Normalized entropy: a comparison with traditional techniques in variable selection[1]

Pedro Macedo,[a] Maria Conceição Costa,[b] and João Pedro Cruz[c]

CIDMA – Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

[a] *Corresponding author: pmacedo@ua.pt*
[b] *Electronic mail: lopescosta@ua.pt*
[c] *Electronic mail: pedrocruz@ua.pt*

Abstract. A variable selection procedure in regression analysis using a normalized entropy measure was firstly proposed in 1996, by Amos Golan, George Judge and Douglas Miller, in the book Maximum Entropy Econometrics – Robust Estimation with Limited Data. To the best of the authors' knowledge, the idea has not been explored in the literature since then, despite many noteworthy advantages that have been pointed out by Amos Golan and coauthors, such as: it is simple to perform, even for a large number of variables (useful in some big data problems); it allows the use of non-sample information (easily incorporated in the optimization structure); and it can be implemented for ill-posed models (frequently observed in real-world problems). Following a recent work that illustrates how normalized entropy can represent a promising approach to identify pure noise models, this paper revises the procedure of normalized entropy, proposes some improvements, and illustrates its performance when compared with some well-known traditional techniques in variable selection problems.

GENERALIZED MAXIMUM ENTROPY AND NORMALIZED ENTROPY

Golan et al. [1] generalized the maximum entropy formalism of Jaynes [2, 3] to linear inverse problems with noise expressed by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ 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 a $(N \times 1)$ vector of random disturbances (errors), usually assumed to have a conditional expected value of zero and representing spherical disturbances.

Assuming that both the unknown parameters and the error terms may be bounded a priori, the linear model in (1) can be represented as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} + \boldsymbol{V}\boldsymbol{w}, \tag{2}$$

where $\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{p}$, with $\boldsymbol{Z}$ a $(K \times KM)$ matrix of support spaces and $\boldsymbol{p}$ a $(KM \times 1)$ vector of unknown probabilities, and $\boldsymbol{e} = \boldsymbol{V}\boldsymbol{w}$, with $\boldsymbol{V}$ a $(N \times NJ)$ matrix of support spaces and $\boldsymbol{w}$ a $(NJ \times 1)$ vector of unknown probabilities. Note that each β_k, $k = 1, 2, \ldots, K$, and each e_n, $n = 1, 2, \ldots, N$, 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. A detailed formulation can be found in Golan [4], Chapter 13.

For a linear regression model expressed by (1), the generalized maximum entropy (GME) estimator is given by

$$\underset{\boldsymbol{p},\boldsymbol{w}}{\operatorname{argmax}} \left\{ -\boldsymbol{p}' \ln \boldsymbol{p} - \boldsymbol{w}' \ln \boldsymbol{w} \right\}, \tag{3}$$

subject to the model constraints, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} + \boldsymbol{V}\boldsymbol{w}$, and the additivity constraints for $\boldsymbol{p}$ and $\boldsymbol{w}$, $\mathbf{1}_K = (\boldsymbol{I}_K \otimes \mathbf{1}'_M)\boldsymbol{p}$ and $\mathbf{1}_N = (\boldsymbol{I}_N \otimes \mathbf{1}'_J)\boldsymbol{w}$, where $\otimes$ represents the Kronecker product. The GME estimator generates the optimal probability vectors $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{w}}$ that can be used to form point estimates of the unknown parameters and the unknown errors through the reparameterizations defined previously. Additional details and properties can be found, for example, in Golan [4] and Mittelhammer et al. [5].

[1] **Final version of this paper is published in AIP Conference Proceedings 2425, 190002 (2022);** `https://doi.org/10.1063/5.0081504`.

Golan et al. [1] defined normalized entropy to measure the information content in a particular model using the GME estimator. For example, normalized entropy for the signal, $\boldsymbol{X}\boldsymbol{\beta}$, is given by

$$S(\widehat{\boldsymbol{p}}) = \frac{-\widehat{\boldsymbol{p}}' \ln \widehat{\boldsymbol{p}}}{K \ln M}. \tag{4}$$

Naturally, $S(\widehat{\boldsymbol{p}}) \in [0,1]$, where $S(\widehat{\boldsymbol{p}}) = 1$ indicates perfect uncertainty and $S(\widehat{\boldsymbol{p}}) = 0$ indicates no uncertainty. (Note that only the GME estimator will be considered here; see Macedo [6] for a comparison between the use of normalized entropy with GME and generalized cross entropy (GCE) estimator.)

Concerning variable selection, it is interesting to note that if all the $\boldsymbol{z}_k$ in $\boldsymbol{Z}$ are defined uniformly and symmetrically around zero, then $S(\widehat{\boldsymbol{p}}_k) \approx 1$ implies $\beta_k \approx 0$, because $\widehat{\boldsymbol{p}}_k$ is uniformly distributed in that case. Thus, a variable corresponding to $S(\widehat{\boldsymbol{p}}_k) \approx 1$ has no information content and should be excluded from the model. Golan et al. [1] considered an exclusion criterion of $S(\widehat{\boldsymbol{p}}_k) > 0.99$. The supports for the parameters defined above must be defined as closed and bounded intervals in which each parameter is restricted to belong, but there is empirical evidence that different supports provide different results in terms of variable selection (see simulation study next). Thus, the challenge is to find the optimal supports that allow the correct identification of relevant variables and simultaneously do not produce excessive shrinkage on coefficients' estimates.

A first proposal to define supports for the GME estimator with the purpose of variable selection is inspired on the ridGME procedure [7]. However, due to the purpose of variable selection, the idea presented here is substantially different: each $\boldsymbol{z}_k$ in $\boldsymbol{Z}$ is uniformly and symmetrically defined around zero with limits established by the absolute maximum values of the ridge estimates, such that

$$\boldsymbol{z}_k = \left[-\left\lceil \left| \max\left\{ \widehat{\boldsymbol{\beta}}_{k_{ridge(\eta)}} \right\} \right| \right\rceil, \left\lceil \left| \max\left\{ \widehat{\boldsymbol{\beta}}_{k_{ridge(\eta)}} \right\} \right| \right\rceil \right]. \tag{5}$$

However, two questions naturally arise: why the absolute maximum values of the ridge estimates? Is it possible to achieve better results in terms of variable selection using other kind of prior information? The answer to the first question can be partially supported by the numerical foundations of ridge regression, since the potencial instability of the least squares estimator can be reduced and the shrinkage process may be controlled through a tuning parameter. Nevertheless, this approach clearly suffers from arbitrariness, which is why a second approach is proposed next.

Similar to traditional algorithms of lasso, elastic net, oscar, among others, which provide results for a set of regularization parameters, a set of supports with decreasing amplitudes in the parameter spaces can be defined, starting with any arbitrary large support such that all $S(\widehat{\boldsymbol{p}}_k)$ are approximately one and ending with tiny supports for those variables that were not selected during the process. The optimal step on the decreasing process of the amplitudes can be identified by the solution that corresponds, for example, to the minimum mean squared error obtained by cross-validation. (Illustration of this approach is left for future research due to space limitations.)

SIMULATION STUDY AND EXAMPLE

The simulation study involves 100 observations and 30 standard normal explanatory variables. Ten coefficients are defined by uniform distributions, namely $U(5,15)$ and $U(1,10)$, and the remaining 20 are zero. Errors are defined by normal distributions, namely $e_i \sim N(0,1)$ and $e_i \sim N(0,9)$, $i = 1,2,\ldots,100$. To define an ill-conditioned design matrix, $\boldsymbol{X}$, with a specific condition number value, namely $cn = 100$, the traditional singular value decomposition is obtained and the singular values in $\boldsymbol{S}$, a diagonal matrix with the same dimension of $\boldsymbol{X}$, are modified such that $cond(\boldsymbol{X}) = cond(\boldsymbol{USV'}) = cn$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are square unitary matrices, and $cond(\boldsymbol{X'X}) = cn^2$. It is important to note that other scenarios were tested and the results presented here correspond to the worst ones, i.e. the scenarios with a lower percentage of identification.

Table I and Table II summarize the results with the percentages of trials where the corresponding variable is included in the model, considering three inclusion criteria of $S(\widehat{\boldsymbol{p}}_k)$, in 500 replications. Regardless the scenario considered with the GME estimator, an important conclusion is that variable selection is not possible by using wide bounds. With a moderate support, such as $[-100,100]$, some correct identifications begin to emerge, but the higher percentages of correct identification appear when is used information from the ridge trace. Probably more important than a criterion for inclusion is a graphical representation of the results. For example, considering the case with lower percentage of identification of the ridge-based approach, the boxplot in Fig. 1 shows the identification of the ten relevant variables in the 500 simulated models.

TABLE I. Percentages of inclusion, [min %, max %], in Model $U(5,15)$, with $cond(\boldsymbol{X}) = 100$.

		$[-500,500]$ $e_i \sim N(0,1)$	$[-100,100]$ $e_i \sim N(0,1)$	ridge-based $e_i \sim N(0,1)$	$[-500,500]$ $e_i \sim N(0,9)$	$[-100,100]$ $e_i \sim N(0,9)$	ridge-based $e_i \sim N(0,9)$
Relevant	$S(\widehat{\boldsymbol{p}}_k) \leq 0.99$	[0.0%, 0.0%]	[14.8%, 20.8%]	[95.4%, 98.4%]	[0.0%, 0.0%]	[19.2%, 24.4%]	[80.6%, 85.6%]
variables	$S(\widehat{\boldsymbol{p}}_k) \leq 0.98$	[0.0%, 0.0%]	[0.0%, 0.0%]	[89.6%, 94.2%]	[0.0%, 0.0%]	[1.2%, 2.4%]	[69.6%, 74.2%]
	$S(\widehat{\boldsymbol{p}}_k) \leq 0.97$	[0.0%, 0.0%]	[0.0%, 0.0%]	[83.4%, 89.2%]	[0.0%, 0.0%]	[0.0%, 0.0%]	[60.2%, 65.2%]
Extraneous	$S(\widehat{\boldsymbol{p}}_k) \leq 0.99$	[0.0%, 0.0%]	[0.0%, 0.0%]	[0.8%, 2.4%]	[0.0%, 0.2%]	[0.0%, 0.2%]	[7.0%, 10.0%]
variables	$S(\widehat{\boldsymbol{p}}_k) \leq 0.98$	[0.0%, 0.0%]	[0.0%, 0.0%]	[0.0%, 0.6%]	[0.0%, 0.0%]	[0.0%, 0.0%]	[1.4%, 3.8%]
	$S(\widehat{\boldsymbol{p}}_k) \leq 0.97$	[0.0%, 0.0%]	[0.0%, 0.0%]	[0.0%, 0.4%]	[0.0%, 0.0%]	[0.0%, 0.0%]	[0.2%, 1.8%]

TABLE II. Percentages of inclusion, [min %, max %], in Model $U(1,10)$, with $cond(\boldsymbol{X}) = 100$.

		$[-500,500]$ $e_i \sim N(0,1)$	$[-100,100]$ $e_i \sim N(0,1)$	ridge-based $e_i \sim N(0,1)$	$[-500,500]$ $e_i \sim N(0,9)$	$[-100,100]$ $e_i \sim N(0,9)$	ridge-based $e_i \sim N(0,9)$
Relevant	$S(\widehat{\boldsymbol{p}}_k) \leq 0.99$	[0.0%, 0.0%]	[0.0%, 0.4%]	[79.0%, 83.8%]	[0.0%, 0.2%]	[1.4%, 3.4%]	[57.2%, 62.8%]
variables	$S(\widehat{\boldsymbol{p}}_k) \leq 0.98$	[0.0%, 0.0%]	[0.0%, 0.0%]	[70.8%, 76.2%]	[0.0%, 0.0%]	[0.0%, 0.2%]	[44.2%, 50.6%]
	$S(\widehat{\boldsymbol{p}}_k) \leq 0.97$	[0.0%, 0.0%]	[0.0%, 0.0%]	[63.8%, 68.8%]	[0.0%, 0.0%]	[0.0%, 0.0%]	[36.4%, 42.6%]
Extraneous	$S(\widehat{\boldsymbol{p}}_k) \leq 0.99$	[0.0%, 0.0%]	[0.0%, 0.0%]	[1.8%, 3.8%]	[0.0%, 0.4%]	[0.0%, 0.0%]	[9.8%, 13.8%]
variables	$S(\widehat{\boldsymbol{p}}_k) \leq 0.98$	[0.0%, 0.0%]	[0.0%, 0.0%]	[0.0%, 1.2%]	[0.0%, 0.2%]	[0.0%, 0.0%]	[3.4%, 6.6%]
	$S(\widehat{\boldsymbol{p}}_k) \leq 0.97$	[0.0%, 0.0%]	[0.0%, 0.0%]	[0.0%, 0.4%]	[0.0%, 0.0%]	[0.0%, 0.0%]	[1.0%, 3.8%]

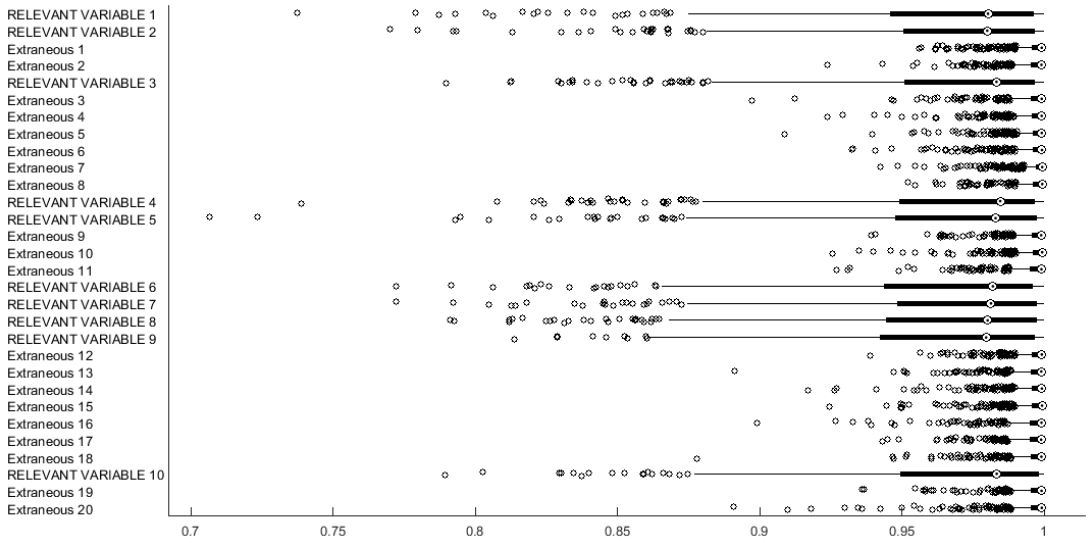

FIGURE 1. Normalized entropy; Model $U(1,10)$, with $cond(\boldsymbol{X}) = 100$, $e_i \sim N(0,9)$.

In the well-known prostate cancer model, the response variable is the level of prostate specific antigen and the eight explanatory variables are cancer volume, prostate weight, age, benign prostatic hyperplasia amount, seminal vesicle invasion, capsular penetration, Gleason score, and percentage Gleason scores 4 or 5. Additional details on this model can be found in the original work of Stamey et al. [8], or in Tibshirani [9], Hastie et al. [10] and Wakefield [11] that use the same data set to illustrate variable selection techniques.

Results for different stepwise methods, all possible subsets, Bayesian model averaging, lasso, among others are available, for example, in Hastie et al. [10] (p. 63) and Wakefield [11] (p. 187). Naturally, there are differences between methods, but two general results emerge between those that eliminate variables: (1) three variables (cancer volume, seminal vesicle invasion and prostate weight) and the constant were selected by almost all the methods; (2) three variables (capsular penetration, Gleason score and percentage Gleason scores 4 or 5) were never selected. The remaining two variables are additionally selected just by one and two methods, respectively.

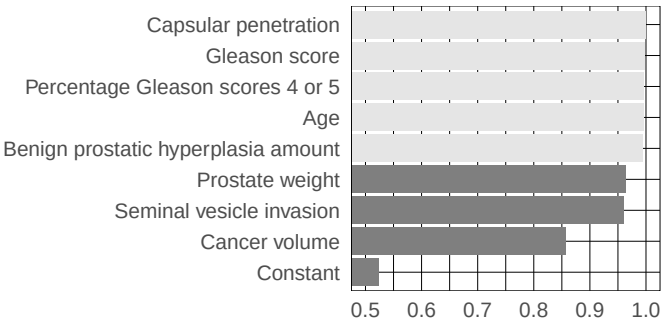

FIGURE 2. Normalized entropy in the prostate cancer model using a ridge-based approach.

Using the supports $[-500, 500]$ and $[-100, 100]$ none of the eight variables is selected. By using the ridge-based approach to select the supports for the GME estimator, and with a simple bar chart as the one in Fig. 2, the information content of each variable is easily checked and ranked. Since variables corresponding to $S(\widehat{\boldsymbol{p}}_k) \approx 1$ have no information content, even without a specific exclusion criterion, the possible relevant and irrelevant variables can be identified. Note that, in comparison with the results of other methods previously mentioned, the two groups of variables are clearly identified in Fig. 2. It is well-known that no single method will work in all kinds of problems and subjective judgement is required in all of them (e.g., significance levels, tuning parameters, cutoff values), but normalized entropy could be a measure of wide application due to its simplicity and intuitive interpretation.

CONCLUSIONS AND FUTURE WORK

A variable selection procedure in regression analysis using normalized entropy is discussed and illustrated in this work. A rule to define a cutoff value in $S(\widehat{\boldsymbol{p}}_k)$ could always be specified by the user following some specific criterion (e.g., interpretation, prediction accuracy, precision), but the information content of each variable may be enough in many real-world scenarios. Following Macedo [6], future research should also be accomplished with the generalized cross entropy estimator, where prior information could be incorporated in the optimization structure with the purpose of variable selection. Maximum entropy has natural connections to artificial intelligence that should be explored in the future. Note that maximum entropy is at the heart of information theory and this, in turn, is in the foundations of computer science. Since artificial intelligence tools just learn from the data, maximum entropy estimation can be important, for example, to provide useful prior information. Comparative studies with specific techniques (e.g., prediction, classification and regression) from recent deep learning tools are needed in future work.

ACKNOWLEDGMENTS

This work is supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT – Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

REFERENCES

1. A. Golan, G. Judge, and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data* (Wiley, Chichester, 1996).
2. E. T. Jaynes, "Information theory and statistical mechanics," Phys. Rev. **106**, 620–630 (1957).
3. E. T. Jaynes, "Information theory and statistical mechanics. II," Phys. Rev. **108**, 171–190 (1957).
4. A. Golan, *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information* (Oxford University Press, New York, 2018).
5. R. Mittelhammer, N. S. Cardell, and T. L. Marsh, "The data-constrained generalized maximum entropy estimator of the GLM: Asymptotic theory and inference," Entropy **15**, 1756–1775 (2013).
6. P. Macedo, "Freedman's paradox: a solution based on normalized entropy," (Springer, Cham, 2020).
7. P. Macedo, "Ridge regression and generalized maximum entropy: an improved version of the Ridge-GME parameter estimator," Commun. Stat. - Simul. Comput. **46**, 3527–3539 (2017).
8. T. Stamey, J. Kabalin, J. McNeal, I. Johnstone, F. Freiha, E. Redwine, and N. Yang, "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients," J. Urol. **141**, 1076–1083 (1989).
9. R. Tibshirani, "Regression shrinkage and selection via the lasso," J. Royal Stat. Soc. B **58**, 267–288 (1996).
10. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer, 2009).
11. J. Wakefield, *Bayesian and Frequentist Regression Methods* (Springer, New York, 2013).