entropy, Normalized Entropy

1 Introduction

Inference and processing of limited information is still one of the most fascinating universal problems. As stated by Amos Golan in [2], a very recent publication, "[...] the available information is most often insufficient to provide a unique answer or solution for most interesting decisions or inferences we wish to make. In fact, insufficient information - including limited, incomplete, complex, noisy and uncertain information - is the norm for most problems across all disciplines."

[⋆] Final version of this work is published in the book Theory and Applications of Time Series Analysis (`https://doi.org/10.1007/978-3-030-26036-1_2`), from the Contributions to Statistics book series.

Also, regardless of the system or question studied, any researcher observes only a certain amount of information or evidence and optimal inference must take into account the relationship between the observable and the unobservable, [3].

Info-Metrics is a constrained optimization framework for information processing, modelling and inference with finite, noisy or incomplete information. It is at the intersection of information theory, statistical methods of inference, applied mathematics, computer science, econometrics, complexity theory, decision analysis, modelling and the philosophy of science, [2].

As Info-Metrics generalizes the Maximum Entropy (ME) principle by Jaynes, [4, 5], which in turn relies on the maximization of Shannon's entropy, the notions of information, uncertainty and entropy are fundamental to the understanding of the methodologies involved. Each scientist and discipline have their own interpretation and definition of information within the context of their research and understanding but, in the context of Info-Metrics, it refers to the meaningful content of data, it's context and interpretation and how to transfer data from one entity to another. As for uncertainty, it arises from a proposition or a set of possible outcomes where none of the choices or outcomes is known with certainty (a proposition is uncertain if it is consistent with knowledge but not implied by knowledge). Therefore, these outcomes are represented by a certain probability distribution. The more uniform the distribution, the higher the uncertainty that is associated with this set of propositions or outcomes. Finally, the concept of entropy reflects what, on average, we expect to learn from observations and it depends on how we measure information. Technically, entropy is a measure of uncertainty of a single random variable. As such, entropy can be viewed as a measure of uniformity.

For a brief discussion of entropy, let us consider the set $\boldsymbol{A} = \{a_1, a_2, \cdots, a_K\}$ to be a finite set and $\boldsymbol{p}$ a proper probability mass function on $\boldsymbol{A}$. The amount of information needed to fully characterize all of the elements of this set consisting of K discrete elements is defined by the Hartley's formula, $I(\boldsymbol{A}_K) = \log_2 K$. Shannon's information content of an outcome a_k is $h(a_k) = h(p_k) \equiv \log_2 \frac{1}{p_k}$. Shannon's entropy reflects the expected information content of an outcome and is defined as

$$H(\boldsymbol{p}) \equiv \sum_{k=1}^{K} p_k \log_2 \frac{1}{p_k} = -\sum_{k=1}^{K} p_k \log_2 p_k = E\left[\log_2\left(\frac{1}{p(X)}\right)\right], \qquad (1)$$

for the random variable X. This information criterion, expressed in bits, measures the uncertainty of X that is implied by $\boldsymbol{p}$. The entropy measure $H(\boldsymbol{p})$ reaches a maximum when $p_1 = p_2 = \cdots = p_K = \frac{1}{K}$ and a minimum with a point mass function. The entropy $H(\boldsymbol{p})$ is a function of the probability distribution $\boldsymbol{p}$ and not a function of the actual values taken by the random variable.

The remainder of the paper is laid out as follows: in Section 2, maximum entropy and generalized maximum entropy estimation are briefly discussed. Section 3 illustrates some traditional aggregation procedures and a new proposal based on normalized entropy. Section 4 presents simulation results. Some conclusions and topics for future research are given in Section 5.

2 Generalized Maximum Entropy Estimator

The ME principle was discussed by Golan, Judge and Miller, [6], in order to develop analytical and empirical methods for recovering the unobservable parameters of a pure linear inverse problem. Considering then

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{p}, \tag{2}$$

where $\boldsymbol{y}$ is the vector $(N \times 1)$ of observations, $\boldsymbol{X}$ is a non-invertible matrix $(N \times K)$ with $N < K$, and $\boldsymbol{p}$ is the vector $(K \times 1)$ of unknown probabilities, the ME principle consists in choosing $\boldsymbol{p}$ that maximizes Shannon's entropy

$$H(\boldsymbol{p}) = -\sum_{k=1}^{K} p_k \ln p_k = -\boldsymbol{p}' \ln \boldsymbol{p}, \tag{3}$$

subject to the data consistency restriction, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{p}$, and the additivity restriction, $\boldsymbol{p}'\boldsymbol{1} = 1$. Formally, the ME estimator is given by

$$\underset{\boldsymbol{p}}{\operatorname{argmax}} \left\{-\boldsymbol{p}' \ln \boldsymbol{p}\right\}, \tag{4}$$

subject to the model consistency and additivity constraints,

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X}\boldsymbol{p} \\ \boldsymbol{1}'\boldsymbol{p} = 1 \end{cases}. \tag{5}$$

There is no closed-form analytical solution, but a numerical approximation can be obtained using the Lagrange multipliers. It can be said that the Jaynes maximum entropy formalism has enabled us to solve the pure inverse problem with this optimization (maximization) procedure, regarding it as an inference problem. The ME principle is the basis for transforming the information in the data into a probabilistic distribution that reflects our uncertainty about individual outcomes.

To extend the ME estimator to the linear regression model represented by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{6}$$

where, as usually, $\boldsymbol{y}$ denotes a $(N \times 1)$ vector of noisy observations, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector of unknown parameters, $\boldsymbol{X}$ is a known $(N \times K)$ matrix of explanatory variables, and $\boldsymbol{e}$ is the $(N \times 1)$ vector of random disturbances (errors), Golan, Judge and Miller, [6], considered each β_k as a discrete random variable with a compact support and $M \geq 2$ possible outcomes and each e_n as a finite and discrete random variable with $J \geq 2$ possible outcomes. The error vector is considered here as another vector of unknown parameters to be estimated simultaneously with the vector $\boldsymbol{\beta}$. In this context, the linear regression model is represented as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} + \boldsymbol{V}\boldsymbol{w}, \tag{7}$$

where

$$\boldsymbol{\beta} = \boldsymbol{Zp} = \begin{bmatrix} \boldsymbol{z}_1' & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{z}_2' & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{z}_K' \end{bmatrix} \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_K \end{bmatrix}, \tag{8}$$

and

$$\boldsymbol{e} = \boldsymbol{Vw} = \begin{bmatrix} \boldsymbol{v}_1' & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{v}_2' & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{v}_N' \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix}. \tag{9}$$

Matrices $\boldsymbol{Z}$ $(K \times KM)$ and $\boldsymbol{V}$ $(N \times NJ)$ are the matrices of support values and vectors $\boldsymbol{p}$ $(KM \times 1)$ and $\boldsymbol{w}$ $(NJ \times 1)$ are the vectors of unknown probabilities to be estimated. Note that each β_k, $k = 1, 2, \dots, K$, and each e_n, $n = 1, 2, \dots, N$, are viewed as expected values of discrete random variables $\boldsymbol{z}_k$ and $\boldsymbol{v}_n$, respectively, with $M \geq 2$ and $J \geq 2$ possible outcomes, within the lower and upper bounds of the corresponding support spaces. Thus, the generalized maximum entropy (GME) estimator is given by

$$\underset{\boldsymbol{p},\boldsymbol{w}}{\operatorname{argmax}} \left\{ -\boldsymbol{p}' \ln \boldsymbol{p} - \boldsymbol{w}' \ln \boldsymbol{w} \right\}, \tag{10}$$

subject to the consistency (with the model) and additivity (for $\boldsymbol{p}$ and $\boldsymbol{w}$) constraints,

$$\begin{cases} \boldsymbol{y} = \boldsymbol{XZp} + \boldsymbol{Vw}, \\ \boldsymbol{1}_K = (\boldsymbol{I}_K \otimes \boldsymbol{1}_M')\boldsymbol{p}, \\ \boldsymbol{1}_N = (\boldsymbol{I}_N \otimes \boldsymbol{1}_J')\boldsymbol{w}, \end{cases} \tag{11}$$

where $\otimes$ represents the Kronecker product. The optimal probability vectors, $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{w}}$, are used to obtain point estimates of the unknown parameters and the unknown errors with $\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{e}} = \boldsymbol{V}\widehat{\boldsymbol{w}}$. Some properties of the GME estimator, such as consistency and asymptotic normality, are discussed in detail, for example, in Mittelhammer, Cardell and Marsh, [7].

3 Large-Scale Data and Aggregation

Large-scale data or big data usually refers to datasets that are large in different ways: many observations, many variables (or both); observations are recorded in different time regimes or are taken from multiple sources. Some difficult issues arise in dealing with this kind of data like, for instance, retaining optimal (or, at least, reasonably good) statistical properties with a computationally efficient analysis; or dealing with inhomogeneous data that does not fit in the classical framework: data is neither i.i.d. (exhibiting outliers or not belonging to same distribution) nor stationary (time-varying effects my be present).

Standard statistical models (linear or generalized linear models for regression or classification) fail to capture inhomogeneity structure in data, compromising estimation and interpretation of model parameters, and, of course, prediction. On the other hand, statistical approaches for dealing with inhomogeneous data (such as varying-coefficient models, mixed effects models, mixture models or clusterwise regression models) are typically very computationally cumbersome.

Ignoring heterogeneity in data, computational burden can be addressed with the following procedure, [8]: firstly, construct g groups from the large-scale data (groups may be overlapping and may not cover all observations in the sample); then, for each group compute an estimator, $\widehat{\boldsymbol{\beta}}_g$, through standard techniques (e.g., OLS, ridge or LASSO); finally, considering the ensemble of estimators, aggregate them into a single estimator, $\widehat{\boldsymbol{\beta}}$.

3.1 Traditional Aggregation Procedures

Several aggregation procedures have been already proposed in literature. Three of them are presented next.

1. Bagging: this procedure results in less computational complexity and even allows for parallel computing. It simply averages the ensemble estimators with equal weight to obtain the aggregated estimator, [8, 9]:

$$\widehat{\boldsymbol{\beta}} := \sum_{g=1}^{G} w_g \widehat{\boldsymbol{\beta}}_g, \tag{12}$$

   where $w_g = \frac{1}{G}$ for all $g = 1, 2, \ldots, G$. The estimates $\widehat{\boldsymbol{\beta}}_g$ are obtained from bootstrap samples, where the groups are sampled with replacement from the whole data. It is a simple procedure and the weights do not depend on the response $\boldsymbol{y}$, but it is not suitable for inhomogeneous data.
2. Stacking: instead of assigning a uniform weight to each estimator, [10] and [11] proposed the aggregated estimator

$$\widehat{\boldsymbol{\beta}} := \sum_{g=1}^{G} w_g \widehat{\boldsymbol{\beta}}_g, \tag{13}$$

   where

$$\boldsymbol{w} := \underset{\boldsymbol{w} \in \boldsymbol{W}}{\operatorname{argmin}} \left\| \boldsymbol{y} - \sum_{g=1}^{G} w_g \widehat{\boldsymbol{y}}_g \right\|_2, \tag{14}$$

   and, using a ridge constraint, $\boldsymbol{W} = \{\boldsymbol{w} : \|\boldsymbol{w}\| \leq s\}$, for some $s > 0$, or using a sign constraint, $\boldsymbol{W} = \{\boldsymbol{w} : \min\limits_{g} w_g \geq 0\}$, or using a convex constraint, $\boldsymbol{W} = \{\boldsymbol{w} : \min\limits_{g} w_g \geq 0 \text{ and } \sum_{g=1}^{G} w_g = 1\}$. The idea is to find the optimal linear or convex combination of all ensemble estimators, but it is also not suitable for inhomogeneous data.

3. Magging: corresponds to maximizing the minimally "explained variance" among all data groups, [8], such that

$$\widehat{\boldsymbol{\beta}} := \sum_{g=1}^{G} w_g \widehat{\boldsymbol{\beta}}_g, \tag{15}$$

where

$$\boldsymbol{w} := \operatorname*{argmin}_{\boldsymbol{w} \in \boldsymbol{W}} \left\| \sum_{g=1}^{G} w_g \widehat{\boldsymbol{y}}_g \right\|_2 , \tag{16}$$

and $\boldsymbol{W} = \{\boldsymbol{w} : \min_{g} w_g \geq 0 \text{ and } \sum_{g=1}^{G} w_g = 1\}$. The idea is to choose the weights as a convex combination to minimize the $\|\cdot\|_2$ of the fitted values, $\widehat{\boldsymbol{y}}$. If the solution is not unique, it is considered the solution with lowest $\|\cdot\|_2$ of the weight vector among all solutions. This procedure was the first that we are aware of that was proposed for heterogeneous data. The main idea is that if an effect is common across all groups, then it cannot be "averaged away" by searching for a specific combination of the weights. The common effects will be present in all groups and will be retained even after the minimization of the aggregation scheme.

We believe the question as to weather the effects are really common across all groups may not be answered straightforwardly. If the groups carry information about the whole dataset and there are inhomogeneities, why should we consider that, with random sub-sampling, all groups are equally informative?

These considerations led us to the idea of choosing the groups according to their "information content".

3.2 Proposed Aggregation Procedure

To measure the information content in a system and to measure the importance of the contribution of each piece of data or constraint in reducing uncertainty, Golan, Judge and Miller, [6], stated that, in the ME formulation, the maximum level of entropy-uncertainty results when the information-moment constraints are not enforced and the distribution of probabilities over the K states is uniform. As each piece of effective data is added, there is a departure from the uniform distribution, which implies a reduction of uncertainty. The proportion of the remaining total uncertainty is measured by the normalized entropy (NE),

$$S(\widehat{\boldsymbol{p}}) = -\frac{\sum_k \widehat{p}_k \ln \widehat{p}_k}{\ln(K)}, \tag{17}$$

where $S(\widehat{\boldsymbol{p}}) \in [0,1]$ and $\ln(K)$ represents maximum uncertainty (the entropy level of the uniform distribution with K outcomes). A value $S(\widehat{\boldsymbol{p}}) = 0$ implies no uncertainty and a value $S(\widehat{\boldsymbol{p}}) = 1$ implies perfect uncertainty. Related to the normalized entropy, the information index (II) is defined as $1 - S(\widehat{\boldsymbol{p}})$ and measures the reduction in uncertainty.

In this work, we propose a new aggregation scheme that is based on identifying the information content of a given group through the calculation of the normalized entropy. The proposed NE aggregated estimator is then

$$\widehat{\boldsymbol{\beta}} := \sum_{g=1}^{G} w_g \widehat{\boldsymbol{\beta}}_g, \tag{18}$$

where w_g is defined by normalized entropy using GME,

$$S(\widehat{\boldsymbol{p}})_g = \frac{-\widehat{\boldsymbol{p}}' \ln \widehat{\boldsymbol{p}}}{K \ln M}, \tag{19}$$

for the signal, $\boldsymbol{X\beta}$, such that $\sum_{g=1}^{G} w_g = 1$. This aggregation procedure is a weighted average of the collection of regression coefficient estimates as in Bagging, Stacking and Magging. The idea is almost as simple as Bagging and it is expected to provide similar results if the data is homogeneous. However, since the weights in (18) will depend on the information content of each group according to (19), or some function of it, the weights will be, in general, non-uniform (as in Stacking and Magging) if the data is inhomogeneous.

Following section reports some simulated situations for which the NE aggregated estimator was calculated and compared to the aggregated estimator based on Bagging.

4 Simulation Study

A linear regression model was considered, where $\boldsymbol{X}$ is the simulated matrix of explanatory variables, drawn randomly from normal distributions; $\boldsymbol{\beta}$ is a vector of parameters, $\boldsymbol{e}$ is the vector of random disturbances, drawn randomly from normal distributions and $\boldsymbol{y}$ is the constructed vector of noisy observations. For this simulation, $\boldsymbol{\beta}$ was considered as

$$\boldsymbol{\beta} = [1.8,\ 1.2,\ -1.4,\ 1.6,\ -1.8,\ 2.0,\ -2.0,\ 0.2,\ -0.4,\ 0.6,\ 0.8]. \tag{20}$$

Necessary reparameterizations were done considering $M = 5$ and $J = 3$ and different matrices $\boldsymbol{Z}$ containing the supports for the parameters. The support matrix $\boldsymbol{V}$ containing the supports for the errors, was set considering symmetric and zero-centred supports using the three-sigma rule with the empirical standard deviation of the noisy observations.

Simulations were done considering $\boldsymbol{X}$ a (20000×11) matrix; $\boldsymbol{\beta}$ a (11×1) vector; $\boldsymbol{e}$ a (20000×1) vector and $\boldsymbol{y}$ a (20000×1) vector. The error distribution was considered to be normal, with mean value zero and standard deviation five. Several matrices $\boldsymbol{X}$ of explanatory variables were simulated, corresponding to different condition numbers (c.n.)[1]. Random sub-sampling with replacement was done considering different number of groups and 50 observations per group. The

[1] Ratio of the largest singular value of $\boldsymbol{X}$, with the smallest singular value.

Euclidean norm of the difference between the aggregated estimator $\widehat{\boldsymbol{\beta}}$ and the true parameter $\boldsymbol{\beta}$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$, is calculated for each simulated case and the results are given in Tables 1 – 5. For each case, three different solutions are presented, namely,

1. NE1: the chosen $\widehat{\boldsymbol{\beta}}$ corresponds to the GME estimate for the group with lower normalized entropy, (NE). This solution does not correspond, in fact, to an aggregated estimator; it corresponds to a chosen estimate amongst all groups;
2. NE2: the chosen $\widehat{\boldsymbol{\beta}}$ corresponds to the weighted average of the GME estimates of all groups, weighted by the information index, II, where II $= 1-$NE;
3. Bgg: the $\widehat{\boldsymbol{\beta}}$ chosen corresponds to Bagging (average of the OLS estimates of all groups).[2]

Table 1. Euclidean norm of the difference $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, with $\boldsymbol{z}_k = [-10, 10]$

n.g.	Solution	c.n.=1337
5	NE1	4.26
	NE2	4.18
	Bgg	181.23

Table 2. Euclidean norm of the difference $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, with $\boldsymbol{z}_k = [-10, 10]$

n.g.	Solution	c.n.=43030
5	NE1	4.22
	NE2	4.25
	Bgg	1432.59

The present results are intended to highlight the overall tendencies we encountered in the simulation study. Many other situations were simulated, with many different matrices of explanatory variables, $\boldsymbol{X}$, corresponding to a wide range of variation regarding the matrix condition number, which, as is well known, is related to the presence of collinearity[3] in the explanatory variables. In this paper, only two extreme cases were chosen to be presented, the first one corresponding to a relatively small condition number (c.n. around 1300) and the second one corresponding to a much higher condition number (c.n. around

[2] It is not considered here the case of a single learning set, as in [9], and the need to take repeated bootstrap samples from it.

[3] The concept is not used here in a literal sense. A discussion about similar notions of this concept is available in Belsley, Kuh and Welsch, [12, pp. 85–98].

Table 3. Euclidean norm of the difference $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, with $\boldsymbol{z}_k = [-10, 10]$

n.g.	Solution	c.n.=1337
5	NE1	4.26
	NE2	4.18
	Bgg	181.23
10	NE1	4.47
	NE2	4.31
	Bgg	171.22
50	NE1	4.45
	NE2	4.30
	Bgg	49.36
100	NE1	5.48
	NE2	4.34
	Bgg	38.74

Table 4. Euclidean norm of the difference $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, with $\boldsymbol{z}_k = [-100, 100]$

n.g.	Solution	c.n.=1337
5	NE1	32.31
	NE2	10.17
	Bgg	214.56

Table 5. Euclidean norm of the difference $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, with $\boldsymbol{z}_k = [-100, 100]$

n.g.	Solution	c.n.=1337	c.n.=43030
5	NE1	32.31	35.54
	NE2	10.17	15.59
	Bgg	214.56	5440.47

43000).

It can be concluded that, for both cases, $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2$ is much lower for any of the normalized entropy methodologies, when compared to Bagging, as can be seen from any of the Tables $1 - 5$.

Comparing Table 1 and Table 2, same number of groups (n.g.=5) and same support vectors for the parameters ($\boldsymbol{z}_k = [-10, 10]$) were considered. The higher condition number in Table 2 results in a much higher $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2$ for the Bagging procedure, whereas the normalized entropy methodologies behave in the same way as with the much lower condition number, revealing that the presence of collinearity does not seem to compromise the results provided by the normalized entropy aggregation procedures. Since the GME estimator is appropriate in the estimation of ill-posed models, including models with ill-conditioned design matrices, these results are not surprising.

Considering Table 3, the analysis was done changing the number of groups in the aggregation. The Bagging procedure tends to provide better results in terms of lower $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2$, as the number of groups rises. This observation does not come as a surprise due to sampling and inferential statistics theory. The normalized entropy methodologies do not seem to follow this behaviour, as the $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2$ remains approximately constant as the number of groups gets higher. This may be considered an advantage of this aggregation procedure, since there is no need for bigger data sets (and consequent higher computational burden) in order to have comparable results in terms of precision.

Finally, Tables 4 and 5 refer to the effect of changing the amplitude of the support vectors, $\boldsymbol{z}_k$. It can be seen that as the support vector $\boldsymbol{z}_k$ changes from $[-10, 10]$, in Table 1, to $[-100, 100]$, in Table 4, all aggregation procedures provide worse results in terms of $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2$. Widening the amplitude of the support vectors results in a less informative probability distribution for the parameters, which should lead to a smaller departure from total uncertainty as compared to the situation where the support vectors are less wide. It is expected, then, that the normalized entropy methodologies provide better results when the amplitude of the support vectors are smaller. The results of the simulation study are in agreement with this interpretation. Nevertheless, when the same analysis is done considering a matrix of explanatory variables $\boldsymbol{X}$, with higher condition number, as presented in Table 5, even though the normalized entropy methodologies provide worse results, as already discussed, the Bagging procedure provides even worse results: while $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2$ changes from 4.25 to 15.59 for the information index weighted average of the GME estimates (solution NE2), the corresponding change for the Bagging procedure is from 1432.59 (which is already a very poor value concerning the precision of the estimates) to 5440.47.

5 Concluding Remarks

The idea of an aggregation procedure based on normalized entropy is promising as it is clear from the simulation study that this approach provides very satisfactory solutions. The normalized entropy methodologies, in particular, the aggregation procedure based on the weighting of the groups by the information index, always results in a $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2$ much lower than the one obtained with Bagging. This discrepancy tends to aggravate in the presence of high collinearity, as that is the case when the explanatory variables matrices, $\boldsymbol{X}$, have high condition numbers. On the other hand, the use of more groups in the aggregation scheme does not seem to improve the overall quality of the estimates obtained through the normalized entropy methodologies, what turns out to be an advantage towards this procedure. These observations suggest that a further and thorough simulation analysis with different error structures or severe inhomogeneities may reveal substantial differences between normalized entropy aggregation schemes and Bagging, eventually penalizing the second. These analysis will be conducted in future work, along with investigation of other scenarios, such as the detection of zero coefficients, non-normal regressors and other violations of the classical framework. Also, the comparison with Magging is a very important analysis that remains to be explored.

Acknowledgments. This research was supported by the Portuguese national funding agency for science, research and technology (FCT), within the Center for Research and Development in Mathematics and Applications (CIDMA), project UID/MAT/04106/2019.

References

1. Golan, A.: On the State of Art of Info-Metrics. In: Uncertainty Analysis in Econometrics with Applications. Huynh, V.N., Kreinovich, V., Sriboonchitta, S., Suriya, K. (eds.), pp. 3–15. Springer-Verlag, Berlin (2013)
2. Golan, A.: Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information. Oxford University Press, New York (2018)
3. Golan, A.: On the Foundations and Philosophy of Info-Metrics. In: Cooper, S.B., Dawar, A., Lowe, B.L. (eds.) CiE2012. LNCS, vol. 7318, pp. 238–245. Springer-Verlag, Heidelberg (2012).
4. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. 106, 620–630 (1957)
5. Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. 108, 171–190 (1957)
6. Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics - Robust Estimation with Limited Data. John Wiley & Sons, Chichester (1996)
7. Mittelhammer, R., Cardell, N. S., Marsh, T. L.: The Data-Constrained Generalized Maximum Entropy Estimator of the GLM: Asymptotic Theory and Inference. Entropy 15, 1756-1775 (2013)
8. Bühlmann, P., Meinshausen, N.: Magging: Maximin Aggregation for Inhomogeneous Large-Scale Data. In: Proceedings of the IEEE 104 (1): Big Data: Theoretical Aspects, pp. 126–135, IEEE Press, New York (2016)
9. Breiman, L.: Bagging Predictors. Mach. Learn. 24, 123–140 (1996)
10. Wolpert, D.: Stacked Generalization. Neural Netw. 5, 241–259 (1992)
11. Breiman, L.: Stacked Regressions. Mach. Learn. 24, 49–64 (1996b)
12. Belsley, D. A., Kuh, E., Welsch, R. E.: Regression Diagnostics - Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, Hoboken, New Jersey (2004)