Regularization with Maximum Entropy and Quantum Electrodynamics: the MERG(E) estimators[1]

Pedro Macedo[a*] Manuel Scotto[a] Elvira Silva[b]

[a]CIDMA – Center for Research and Development in Mathematics and Applications,
Department of Mathematics, University of Aveiro,
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
[b]CEF.UP, Faculty of Economics, University of Porto,
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
[*]Corresponding author: pmacedo@ua.pt

It is well-known that under fairly conditions linear regression becomes a powerful statistical tool. In practice, however, some of these conditions are usually not satisfied and regression models become ill-posed, implying that the application of traditional estimation methods may lead to non-unique or highly unstable solutions. Addressing this issue, in this paper a new class of maximum entropy estimators suitable for dealing with ill-posed models, namely for the estimation of regression models with small samples sizes affected by collinearity and outliers, is introduced. The performance of the new estimators is illustrated through several simulation studies.

Keywords: collinearity, linear regression, maximum entropy, micronumerosity, outliers, quantum electrodynamics
MSC 2010: 47A52, 62J05, 62J07, 81V10, 94A17

[1] Final version of this paper is published in **Communications in Statistics – Simulation and Computation, 45: 1–15, 2016** (`http://dx.doi.org/10.1080/03610918.2014.957838`).

1 Introduction

The maximum entropy (ME) formalism was first established by Jaynes (1957a,b) based on physics (the Shannon entropy and statistical mechanics) and statistical inference. In a linear pure inverse model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{y}$ denotes a known $(N \times 1)$ vector of observations, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector of unknown parameters, and $\boldsymbol{X}$ is a known $(N \times K)$ matrix, considering $\boldsymbol{\beta} := \boldsymbol{p}$ a vector of probabilities, the ME principle provides the probability distribution for which the current state of knowledge is sufficient to determine the probability assignment.

Definition 1.1. *In a linear pure inverse model, where $\boldsymbol{\beta} := \boldsymbol{p}$ is a vector of probabilities, the ME formalism is given by*

$$\underset{\boldsymbol{p}}{\operatorname{argmax}} \left\{ -\boldsymbol{p}' \ln \boldsymbol{p} \right\}, \tag{1}$$

subject to the model (or data consistency) constraint, $\boldsymbol{X}\boldsymbol{p} = \boldsymbol{y}$, and the additivity (or normalization) constraint, $\boldsymbol{1}'\boldsymbol{p} = 1$, where $\boldsymbol{1}$ is a $(K \times 1)$ vector of ones, and $\boldsymbol{p} > \boldsymbol{0}$ is a $(K \times 1)$ vector of probabilities.

The ME principle provides a tool to make the best prediction (i.e., the one that is the most strongly indicated) from the available information. Provided that the entropy function in (1) is maximized without model constraint a solution from a uniform distribution is obtained. In this case, it is worth noticing that the ME principle can be seen as an extension of Bernoulli's principle of insufficient reason; e.g., Jaynes (1957a). The ME principle is often used for solving ill-posed problems in physics (Golan and Dose, 2001), biology (Galleani and Garello, 2010), communication engineering (Vila et al., 2011), statistics (Park and Bera, 2009) and economics (Dionísio et al., 2008).

Since statistical data are frequently limited and affected by collinearity implying that the associated statistical models may be ill-posed, Golan et al. (1996) generalized the ME formalism specified in Definition 1.1 to linear inverse problems with noise, expressed by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{2}$$

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 $\boldsymbol{u}$ is a $(N \times 1)$ vector of random disturbances (errors), usually assumed to have a conditional expected value of zero and representing spherical disturbances. The generalized maximum entropy (GME) estimator is defined below using matrix form; e.g., Golan et al. (1996).

Definition 1.2. *For the linear regression model in (2) the GME estimator is given by*

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

subject to the model constraint

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} + \boldsymbol{V}\boldsymbol{w}, \tag{4}$$

and the additivity constraints for $\boldsymbol{p}$ *and* $\boldsymbol{w}$*, respectively,*

$$\begin{aligned} \mathbf{1}_K &= (\boldsymbol{I}_K \otimes \mathbf{1}'_M)\boldsymbol{p}, \\ \mathbf{1}_N &= (\boldsymbol{I}_N \otimes \mathbf{1}'_J)\boldsymbol{w}, \end{aligned} \tag{5}$$

where $\otimes$ *represents the Kronecker product,* $\mathbf{1}$ *is a column vector of ones with a specific dimension,* $\boldsymbol{I}$ *is an identity matrix with a specific dimension,* $\boldsymbol{Z}$ *and* $\boldsymbol{V}$ *are the matrices of supports, and* $\boldsymbol{p} > \mathbf{0}$ *and* $\boldsymbol{w} > \mathbf{0}$ *are probability vectors to be estimated.*

In recent years, the GME estimator greatly contributed to the development of the ME literature. In view of the fact that ill-posed real-world problems seem to be the rule rather than the exception in applied mathematics and statistics, the GME estimator has acquired special importance in the toolkit of statistical techniques, by allowing statistical formulations free of restrictive and unnecessary assumptions. In particular, this estimator is widely used in linear regression models in which (a) the design matrix is ill-conditioned (collinearity), (b) the number of unknown parameters exceeds the number of observations (under-determined models), and (c) in regression models with small samples sizes (micronumerosity).

The supports in matrices $\boldsymbol{Z}$ and $\boldsymbol{V}$ are defined as being closed and bounded intervals within which each parameter or error is restricted to lie, implying that researchers need to provide exogenous information which, unfortunately, it is not always available. This is considered the main weakness of the GME estimator and the main reason for which some statisticians reject this estimator; see, for example, Caputo and Paris (2008) for further details. The number of points M and J in the supports is less controversial. Based on the experiments conducted by Golan et al. (1996), $M = 5$ and $J = 3$ are commonly used in the literature, since there is likely no significant improvement in the estimation with more points in supports. Obviously, as the number of points in the supports increases the computational effort required increases as well.

Some applications of the GME estimator can be found in Macedo et al. (2014), Ferreira et al. (2010), Campbell et al. (2008), Tonini and Jongeneel (2008), Campbell and Hill (2006), Lansink et al. (2001), Lence and Miller (1998), Paris and Howitt (1998), among others.

Giving heed to the problem of the definition of support intervals, Paris (2001) developed the maximum entropy Leuven (MEL) estimator based on some ideas borrowed from the theory of light (quantum electrodynamics) of Feynman (1985). Recently, Macedo et al. (2010a) introduced a general class of estimators based on the MEL estimator, information theory and robust regression techniques hereafter denoted by MERG estimators. For completeness and reader's convenience the definition of the MERG estimators is formalized below.

Definition 1.3. *The MERG estimators of $\boldsymbol{\beta}$ in model (2) are given by*

$$\underset{\boldsymbol{p}_\beta, L_\beta, \boldsymbol{r}}{\operatorname{argmin}} \left\{ \sum_{k=1}^{K} H_1(p_{\beta_k}) + H_2(L_\beta) + \sum_{n=1}^{N} \phi(r_n) \right\}, \tag{6}$$

subject to the model constraint,

$$y_n = \sum_{k=1}^{K} x_{nk}\beta_k + r_n, \tag{7}$$

and the two constraints inspired on the theory of light,

$$L_\beta = \sum_{k=1}^{K} \beta_k^2 \quad \text{and} \quad p_{\beta_k} = \frac{\beta_k^2}{L_\beta}, \tag{8}$$

where $0 < p_{\beta_k} < 1$ is the probability of the parameter β_k, $\sum_n \phi(r_n)$ is a function of the regression residuals, and the functions $\sum_k H_1(p_{\beta_k})$ and $H_2(L_\beta)$ are both entropy measures (e.g., Shannon, Rényi or Tsallis entropies).

The MERG estimators (where the MEL is a particular case) are inspired by the theory of light (quantum electrodynamics). The analogy between the theory of light and statistical analysis can be found in Paris (2001, p. 3). This analogy justifies the following assumption.

Assumption 1.1. *As in the theory of light, where the probability of a photomultiplier being hit by a photon reflected from a sheet of glass equals the square of its amplitude, the probability of β_k is given by the square of Δ (i.e., the amplitude or normalized dimensionality), where*

$$\Delta = \frac{\beta_k}{\sqrt{L_\beta}}. \tag{9}$$

The amplitude of a photon is denoted by a vector that summarizes the various ways in which a photon can reach the photomultiplier. Feynman (1985, pp. 17–35) explains this idea with simple experiments to measure the partial reflection of light by a single or two surfaces of glass. By using arrows representing each possible way in which a photon can reach a given photomultiplier, the author illustrates how to define the "final arrow" (a sum vector) whose square represents the probability of reflection.

By considering different specifications for the components in (6) several MERG estimators are obtained; see Table 1 where the superscripts "R" and "T" denote Rényi and Tsallis' entropies, respectively, and α is the order of the entropy measure. Table 1 does not contain $H_2(L_\beta)$ since this function is defined by the Shannon entropy measure for all MERG estimators. Note that MERG1 corresponds to the MEL estimator.

The MERG estimators represent a non-standard approach to the collinearity problem in linear regression models although they may be also regarded as belonging to the class of regularization methods. Thus, it is interesting to compare the MERG estimators with some other traditional regularization methods that are related to (or make use of) maximum entropy; e.g., Gamboa and Gassiat (1997) and Donoho et al. (1992). For example, the estimator in Donoho et al. (1992), presented in Definition 1.4 below, exhibits some similarities with MERG estimators and in particular with the MEL estimator.

Table 1: MERG estimators.

	$\sum_k H_1(p_{\beta_k})$	$\sum_n \phi(r_n)$
MERG1	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$\sum_n r_n^2$
MERG2	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$\sum_{n=1}^{[(1-\rho)N]+1} r_{(n:N)}^2$
MERG3	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$\sum_n \|r_n\|$
MERG4	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$med_n\, r_n^2$
MERG1$_\alpha^R$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$\sum_n r_n^2$
MERG1$_\alpha^T$	$\frac{1}{1-\alpha} \left(1 - \sum_k (p_{\beta_k})^\alpha\right)$	
MERG2$_\alpha^R$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$\sum_{n=1}^{[(1-\rho)N]+1} r_{(n:N)}^2$
MERG2$_\alpha^T$	$\frac{1}{1-\alpha} \left(1 - \sum_k (p_{\beta_k})^\alpha\right)$	
MERG3$_\alpha^R$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$\sum_n \|r_n\|$
MERG3$_\alpha^T$	$\frac{1}{1-\alpha} \left(1 - \sum_k (p_{\beta_k})^\alpha\right)$	
MERG4$_\alpha^R$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$med_n\, r_n^2$
MERG4$_\alpha^T$	$\frac{1}{1-\alpha} \left(1 - \sum_k (p_{\beta_k})^\alpha\right)$	

Definition 1.4. *The ME estimator of $\boldsymbol{\beta}$ in model (2) is given by*

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\lambda \sum_{k=1}^{K} \beta_k \log \beta_k \right\}, \tag{10}$$

where λ is a regularization parameter and β_k must be non-negative for all k.

In addition to the robust regression methods and the different entropy measures used in some MERG estimators, the main diference between the MERG estimators and the ME estimator in Donoho et al. (1992) lies on the probability specification of the parameters inspired by quantum electrodynamics. With this strategy developed by Paris (2001) the MERG estimators are not restricted to problems with $\boldsymbol{\beta} \in \mathbb{R}_+^K$. Although based on Assumption 1.1, the MERG estimators can be considered as a generalization of the ME estimator in Donoho et al. (1992).

Moreover, an interesting comparison can be made between the ridge regression estimator (Hoerl and Kennard, 1970), which falls within the class of regularization methods, the MERG estimators and the ME estimator in Donoho et al. (1992). Note that while Donoho and co-authors' ME estimator just replaces the quadratic penalty $\sum_k \beta_k^2$ in the ridge regression estimator by $\sum_k \beta_k \log \beta_k$, the MERG estimators replace:

- the loss function $\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$ in the ridge regression estimator by other functions of the regression residuals, namely the least absolute deviations (LAD), the least trimmed squares (LTS) and the least median of squares (LMS);

- the penalty function $\sum_k \beta_k \log \beta_k$ with other entropy measures, namely Rényi or Tsallis entropies, and, with Assumption 1.1, they are not confined to problems with $\boldsymbol{\beta} \in \mathbb{R}_+^K$.

Macedo et al. (2010a) argued that MERG estimators rivals (and in some cases outperforms) with some traditional competitors in linear regression models with small samples sizes affected by collinearity and outliers, and recommend more simulation studies to assess the performance of the MERG estimators.

In this paper, several simulation studies are carried out for the MERG estimators. Furthermore, since the ideas from quantum electrodynamics used in the MERG estimators may not be always valid in different regression models, the violation of Assumption 1.1 motivates the extension of the MERG estimators presented in Section 3.

The remainder of the paper is laid out as follows: in Section 2, the performance of the MERG estimators is assessed through several simulation studies. In Section 3, an extension of the MERG estimators is presented. Finally, some conclusions are given in Section 4.

2 MERG estimators: simulation studies

In this section, the performance of the MERG estimators is compared with other possible competitors. The set of possible competitors includes the ordinary least squares (OLS), LTS, LAD, LMS, GME, the ridge regression, the iteratively reweighted least squares (IRLS), the Liu-type (Liu, 2003) and the RR-MM estimator (Maronna, 2011). For the sake of completeness, the latter estimators are defined next.

Definition 2.1. *The IRLS estimator of $\boldsymbol{\beta}$ in model (2) is given by*

$$\widehat{\boldsymbol{\beta}}_{IRLS}^{(i+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(i)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(i)}\boldsymbol{y}, \tag{11}$$

where $\boldsymbol{W}^{(i)}$ is a $(N \times N)$ diagonal matrix of weights (of the residuals) in the ith iteration.

The weights assigned to the residuals at each iteration are calculated by applying robust criterion functions (Tukey's biweight, Andrews' wave, Huber, among others) to the residuals from the previous iteration; see e.g., Maronna et al. (2006).

Definition 2.2. *The Liu-type estimator of $\boldsymbol{\beta}$ in model (2) is given by*

$$\widehat{\boldsymbol{\beta}}_{\eta,d} = (\boldsymbol{X}'\boldsymbol{X} + \eta\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{y} - d\widehat{\boldsymbol{\beta}}), \tag{12}$$

where $\eta > 0$ and $d \in \mathbb{R}$ are tuning parameters, $\boldsymbol{I}$ is a $(K \times K)$ identity matrix, and $\widehat{\boldsymbol{\beta}}$ is any estimator of $\boldsymbol{\beta}$.

Different choices of η and d are discussed in Liu (2003, pp. 1013–1014).

Definition 2.3. *The RR-MM estimator of $\boldsymbol{\beta}$ in model (2), that combines ridge regression and MM estimation, is given by*

$$\widehat{\boldsymbol{\beta}}_{RR-MM} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \widehat{\sigma}_{ini}^2 \sum_{n=1}^{N} \rho\left(\frac{r_n(\boldsymbol{\beta})}{\widehat{\sigma}_{ini}}\right) + \eta\|\boldsymbol{\beta}_1\|^2 \right\}, \tag{13}$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$, $\widehat{\sigma}_{ini}$ is an M-scale estimate, ρ is the Tukey's biweight ρ-function and η is a penalty parameter (ridge parameter).

It is worth to mention here that the RR-MM estimator is one of the most recent, complete and powerful estimators in the literature for the estimation of regression models affected by collinearity and outliers. Note that if $\widehat{\sigma}_{ini}^2 \sum_n \rho\left(r_n(\boldsymbol{\beta})/\widehat{\sigma}_{ini}\right)$ is replaced by $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ in Definition 2.3 then the RR-MM estimator reduces to the ridge regression estimator. Maronna (2011, pp. 48–49) discusses the good performance of the RR-MM estimator as well as the drawbacks of other competitors namely the estimators proposed by Simpson and Montgomery (1996) and Silvapulle (1991).

In the first simulation study, a pseudo-random number generator is used to define a (40×3) matrix $\boldsymbol{X}$ from a normal distribution with zero mean and unit standard deviation. Using singular value decomposition, the singular values of $\boldsymbol{X}$ in the diagonal matrix obtained from the decomposition are changed to define a matrix $\boldsymbol{X}_1$ with any desired condition number specified *a priori*. In this experiment, $\text{cond}_2\, \boldsymbol{X}_1 = 500$, where cond_2 represents the 2-norm condition number. Finally, a column of ones is added to $\boldsymbol{X}_1$ to define a (40×4) matrix $\boldsymbol{X}_2$, whose $\text{cond}_2\, \boldsymbol{X}_2 \approx 1600$.

The model is defined as $\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{u}$, where $\boldsymbol{\beta} = (0.7, 0.1, -0.8, 0.5)'$ and $\boldsymbol{u}$ is a vector of $N(0,1)$ errors added to form the vectors of noisy observations $\boldsymbol{y}$ in each Monte Carlo trial. To create a small proportion of regression outliers, the first two elements in the second column of $\boldsymbol{X}_2$ are replaced with pseudo-random values drawn from a uniform distribution $U(10, 15)$. After incorporating the outliers $\text{cond}_2\, \boldsymbol{X}_2 \approx 98$. In this case, outliers diminish the collinearity problem, i.e., the magnitude of the relation between the independent variables is reduced.

In each Monte Carlo trial, the first two elements of $\boldsymbol{y}$ are replaced with pseudo-random values drawn from a uniform distribution $U(10, 15)$. For the set of 1000 trials performed, the mean squared error loss (MSEL), with $\text{SEL} = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2$, is the measure used to evaluate the performance of the LTS (with a trimming proportion $\rho = 0.1$), LAD, LMS, IRLS,[2] OLS, ridge, Liu-type, RR-MM, GME and some MERG estimators (the MERG2 class with $\rho = 0.1$); see Table 2.

The MERG estimators with Tsallis and Rényi's entropies are performed assuming that $\alpha = 4$. The GME estimators are performed with two different supports for the parameters, $[-10, 10]$ and $[-5, 5]$, with $M = 5$. The supports for the errors are defined by the 3σ rule with $J = 3$. The ridge estimators are performed with the Ridge-GME (Macedo et al., 2010b), KM5 (Muniz and Kibria, 2009) and HKB (Hoerl et al., 1975) estimators. The ridge interval for the Ridge-GME estimator is defined as $\eta \in]0, 1]$. The corresponding GME estimator is performed using the support $[-5, 5]$ for the parameters (with $M = 5$) and the 3σ rule to

[2] The weights at each iteration are calculated by applying Tukey's biweight function to the residuals from the previous iteration.

Table 2: MSEL in the simulation study with outliers and collinearity.

	MSEL		MSEL		MSEL
MERG1	2.7461	MERG1_4^R	2.2679	OLS	226.6157
MERG2	1.4641	MERG2_4^R	1.4407	ridge (HKB)	44.3252
MERG3	2.5598	MERG3_4^R	2.0023	ridge (KM5)	3.0146
MERG4	2.3033	MERG4_4^R	1.3161	ridge (Ridge-GME)	3.6274
MERG1_4^T	2.2679	LTS	116.7541	GME $[-10, 10]$	12.2597
MERG2_4^T	1.4407	IRLS	153.8387	GME $[-5, 5]$	4.1983
MERG3_4^T	1.9822	LAD	198.3922	Liu-type	2.1369
MERG4_4^T	1.2866	LMS	105.3665	RR-MM	96.6137

define the support for the errors (with $J = 3$).

Three important conclusions emerge from this simulation study. First, the MERG estimators perform very well and rival with two ridge estimators (using the KM5 and Ridge-GME parameter estimates), the Liu-type estimator[3] and the GME estimator (with the support of smaller amplitude).

Second, outliers can mask the presence of collinearity (in the present analysis they reduce collinearity although the problem still remains) which means that the use of traditional robust estimators without a careful diagnostic of collinearity problems can be misleading. Note that the LTS, IRLS, LAD and LMS estimators perform poorly in this experiment.

Third, the MERG estimators outperform the RR-MM estimator which performs poorly in this simulation study.[4] This is an important result because it highlights the performance of the MERG estimators for the combined collinearity-outliers problem in regression analysis with small samples sizes.

Another simulation study, similar to the previous one, is performed with a sample size of

[3] Other ridge and Liu-type estimators are also considered, but the results (not reported here) are very poor. These estimators depend on some parameters that must be estimated from the sample and the results are sensitive to the quality of these parameter estimates.

[4] All the MSEL values presented for the RR-MM estimator (in Table 2 and others where it is used) are calculated using a 10% upper trimmed average; see Maronna (2011, p. 49). The real values of MSEL are higher.

$N = 30$, $\mathrm{cond}_2\, \boldsymbol{X}_2 \approx 200$ and an outlier contamination of 20% only in $\boldsymbol{y}$. In the LTS estimator and the MERG2 class estimators, $\rho = 0.25$ (the usual default value in statistical software using LTS) is used. The results from this experiment are qualitatively the same as the ones presented in Table 2. The best results (with MSEL less than 200) are achieved by the MERG estimators, with MSEL values ranging between 0.8359 (MERG3_4^T) and 6.8959 (MERG1), whereas the MSEL for the RR-MM estimator is 29.7082. For comparison purposes, the MSEL for the OLS estimator is approximately 14236 in this simulation study.

From the results above it becomes clear that the estimation of regression models affected by outliers and collinearity is a difficult task. The interaction among different proportions of outliers contamination, different types of outliers and different magnitudes in the relations among regressors makes the estimation of regression models a very difficult task. Even the best estimators, such as the RR-MM estimator, suffer from this interaction. Surprisingly, the MERG estimators reveal a high stability in different scenarios. Furthermore, the MERG estimators are very easy to compute and no relevant prior information is needed in order to implement them.

Based on the first simulation study (with $N = 40$) another simulation study is conducted with $\mathrm{cond}_2\, \boldsymbol{X}_2 \approx 5$ and only two outliers in $\boldsymbol{y}$. Table 3 presents the results.

Table 3: MSEL in the simulation study with outliers.

	MSEL		MSEL		MSEL
MERG1	1.6756	MERG3_4^T	0.8525	LTS	3.5814
MERG2	0.9978	MERG4_4^T	1.2930	IRLS	3.1911
MERG3	0.9012	MERG1_4^R	1.7069	LAD	5.1867
MERG4	1.1957	MERG2_4^R	0.9774	LMS	13.7382
MERG1_4^T	1.7069	MERG3_4^R	0.9026	OLS	24.4410
MERG2_4^T	0.9774	MERG4_4^R	1.2723		

Note that the LTS, IRLS and LAD estimators perform well whereas, unexpectedly, the LMS estimator performs poorly. However, this result is not entirely surprising since this estimator is usually not recommended as a stand-alone regression procedure. Finally, the MERG estimators reveal the best performance in this experiment.

Another experiment was carried out with $N = 40$, $\text{cond}_2 \boldsymbol{X}_2 \approx 250$ and no outliers. As expected, the OLS estimator performs poorly in this ill-conditioned model. Results for the MERG estimators and some competitors are presented in Table 4. Although there are no outliers the MERG class of estimators with robust methods can be applied. In this experiment, $\rho = 0.05$ is used in the MERG2 class of estimators.

Table 4: MSEL in the simulation study with collinearity.

	MSEL		MSEL		MSEL
MERG1	0.7801	MERG3$_4^T$	0.7114	OLS	1548.9017
MERG2	0.8324	MERG4$_4^T$	1.0739	ridge (KM5)	0.9870
MERG3	0.7057	MERG1$_4^R$	0.6833	ridge (Ridge-GME)	0.9937
MERG4	1.0816	MERG2$_4^R$	0.7107	GME $[-10, 10]$	1.5372
MERG1$_4^T$	0.6833	MERG3$_4^R$	0.7359	GME $[-5, 5]$	1.1533
MERG2$_4^T$	0.7107	MERG4$_4^R$	1.1054	Liu-type	1.3625

In addition, these simulation studies with collinearity and/or outliers were replicated under different conditions, namely for $N = 20$ and 30, $K = 3, 4$ and 5, and different combinations of β_k ranging from -3 to 3. For these additional experiments, the MERG estimators reveal, in general, a good performance and appear to be an interesting choice in models with small samples sizes affected by outliers and collinearity simultaneously or separately.[5] Regarding the choice of tuning parameters, the usual cross-validation technique is suggested to estimate prediction errors and to compare different estimators. See, for example, Hastie et al. (2009, pp. 241–249).

The MERG estimators avoid the possible subjective exogenous information needed in the GME estimator to define the support intervals for the parameters and errors in linear regression models. However, it is important to mention that the ideas from quantum

[5] Results from different empirical applications are not shown here due to space limitations, but are provided upon request to the authors.

electrodynamics used in the MERG estimators may not be always valid in different regression models. The violation of Assumption 1.1 motivates a possible extension of the MERG estimators which is presented next.

3 An extension of the MERG estimators: the MERGE estimators

In this section, an extension of the MERG estimators is presented. For simplicity in notation this extension is denoted as MERGE estimators. This acronym is the initials of the words *maximum entropy robust regression group extended* and reflects the ambitious objective to *merge* several estimators in one group with high performance in linear regression models with small samples sizes affected by collinearity and outliers.

Why this extension of MERG estimators? First, it is not yet fully understood whether the theory of light in Feynman (1985), used in the MERG estimators, is always valid in different regression models. More research is needed to assess the reasonability of the Assumption 1.1. Second, there are regression models where the supports of the parameters can be defined by the researcher's experience and/or provided by the theory (e.g., in economics to estimate the marginal propensity to consume in a Keynesian consumption function). Third, the use of the cross-entropy formalism and the possibility of imposing parameter inequality restrictions through the parameter support matrix (as in the GME estimator case) are easily handled with this extension. Fourth, the supports for the errors used in the GME estimator are not needed with the MERGE estimators.

Definition 3.1. *The MERGE estimators of $\boldsymbol{\beta}$ in model (2) are given by*

$$\underset{\boldsymbol{p},\boldsymbol{r}}{\operatorname{argmin}} \left\{ (1-\theta) \sum_{k=1}^{K} \sum_{m=1}^{M} H(p_{km}) + \theta \sum_{n=1}^{N} \phi(r_n) \right\}, \tag{14}$$

subject to the model constraint and the additivity constraint

$$\begin{cases} y_n = \displaystyle\sum_{k=1}^{K}\sum_{m=1}^{M} x_{nk} z_{km} p_{km} + r_n \\ \displaystyle\sum_{m=1}^{M} p_{km} = 1, k = 1, 2, \ldots, K \end{cases}, \tag{15}$$

where $p_{km} > 0$, $k = 1, 2, \ldots, K$, $m = 1, 2, \ldots, M$, are probabilities, z_{km} are the supports for the parameters, the function $\sum_k \sum_m H(p_{km})$ is an entropy measure (e.g., Shannon, Rényi or Tsallis entropies) and $\sum_n \phi(r_n)$ is a function of the regression residuals.

In the objective function (14), the parameter $\theta \in (0, 1)$ assigns different weights on the components of the objective function in order to reflect greater prediction or precision reliability in the estimation.[6] This is in contrast with the MERG estimators (Definition 1.3) which implicitly assume equal weights in the components of the objective function. Note that unlike the MERG, Assumption 1.1 is not required in the definition of the MERGE estimators and no supports are needed for the error component in contrast to the GME estimator. It is also important to stress that the penalty function in the MERGE estimators is now composed by an entropy measure with supports for the parameters.

Table 5: MERGE estimators.

	MERGE1	MERGE2	MERGE3	MERGE4
$\sum_{n=1}^{N} \phi(r_n)$	$\sum_{n=1}^{N} r_n^2$	$\sum_{n=1}^{[(1-\rho)N]+1} r_{(n:N)}^2$	$\sum_{n=1}^{N} \lvert r_n \rvert$	$med_n\, r_n^2$

Table 5 presents different MERGE estimators using the OLS, LTS, LAD and LMS estimators assuming $H(p_{km}) = p_{km} \ln p_{km}$. When the Tsallis and Rényi entropies are adopted in the objective function (14), the notation is similar to the one used in the MERG estimators, i.e., MERGEi_α^T or MERGEi_α^R, respectively, considering $i = 1, 2, 3, 4$ and α the order of the entropy measure.

[6] The idea of a weighted GME objective function, which is followed here for the MERGE estimators, is proposed in Wu (2009).

To define the MERGE estimators in a more general framework, the term $\sum_n \phi(r_n)$ in (14) can be generalized by using an S-estimator; see Definition 3.2. The S-estimators are regression, scale and affine equivariant and by a convenient choice of the constants involved their breakdown point can attain 50%. Note also that by allowing different types of dispersion measures, the OLS, LTS, LAD and LMS estimators are S-estimators; e.g., Maronna et al. (2006).

Definition 3.2. *The MERGE estimators of $\boldsymbol{\beta}$ in model (2) are given by*

$$\underset{\boldsymbol{p},s}{\operatorname{argmin}} \left\{ (1-\theta) \sum_{k=1}^{K} \sum_{m=1}^{M} H(p_{km}) + \theta s(r_1(\boldsymbol{\beta}), r_2(\boldsymbol{\beta}), \ldots, r_N(\boldsymbol{\beta})) \right\}, \tag{16}$$

subject to

$$\begin{cases} y_n = \sum_{k=1}^{K} \sum_{m=1}^{M} x_{nk} z_{km} p_{km} + r_n \\ \sum_{m=1}^{M} p_{km} = 1, k = 1, 2, \ldots, K \\ \sum_{n=1}^{N} \rho\left(\frac{r_n}{s}\right) = N\tau \end{cases}, \tag{17}$$

where $p_{km} > 0$, $k = 1, 2, \ldots, K$, $m = 1, 2, \ldots, M$, are probabilities, z_{km} are the supports for the parameters, the function $\sum_k \sum_m H(p_{km})$ is an entropy measure (e.g., Shannon, Rényi or Tsallis entropies), r_n are the residuals, $\rho(\cdot)$ is a function to be selected (e.g., the Tukey's biweight ρ-function) and τ is a consistency constant.

As for the GME or MERG estimators, MERGE estimators have no closed-form solution and the solution must be found numerically. The Lagrangian function and the first-order optimality conditions for the MERGE1 estimator are presented next. The same procedure can be followed for the other MERGE estimators. In matrix form, the Lagrangian function is given by

$$L(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = (1-\theta)\boldsymbol{p}' \ln \boldsymbol{p} + \theta \boldsymbol{r}'\boldsymbol{r} + \boldsymbol{\lambda}'(\boldsymbol{y} - \boldsymbol{XZp} - \boldsymbol{r}) + \boldsymbol{\mu}'(\boldsymbol{1}_K - (\boldsymbol{I}_K \otimes \boldsymbol{1}'_M)\boldsymbol{p}), \tag{18}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are, respectively, a $(N \times 1)$ and a $(K \times 1)$ vectors of Lagrange multipliers on the corresponding constraints. The first-order optimality conditions are

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{p}} = (1-\theta)(\ln \boldsymbol{p} + \mathbf{1}_{KM}) - \boldsymbol{Z}'\boldsymbol{X}'\boldsymbol{\lambda} - (\boldsymbol{I}_K \otimes \mathbf{1}_M)\boldsymbol{\mu} = \mathbf{0}, \tag{19}$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{r}} = 2\theta \boldsymbol{r} - \boldsymbol{\lambda} = \mathbf{0}, \tag{20}$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\lambda}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} - \boldsymbol{r} = \mathbf{0}, \tag{21}$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\mu}} = \mathbf{1}_K - (\boldsymbol{I}_K \otimes \mathbf{1}'_M)\boldsymbol{p} = \mathbf{0}. \tag{22}$$

In real-world empirical applications and also for inference purposes the bootstrap method is recommended to inferring statistical properties over the MERGE estimators since these estimators like the GME and MERG estimators are suitable for regression models with micronumerosity. The bootstrap estimator for the asymptotic covariance matrix can be computed as

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\text{MERGE}}) = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\boldsymbol{\beta}}_{\text{MERGE}}(t) - \widehat{\boldsymbol{\beta}}_{\text{MERGE}})(\widehat{\boldsymbol{\beta}}_{\text{MERGE}}(t) - \widehat{\boldsymbol{\beta}}_{\text{MERGE}})', \tag{23}$$

where $\widehat{\boldsymbol{\beta}}_{\text{MERGE}}$ is the MERGE estimator and $\widehat{\boldsymbol{\beta}}_{\text{MERGE}}(t)$ is the tth MERGE estimate of $\boldsymbol{\beta}$ based on a sample of N observations drawn with replacement from the original sample.

3.1 Simulation study

The following simulation study illustrates the performance of the MERGE estimators in the estimation of linear regression models with small samples sizes affected by outliers and collinearity. The main purpose is to illustrate that the MERGE estimators may outperform the MERG estimators rather than to provide a full comparison between these estimation techniques and other methods. However, results are also presented for the RR-MM estimator, the main competitor of MERG and MERGE estimators (MERG(E) in short) in regression models with collinearity and outliers, and the OLS estimator.

As in the experiments previously showed in Section 2, a pseudo-random number generator is used as well as the singular value decomposition to define matrices with a desired condition number. Different $(N \times K)$ matrices $\boldsymbol{X}$ are generated from a normal distribution with zero mean and unit standard deviation. Changing the singular values obtained from the decomposition, different matrices $\boldsymbol{X}_1$ with $\text{cond}_2\, \boldsymbol{X}_1 = 150$ are defined. To create a proportion δ of outliers, a similar strategy in Golan and Perloff (2002) and Ferretti et al. (1999) is followed, i.e., different models (without intercept) given by $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{u}$ are defined, where $\boldsymbol{y} = \boldsymbol{u}$ is randomly generated from a normal distribution with zero mean and unit standard deviation. For $N\delta$ observations, the elements in $\boldsymbol{y}$ are replaced by the value 6 and the corresponding elements in the first column of $\boldsymbol{X}_1$ by the value 10 (being this the $\boldsymbol{X}_2$ matrix).

This simulation study considers the following possibilities: $N = 10$ and $N = 30$; $K = 3$ and $K = 5$; $\delta = 0.1$ and $\delta = 0.3$. The MERGE estimators are performed taking $\theta = 0.5$, and two different supports for the parameters, namely $[-5, 5]$ and $[-1, 1]$, both with $M = 5$. For the 1000 trials performed, the MSEL is the measure used to evaluate the performance of different estimators. Tables 6 and 7 present the results.

This simulation study reveals that the MERGE estimators may outperform the MERG estimators although this performance seems to depend on the amplitude of the supports. Thus, as a precaution, the MERGE estimators should be used only in cases of fully correct prior information about the parameters (supports of smaller amplitude).[7] The worst results for the MERGE estimators are almost exclusively for those estimators using the OLS and LMS estimators, i.e., the MERGEi, MERGEi_4^R and MERGEi_4^T, for $i = 1, 4$. In contrast, the lower values of MSEL are almost exclusively achieved by MERGE estimators using the LTS and LAD estimators.

Furthermore, the comparison between the MERGE estimators and the RR-MM estima-

[7] The exogenous parameter weighting between signal and noise is $\theta = 0.5$ in this study. Naturally, this parameter can be changed in order to reflect different weights in the components of the objective function. The impact of this choice is left for future research.

Table 6: MSEL for the estimators in the simulation study ($N = 10$).

		$K = 3$		$K = 5$	
		$\mathrm{cond}_2\, \boldsymbol{X}_2$		$\mathrm{cond}_2\, \boldsymbol{X}_2$	
		$(\approx)\, 203$	$(\approx)\, 95$	$(\approx)\, 108$	$(\approx)\, 96$
		$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.3$
OLS		559.1269	45.2705	149.8799	22.5019
RR-MM		154.0145	8.3637	109.0394	14.1674
MERG	1st best	0.8673	0.7271	1.0039	0.7147
	2nd best	0.9130	0.7455	1.0265	0.8303
	3rd best	0.9285	0.8669	1.0463	0.8722
	worst	2.8191	1.1978	1.7521	2.7399
MERGE $[-5, 5]$	1st best	0.1179	0.1232	0.2117	0.2286
	2nd best	0.1250	0.1412	0.2230	0.3057
	3rd best	1.3939	1.4748	2.9431	3.2880
	worst	11.4988	10.7308	16.5803	10.1718
MERGE $[-1, 1]$	1st best	0.1034	0.1472	0.0884	0.2615
	2nd best	0.1417	0.2923	0.2531	0.3354
	3rd best	0.2426	0.4533	0.5441	0.4734
	worst	1.8883	1.8337	3.0394	2.8983

tor[8] depends on the sample size N. For very small samples, such as $N = 10$, almost all the MERGE estimators outperform the RR-MM estimator. However, for $N = 30$ it is only possible to conclude that in general the MERGE estimators rival with the RR-MM estimator.

4 Conclusions

The main weakness of the GME estimator is that support intervals (exogenous information not always available) for the parameters and error vectors are needed. To tackle this problem Paris (2001) developed the MEL estimator based on some ideas from quantum

[8] The MSEL values for the RR-MM estimator are calculated using a 10% upper trimmed average.

Table 7: MSEL for the estimators in the simulation study ($N = 30$).

		$K = 3$		$K = 5$	
		$\operatorname{cond}_2 \boldsymbol{X}_2$		$\operatorname{cond}_2 \boldsymbol{X}_2$	
		$(\approx)\,68$	$(\approx)\,146$	$(\approx)\,255$	$(\approx)\,101$
		$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.3$
OLS		24.6791	29.1554	273.1555	14.1823
RR-MM		1.8419	0.4840	0.6286	0.4145
MERG	1st best	0.7618	0.6332	0.8927	0.8257
	2nd best	0.9432	0.6973	1.0067	0.8649
	3rd best	0.9669	0.8127	1.0410	1.0292
	worst	1.4752	1.8351	2.1807	1.6949
MERGE $[-5, 5]$	1st best	0.1363	0.1316	0.2176	0.2101
	2nd best	0.1522	0.1614	0.2295	0.2541
	3rd best	1.2663	1.4268	2.4975	2.6356
	worst	12.7448	6.9831	19.7499	6.3764
MERGE $[-1, 1]$	1st best	0.1012	0.1533	0.2691	0.2537
	2nd best	0.2090	0.3199	0.2936	0.3271
	3rd best	0.3919	0.3785	0.6495	0.4090
	worst	1.9006	1.5951	2.8381	3.0892

electrodynamics in Feynman (1985), the Shannon entropy measure and the OLS estimator.

Considering the same framework as in the MEL estimator, the MERG estimators are defined through a general expression in which Rényi and Tsallis' entropies can be applied as well as different robust regression estimators. The MERG estimators (which include the MEL estimator as a particular case) have two important features: are easy to compute and no relevant prior information is needed to implement them. In this paper, several simulation studies illustrate an excellent performance of the MERG estimators.

The MERGE estimators introduced in this work are a possible extension of the MERG estimators. The simulation study reveals that the MERGE estimators may outperform the MERG estimators and the RR-MM estimator in linear regression models with small samples sizes affected by collinearity and outliers.

Based on the experiments conducted in this paper (and others not reported here), the MERG2 and MERG3 class of estimators are probably the most adequate choices in the case of no prior information on the parameters. Moreover, the MERGE2 and MERGE3 class of estimators are likely the most proper choices if there is correct prior information on the parameters. However, some questions on the MERG(E) estimators remain open, namely: which entropy measure should be used?, and what should be the order of the entropy measure? It seems that, in some cases, the MERG(E) estimators with the Tsallis and Rényi entropies provide a lower MSEL than the estimators with the Shannon entropy. Further research is necessary in order to identify the most proper estimator in each case.

Acknowledgements

The authors thank the referee for comments that greatly improved the article. This work is supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT - Fundação para a Ciência e a Tecnologia"), within project PEst-OE/MAT/UI4106/2014. Pedro Macedo was also supported by the grant SFRH/BD/40821/2007 from FCT.

References

Campbell, R., Rogers, K., and Rezek, J. (2008). Efficient frontier estimation: a maximum entropy approach. *Journal of Productivity Analysis*, 30(3):213–221.

Campbell, R. C. and Hill, R. C. (2006). Imposing parameter inequality restrictions using the principle of maximum entropy. *Journal of Statistical Computation and Simulation*, 76(11):985–1000.

Caputo, M. R. and Paris, Q. (2008). Comparative statics of the generalized maximum entropy estimator of the general linear model. *European Journal of Operational Research*, 185(1):195–203.

Dionísio, A., Reis, A. H., and Coelho, L. (2008). Utility function estimation: the entropy approach. *Physica A*, 387(15):3862–3867.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society, Series B*, 54(1):41–81.

Ferreira, P., Dionísio, A., and Pires, C. (2010). Adopt the euro? The GME approach. *Journal of Economic Interaction and Coordination*, 5(2):231–247.

Ferretti, N., Kelmansky, D., Yohai, V. J., and Zamar, R. H. (1999). A class of locally and globally robust regression estimates. *Journal of the American Statistical Association*, 94(445):174–188.

Feynman, R. P. (1985). *QED - The Strange Theory of Light and Matter.* Penguin Group, London.

Galleani, L. and Garello, R. (2010). The minimum entropy mapping spectrum of a DNA sequence. *IEEE Transactions on Information Theory*, 56(2):771–783.

Gamboa, F. and Gassiat, E. (1997). Bayesian methods and maximum entropy for ill-posed inverse problems. *The Annals of Statistics*, 25(1):328–350.

Golan, A. and Dose, V. (2001). A generalized information theoretical approach to tomographic reconstruction. *Journal of Physics A*, 34(7):1271–1283.

Golan, A., Judge, G., and Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* John Wiley & Sons, Chichester.

Golan, A. and Perloff, J. M. (2002). Comparison of maximum entropy and higher-order entropy estimators. *Journal of Econometrics*, 107(1-2):195–211.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics - Simulation and Computation*, 4(2):105–123.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review*, 108(2):171–190.

Lansink, A. O., Silva, E., and Stefanou, S. (2001). Inter-firm and intra-firm efficiency measures. *Journal of Productivity Analysis*, 15(3):185–199.

Lence, S. H. and Miller, D. J. (1998). Estimation of multi-output production functions with incomplete data: a generalised maximum entropy approach. *European Review of Agricultural Economics*, 25(2):188–209.

Liu, K. (2003). Using Liu-type estimator to combat collinearity. *Communications in Statistics - Theory and Methods*, 32(5):1009–1020.

Macedo, P., Scotto, M., and Silva, E. (2010a). A general class of estimators for the linear regression model affected by collinearity and outliers. *Communications in Statistics - Simulation and Computation*, 39(5):981–993.

Macedo, P., Scotto, M., and Silva, E. (2010b). On the choice of the ridge parameter: a maximum entropy approach. *Communications in Statistics - Simulation and Computation*, 39(8):1628–1638.

Macedo, P., Silva, E., and Scotto, M. (2014). Technical efficiency with state-contingent production frontiers using maximum entropy estimators. *Journal of Productivity Analysis*, 41(1):131–140.

Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics*, 53(1):44–53.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics - Theory and Methods*. John Wiley & Sons, Chichester.

Muniz, G. and Kibria, B. M. G. (2009). On some ridge regression estimators: an empirical comparisons. *Communications in Statistics - Simulation and Computation*, 38(3):621–630.

Paris, Q. (2001). Multicollinearity and maximum entropy estimators. *Economics Bulletin*, 3(11):1–9.

Paris, Q. and Howitt, R. E. (1998). An analysis of ill-posed production problems using maximum entropy. *American Journal of Agricultural Economics*, 80(1):124–138.

Park, S. Y. and Bera, A. K. (2009). Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230.

Silvapulle, M. J. (1991). Robust ridge regression based on an M-estimator. *Australian Journal of Statistics*, 33(3):319–333.

Simpson, J. R. and Montgomery, D. C. (1996). A biased-robust regression technique for the combined outlier-multicollinearity problem. *Journal of Statistical Computation and Simulation*, 56(1):1–22.

Tonini, A. and Jongeneel, R. (2008). Modelling dairy supply for Hungary and Poland by generalised maximum entropy using prior information. *European Review of Agricultural Economics*, 35(2):219–246.

Vila, M., Bardera, A., Feixas, M., and Sbert, M. (2011). Tsallis mutual information for document classification. *Entropy*, 13(9):1694–1707.

Wu, X. (2009). A weighted generalized maximum entropy estimator with a data-driven weight. *Entropy*, 11:917–930.