

RidGME Estimation and Inference in Ill-Conditioned Models

Pedro Macedo[1], Jorge Cabral[1], Vera Afreixo[1], Francisco Macedo[2], and Mario Angelelli[3](✉)

[1] Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, Aveiro 3810-193, Portugal
{pmacedo,jorgecabral,vera}@ua.pt

[2] Department of Informatics Engineering, Faculty of Science and Technology, University of Coimbra, Pólo II, 3030-290 Coimbra, Portugal
fmacedo@student.uc.pt

[3] Department of Human and Social Sciences, Centre of Applied Mathematics and Physics for Industry (CAMPI), University of Salento, Via Valesio, 73100 Lecce, Italy
mario.angelelli@unisalento.it

Abstract. The selection of statistical methods that can operate even on ill-conditioned problems, providing accurate estimation and supporting general inference tasks, is a central need in several applications in different domains. In this setting, two major approaches are the methods from the bridge estimation family, such as ridge regression, and the entropy-based methods, such as generalized maximum entropy estimation. Regarding these two methods, on the one hand, the appropriate choice of the shrinkage parameter in ridge regression is not always an easy task, and, on the other hand, the choice of supports for the parameters of the model is a usual difficulty with the generalized maximum entropy estimator. The RidGME estimator proposed in this work, using generalized maximum entropy with supports defined with information from the ridge trace, mitigates the weaknesses of both methodologies, revealing itself as an alternative to ridge regression on ill-conditioned models. The results from an empirical application reveal that the RidGME estimator is competitive with the *best* optimal choice of ridge estimators and outperforms the ordinary least squares estimator in a cross-validation root mean squared error sense. Its implementation is illustrated, and computational codes are made available to practitioners.

Keywords: Generalized maximum entropy · Ridge regression · Ridge trace · Shrinkage estimation

1 Introduction

The quality of analyses exploring relationships between multiple variables may be critically affected by redundant information shared by such variables. Specifically, the presence of strong collinearity in regression models makes them ill-conditioned (or ill-posed in a broader sense) [4, pp. 85–98]. In turn, this situation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15887, pp. 300–313, 2026.
https://doi.org/10.1007/978-3-031-97589-9_21

may invalidate or prevent the use of traditional estimation methods, as they may lead to highly unstable estimates in this circumstance.

Ridge regression proposed by Hoerl and Kennard [9] is an estimation procedure to handle collinearity without removing variables from the regression model; although biased, the ridge estimator can be substantially more precise than unbiased estimators under collinearity problems. Regarding the usual concern with the use of biased estimators, note that the unknown parameters in a regression model are assumed finite to be well-defined. Such parameters constitute the components of a vector $\boldsymbol{\beta}$ and, under the finiteness hypotesis, we can assume an upper bound on $\boldsymbol{\beta}'\boldsymbol{\beta}$ and apply the ordinary least squares (OLS) estimator subject to this restriction. This *constrained* optimization problem returns the general form of the ridge regression [21, p. 470]. Although the standardization of the variables is usually recommended to practitioners, this decision should always be evaluated on a case-by-case basis, depending on the data (e.g., measurement scales), the theoretical model (e.g., the model includes an intercept or not; whether the intercept is involved in the collinearity issue or not), and the purpose of the regularization process (e.g., prediction, precision, interpretation).

The selection of the shrinkage (tuning or ridge) parameter that controls the additional constraint on the coefficient vector's norm remains a challenge in ridge regression. In addition to other possibilities (e.g., formal methods, cross-validation), one straightforward approach is based on the analysis of the ridge trace – a plot with the path of each estimate as a function of several possible values for the shrinkage parameter –, i.e., the analysis of the paths of the estimates [8,9,16,25]. Then a value for the shrinkage parameter is chosen in the region where the estimates "become more stable," but the formalization of this notion of stability is usually considered somewhat arbitrary. It is commonly accepted that this choice should be made in the interval between zero and one for the shrinkage parameter, providing a compromise between the reduction of variance and the increase of bias [16]. This means that, assuming this interval (or even a larger one), each parameter space is clearly established with the analysis of the ridge trace.

On the other hand, a different approach to deal with collinearity (and, in general, with ill-posed models) was proposed by Golan et al. [7] through the *generalized maximum entropy* (GME) estimator. This approach extends the maximum entropy (ME) principle established by Jaynes [10,11] based on Shannon entropy [23]. The concept of entropy appeared in the 19th century as a central notion at the foundations of thermodynamics, later connected with a probabilistic formulation in statistical physics [12]. The concept of entropy acquires a new meaning beyond its original purpose in 1948, with Claude Shannon [23], as a measure of information (or uncertainty) in information theory. The GME and other information-theoretic estimation techniques are included in a research area at the intersection of statistics, computer science, and decision theory, that is, Info-Metrics [6]. Some concerns regarding the use of the GME estimator are usually related to the specification of the support spaces for the parameters of the regression model, an issue that remains an open question from a formal statisti-

cal perspective [5–7,19], but also has practical interpretations, e.g., in relation to inequivalent knowledge representations as a source of epistemic uncertainty [3].

The Ridge-GME shrinkage parameter estimator [13,15] is a procedure that combines the ridge and the GME estimators, in order to choose an adequate shrinkage parameter. The ridge trace provides information for the selection of the supports for the parameters in the regression model, which is the main difficulty with the GME estimator; on the other hand, the use of the GME estimator tries to solve the arbitrariness in the selection of the shrinkage parameter based on the visual analysis of the ridge trace. This work moves the initial purpose of the Ridge-GME shrinkage parameter estimator from choosing the shrinkage parameter for ridge regression to the objective of being itself as an alternative to ridge regression estimation, including confidence intervals for inference purposes.

The remainder of the paper is laid out as follows: in Sect. 2, a brief overview of OLS, ridge, GME and Ridge-GME estimators is presented. The new RidGME estimator is discussed in Sect. 3. An illustration is accomplished in Sect. 4. Finally, some concluding remarks are given in Sect. 5.

2 Overview of the Estimators

The linear regression model is usually defined by

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

where $\boldsymbol{y}$ denotes a $(N \times 1)$ vector of noisy observations, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector of unknown parameters to be estimated, $\boldsymbol{X}$ is a known $(N \times K)$ matrix of explanatory variables, and $\boldsymbol{u}$ is a $(N \times 1)$ vector of random errors, usually assumed to have a conditional expected value of 0 and representing spherical disturbances, i.e., $E[\boldsymbol{u}|\boldsymbol{X}] = \boldsymbol{0}$ and $E[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}] = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $(N \times N)$ identity matrix and σ^2 is the error variance. Different instances of the generating model (1) can be considered, for example, including the intercept – if this aligns with the theoretical framework and the nature of data –, by extending both the parameter vector $\overline{\boldsymbol{\beta}} := (\beta_0, \boldsymbol{\beta})$ and the matrix of explanatory variables $\overline{\boldsymbol{X}} := (\boldsymbol{1}_{N\times 1}\, \boldsymbol{X})$, where $\boldsymbol{1}$ is the $(N \times 1)$-dimensional vector with all components equal to 1.

From these premises, we can formalize the different parameter estimators $\widehat{\boldsymbol{\beta}}$ considered in this work:

Definition 1. *The ordinary least squares estimator, $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, of $\boldsymbol{\beta}$ in model (1) is given by*

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2}$$

As mentioned in the Introduction section, the ridge estimator, by adding a small shrinkage parameter to the diagonal of $\boldsymbol{X}'\boldsymbol{X}$, introduces some bias, although it reduces the variance of the OLS estimator.

Definition 2. *The ridge estimator, $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$, of $\boldsymbol{\beta}$ in model (1) is given by*

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{3}$$

where $\lambda \geq 0$ *denotes the shrinkage parameter and* $\boldsymbol{I}$ *is the* $(K \times K)$ *identity matrix.*

Among the advantages of the OLS and ridge estimators, we briefly discuss their closed forms (2) and (3). The real and symmetric matrices $\boldsymbol{X}'\boldsymbol{X}$ and $\lambda \cdot \boldsymbol{I}$ can be simultaneously diagonalized, and the eigenvalues of their sum $\boldsymbol{X}'\boldsymbol{X} + \lambda \cdot \boldsymbol{I}$ have the form $\Lambda_k + \lambda$, where Λ_k denotes the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$, $k \in \{1, \ldots, K\}$. Therefore, in cases where $\Lambda_k \approx 0$ for some $k \in \{1, \ldots, K\}$, i.e., when $\boldsymbol{X}'\boldsymbol{X}$ is (nearly) singular, we can choose a *generic* value of the shrinkage parameter λ to get a non-singular sum $\boldsymbol{X}'\boldsymbol{X} + \lambda \cdot \boldsymbol{I}$. This means that at least one of the two estimators (2) or (3) is well-defined; furthermore, when they both exist, it follows from Definition 2 that the ridge estimator approaches the OLS estimator when $\lambda \to 0$, while it approaches the zero vector when $\lambda \to \infty$.

On the other hand, we stress once again that, when $\boldsymbol{X}'\boldsymbol{X}$ is (nearly) singular, ridge regression introduces arbitrariness in the estimates due to the freedom in choosing λ. The variance of $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$ decreases as λ increases, but the bias increases with λ. Since the range of values for which the ridge estimator is superior to the OLS estimator depends on the unknown $\boldsymbol{\beta}$ and σ^2, the challenge is to select an estimate of λ such that the ridge estimator has a smaller mean squared error (MSE) than the OLS estimator. As we will see, this arbitrariness can be handled efficiently through maximum entropy methods, specifically, the GME estimator introduced in the following.

Definition 3. *The generalized maximum entropy estimator,* $\widehat{\boldsymbol{\beta}}_{\text{GME}}$*, of* $\boldsymbol{\beta}$ *in model (1) is given by*

$$\widehat{\boldsymbol{\beta}}_{\text{GME}} = \boldsymbol{Z}\widehat{\boldsymbol{p}}, \tag{4}$$

where $\widehat{\boldsymbol{p}}$ *is obtained from the optimization problem described by*

$$\underset{\boldsymbol{p},\boldsymbol{w}}{\operatorname{argmax}} \left\{ -\boldsymbol{p}' \ln \boldsymbol{p} - \boldsymbol{w}' \ln \boldsymbol{w} \right\}, \tag{5}$$

subject to the model constraints,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{p} + \boldsymbol{V}\boldsymbol{w}, \tag{6}$$

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

$$\mathbf{1}_K = (\boldsymbol{I}_K \otimes \mathbf{1}'_M)\boldsymbol{p}, \quad \mathbf{1}_N = (\boldsymbol{I}_N \otimes \mathbf{1}'_J)\boldsymbol{w}, \tag{7}$$

where $\otimes$ *represents the Kronecker product,* $\boldsymbol{Z}$ *and* $\boldsymbol{V}$ *are the matrices of supports, and* $\boldsymbol{p} > \mathbf{0}$ *and* $\boldsymbol{w} > \mathbf{0}$ *(componentwise) are probability vectors to be estimated.*

The algebraic structure for $\boldsymbol{\beta}$ in Definition 3 is given by

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

where $\boldsymbol{Z}$ is the $(K \times KM)$ matrix of supports for $\boldsymbol{\beta}$ and $\boldsymbol{p}$ is the $(KM \times 1)$ vector of unknown probabilities; here, $\boldsymbol{Z}$ is a block-diagonal matrix where each row corresponds to a parameter in the model, and the associated block on the principal diagonal comprises a finite number (M) of points as representatives of a closed and bounded interval in which the parameter is expected to be contained. Regarding the error terms, they are modelled as follows:

$$\boldsymbol{u} = \boldsymbol{V}\boldsymbol{w} = \begin{bmatrix} \boldsymbol{v}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{v}_2' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{v}_N' \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix}, \tag{9}$$

where J is the number of points in each support of $\boldsymbol{V}$, $\boldsymbol{V}$ is the $(N \times NJ)$ matrix of supports for $\boldsymbol{u}$ (representing a closed and bounded interval in which each error is expected to be contained), and $\boldsymbol{w}$ is the $(NJ \times 1)$ vector of unknown probabilities. The number of points, M and J, are usually between three and five because there is no significant improvement in the estimation with more points in the supports [7].

The supports in matrices $\boldsymbol{Z}$ and $\boldsymbol{V}$ should be constructed with prior information, which unfortunately is not always available. As mentioned before, the main criticism in this regard is the specification of the support spaces for the model parameters, as the bounds and the center of supports are problem-specific and should be chosen with care. If no prior information exists, which is the most common scenario in practice, the supports are usually defined with equally spaced points between arbitrarily chosen lower and upper bounds, with the centre set at zero. Due to the lack of a universally accepted criterion, arbitrariness in the choice of bounds is usually managed based on theoretical constraints, data from previous works, or confidence intervals with high confidence levels from other estimation techniques.

Despite their differences, both the ridge and GME estimators can be applied under very mild conditions, even for ill-conditioned problems. Having two distinct methods provides us with a "*self-consistency*" criterion that guides the selection of the ridge parameter, mitigating its indeterminacy by minimizing the distance between the two estimators.

Definition 4. *The Ridge-GME shrinkage parameter estimator of* λ *is given by*

$$\widehat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \left\| \boldsymbol{Z}\widehat{\boldsymbol{p}} - \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} \right\|_{\infty}, \tag{10}$$

where the vector $\widehat{\boldsymbol{p}}$ *is obtained from the GME estimator, the matrix* $\boldsymbol{Z}$ *is a* $(K \times KM)$ *matrix of supports with limits given by*

$$\boldsymbol{z}_k = \left[\min \left\{ \widehat{\boldsymbol{\beta}}_{k_{\text{ridge}(\lambda)}} \right\}, \max \left\{ \widehat{\boldsymbol{\beta}}_{k_{\text{ridge}(\lambda)}} \right\} \right], \tag{11}$$

for all the possible values of λ *in the domain of ridge values, which is considered* $[0, 1]$ *by default.*

With the Ridge-GME shrinkage parameter estimator, an estimate for λ is obtained without the need for any arbitrary choice from visual analysis of the ridge trace.

As a final remark, Golan et al. [7] defined normalized entropy for the signal $\boldsymbol{X}\boldsymbol{\beta}$ in the model (1) as

$$S(\widehat{\boldsymbol{p}}) = \frac{-\widehat{\boldsymbol{p}}' \ln \widehat{\boldsymbol{p}}}{K \ln M}, \tag{12}$$

where $S(\widehat{\boldsymbol{p}}) = 1$ indicates perfect uncertainty and $S(\widehat{\boldsymbol{p}}) = 0$ indicates no uncertainty. Even if this normalization involves a constant scaling factor, it affects the optimization process by changing the relative weights attributed to the signal and error terms in the GME objective function. Furthermore, it is important to note that there is empirical evidence that different supports for $\boldsymbol{Z}$ and for $\boldsymbol{V}$, as well as different values for M and J, provide different results in terms of variable selection, which means that variable selection through normalized entropy should be interpreted and tested with caution [14]. More generally, in a probabilistic setting, different supports describe different event spaces, and a lack of information on a well-defined event space can be framed as a form of ambiguity [17, Sec. 1]. In turn, inconsistency may arise when combining or comparing results from different specifications, as previous studies explored both at a conceptual level, e.g., in knowledge representations [3], and in practical applications requiring the definition of robust measures for ranking accuracy under such forms of ambiguity [2]. These implications further stress the role of an appropriate choice of supports, taking into account the sources of uncertainty and their potential effects. In this work, we take advantage of the combination of ridge and GME estimators to bound the supports in a data-driven way, devoting to future work the exploration of other information measures, such as normalized entropy, and their use for variable selection within the RidGME estimator.

3 RidGME Estimator

Based on the previous discussion, we can now present the RidGME estimator, as an alternative to ridge regression in ill-conditioned models, starting with the choice of supports for both parameters and errors, which are the basic objects to enable the GME estimation step. This choice may substantially affect the estimation process; in our proposal, the derivation of the supports takes advantage of the ridge estimation step as a robust data-driven procedure, also mitigating the arbitrariness of this choice.

3.1 Step 1: Definition of Parameter and Error Supports

The first step involves choosing the domain $\mathcal{D}_\lambda$ for the shrinkage parameter λ in the ridge regression. We can consider the whole $[0, 1]$ range; alternatively, we can adopt a more parsimonious criterion and select an interval $[a, b] \subseteq [0, 1]$ based on the evaluation of the ridge trace. However, the case $\lambda = 0$ trivially returns the standard OLS estimates, which may be affected by the same collinearity

issues we are addressing. Then, we exclude this occurrence by selecting a small regularization value $\varepsilon > 0$ for the domain $\mathcal{D}_\lambda \subseteq [\varepsilon, 1]$ of the shrinkage parameters involved in this step.

The information from the ridge trace is instrumental in finding, for each $k \in \{1, \ldots, K\}$, the bounds

$$\begin{aligned} z_{\max}^{(\beta_k)} &:= \max\{\widehat{\beta}_k^{(\lambda)},\ \lambda \in \mathcal{D}_\lambda\}, \\ z_{\min}^{(\beta_k)} &:= \min\{\widehat{\beta}_k^{(\lambda)},\ \lambda \in \mathcal{D}_\lambda\}. \end{aligned} \tag{13}$$

These bounds allow us to define the data-driven support in the RidGME algorithm. First, we choose the number of support points for both parameters and error terms, say M and J, respectively. Then, we can identify two specifications for the parameter supports:

– the *asymmetric support* specification, setting

$$Z_{\text{asym}}^{(k)} := \left\{ z_{\min}^{(\beta_k)} + \frac{z_{\max}^{(\beta_k)} - z_{\min}^{(\beta_k)}}{M-1} \cdot t,\ t \in \{0, \ldots, M-1\} \right\},\ k \in \{1, \ldots, K\}; \tag{14}$$

– the *symmetric support* specification, setting

$$\begin{aligned} z_\star^{(k)} &:= \max\left\{ \left| z_{\min}^{(\beta_k)} \right|, \left| z_{\max}^{(\beta_k)} \right| \right\}, \\ Z_{\text{symm}}^{(k)} &:= \left\{ -z_\star^{(k)} + \frac{2t}{M-1} \cdot z_\star^{(k)},\ t \in \{0, \ldots, M-1\} \right\},\ k \in \{1, \ldots, K\}. \end{aligned} \tag{15}$$

Both the specifications generate equally spaced points in the supports; the former (14) provides narrower supports, hence, it tends to reduce the possible variability of parameter estimates in the subsequent GME step, as they are obtained as convex combinations of the points in the support. Also note that each parameter has its own support, which aligns with the data-driven approach proposed in this work.

Even with respect to the support of error terms, we can use two distinct specifications:

– the first approach is the extension of the well-known three-sigma rule [20] to a different number of points (e.g., user-defined) of the support. Formally, we define the support for error terms as

$$V^{(n)} := \left\{ -3\text{sd}(\boldsymbol{y}) + \frac{2\,s}{J-1} \cdot 3\text{sd}(\boldsymbol{y}),\ s \in \{0, \ldots, J-1\} \right\},\ n \in \{1, \ldots, N\}, \tag{16}$$

  where $\text{sd}(\boldsymbol{y})$ stands for the standard deviation of $\boldsymbol{y}$;
– the second approach leverages the information in the previous phase. Indeed, after conducting ridge regression with multiple values of λ to generate the parameter supports, we can also use the estimates to evaluate the residuals

$$\boldsymbol{\eta}_\lambda := \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{ridge}(\lambda)}, \tag{17}$$

and generate the support as

$$\eta_\star^{(n)} := \max\{|\boldsymbol{\eta}_\lambda|, \lambda \in \mathcal{D}_\lambda\}\,,\ n \in \{1,\ldots,N\},$$

$$V^{(n)} := \left\{-\eta_\star^{(n)} + \frac{2\,s}{J-1}\cdot\eta_\star^{(n)},\ s \in \{0,\ldots,J-1\}\right\},\ n \in \{1,\ldots,N\}. \quad (18)$$

The different approaches to estimating the supports reflect the distinguished roles of the parameters (signal) and error terms (statistical noise). Such a distinction is instrumental in different applications [13], even when data exhibit a temporal dimension and a specific update rule is needed for probability distributions to take into account memory effects for parameters, e.g., in the case of streaming data [1].

In terms of numerical implementation, we point out that the maximization in (13) and (18) is carried out over a *discretization* of the support $\mathcal{D}_\lambda \subseteq [\varepsilon, 1]$, namely, we select a given number of points, n_{ridge}, whose default value is set equal to 1000, and generate an equally spaced grid of n_{ridge} points in $\mathcal{D}_\lambda$.

3.2 Step 2: GME Estimation and Residual Bootstrapping

This second step carries out GME estimation based on the supports $Z^{(k)}$, $k \in \{1,\ldots,K\}$, and $V^{(n)}$, $n \in \{1,\ldots,N\}$, obtained in the previous step. The strategy for this step is the same independent on the choice of asymmetric (14) or symmetric (15) supports for the parameters, as well as variable-based (16) or residual-based (18) supports for the error terms.

Specifically, this step combines GME estimation with the well-known procedure for residual bootstrapping:

- conduct GME estimation by solving the optimization problem (5), with consistency constraints (6) and the normalization conditions (7). From the probability distributions $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{w}}$ obtained from this optimization step, we get the GME parameter estimates $\widehat{\boldsymbol{\beta}}_{\text{GME}}$ through (4);
- calculate the estimated responses $\widehat{\boldsymbol{y}}_{\text{GME}} := \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{GME}}$ and the corresponding GME residuals $\boldsymbol{\eta}_{\text{GME}} := \boldsymbol{y} - \widehat{\boldsymbol{y}}_{\text{GME}}$;
- conduct residual bootstrapping by selecting (with replacement) $n_{\text{boot}} = 1000$ N-dimensional samples from the vector of residuals $\boldsymbol{\eta}_{\text{GME}}$. For each $b \in \{1,\ldots,n_{\text{boot}}\}$, denote the corresponding residual sample as $\boldsymbol{\eta}_{\text{GME}}^{(b)}$, which serves as a means to construct a new vector of observations $\boldsymbol{y}_{\text{boot}}^{(b)} := \widehat{\boldsymbol{y}}_{\text{GME}} + \boldsymbol{\eta}_{\text{GME}}^{(b)}$. Then, iterate the GME algorithm to obtain a new parameter estimate $\widehat{\beta}_{\text{GME}}^{(b)}$ starting with the same supports $Z^{(k)}$, $k \in \{1,\ldots,K\}$, and $V^{(n)}$, $n \in \{1,\ldots,N\}$, along with the same data matrix $\boldsymbol{X}$ while using the new vector of observations $\boldsymbol{y}_{\text{boot}}^{(b)}$ for the b-th iteration;
- from the n_{boot} iteration of the residual bootstrap procedure, we obtain an empirical distribution for all quantities of interest, including the parameter GME estimates $\widehat{\beta}_{k,\text{boot}}^{(b)}$, $b \in \{1,\ldots,n_{\text{boot}}\}$, $k \in \{1,\ldots,K\}$. The percentile method is used to infer confidence intervals at a given significance level (e.g., two-sided 95% CI).

We point out that our data-driven procedure derives the supports for the parameters and error terms from the original data; in particular, it relies on the original vector of observations y_n, $n \in \{1, \ldots, N\}$. Therefore, the aforementioned procedure for this step of the RidGME algorithm implicitly relies on the assumption that the distributional properties of the ridge and GME residuals hold approximately even for the bootstrapped distribution in terms of support estimation. On the other hand, the main focus of this work is on potential collinearity issues, which are encoded in the data matrix $\boldsymbol{X}$ that is not modified in the bootstrapping step.

4 Empirical Application

The well-known Portland cement model [18,22,24] is used to illustrate the RidGME estimator. The response variable is the heat evolved per gram of cement, and the four explanatory variables, β_k, $k = 1, 2, 3, 4$, are the amounts of tricalcium aluminate, tricalcium silicate, tetracalcium aluminoferrite and β-dicalcium silicate. The data set includes 13 observations. The original linear model (without intercept) has a $\text{cond}_2\, \boldsymbol{X} \approx 21$, where $\boldsymbol{X}$ is the matrix of explanatory variables and cond_2 is the 2-norm condition number. Interestingly, since the sum of each row is approximately equal to 100 in the original $\boldsymbol{X}$ matrix, adding an intercept, β_0, to the original linear model is obtained $\text{cond}_2\, \boldsymbol{X} \approx 6056$, with $\boldsymbol{X}$ representing now the matrix of explanatory variables with the first column of ones.

Table 1 presents the OLS estimates for the model with and without intercept, β_0, including the bootstrap confidence intervals at 95% (CI95). All values are rounded to four decimals.

Table 1. OLS estimates for the models with and without intercept.

	$\widehat{\beta}_k$	CI95	$\widehat{\beta}_k$	CI95
$\widehat{\beta}_0$	–	–	62.4054	$(-43.0052, 166.1128)$
$\widehat{\beta}_1$	2.1930	$(1.8797, 2.5288)$	1.5511	$(0.4563, 2.6931)$
$\widehat{\beta}_2$	1.1533	$(1.0688, 1.2315)$	0.5102	$(-0.5596, 1.5814)$
$\widehat{\beta}_3$	0.7585	$(0.4923, 1.0378)$	0.1019	$(-0.9918, 1.2609)$
$\widehat{\beta}_4$	0.4863	$(0.4202, 0.5550)$	-0.1441	$(-1.1881, 0.9025)$

Table 2 and Table 3 present the estimates and bootstrap confidence intervals at 95% (CI95) from the RidGME estimator for the parameters of the model with intercept, considering $\varepsilon = 0$. In Table 2, the RidGME estimator is performed with asymmetric supports (14), considering the three-sigma rule (16), on the left, and the residuals from ridge (18), on the right, to define the supports for the error component. On the other hand, in Table 3, the RidGME estimator is performed with symmetric supports (15), considering the three-sigma rule (16),

on the left, and the residuals from ridge (18), on the right, to define the supports for the error component. All values are rounded to four decimals.

Table 2. RidGME estimates; asymmetric; 3σ (left), residuals ridge (right); $\varepsilon = 0$.

	$\widehat{\beta}_k$	CI95	$\widehat{\beta}_k$	CI95
$\widehat{\beta}_0$	31.2334	(30.5811, 31.9154)	31.9373	(27.3898, 34.7482)
$\widehat{\beta}_1$	1.8685	(1.8658, 1.8714)	1.8788	(1.7780, 1.9894)
$\widehat{\beta}_2$	0.8324	(0.8263, 0.8395)	0.8281	(0.7668, 0.9170)
$\widehat{\beta}_3$	0.4271	(0.4239, 0.4304)	0.4113	(0.3144, 0.5175)
$\widehat{\beta}_4$	0.1714	(0.1644, 0.1789)	0.1621	(0.1043, 0.2339)

Table 3. RidGME estimates; symmetric; 3σ (left), residuals ridge (right); $\varepsilon = 0$.

	$\widehat{\beta}_k$	CI95	$\widehat{\beta}_k$	CI95
$\widehat{\beta}_0$	44.1554	(43.5700, 46.3365)	47.4432	(41.1803, 51.0497)
$\widehat{\beta}_1$	0.9155	(0.5672, 0.8608)	1.6544	(1.3682, 1.8530)
$\widehat{\beta}_2$	0.7838	(0.7534, 0.8376)	0.6834	(0.6175, 0.8130)
$\widehat{\beta}_3$	0.0629	(0.0583, 0.1169)	0.1978	(−0.0302, 0.3502)
$\widehat{\beta}_4$	0.0937	(0.0783, 0.1339)	0.0174	(−0.0366, 0.1180)

With the same structure as before, Table 4 and Table 5 present the estimates and bootstrap confidence intervals at 95% (CI95) from the RidGME estimator for the parameters of the model with intercept, considering now $\varepsilon = 0.01$. All values are rounded to four decimals.

Table 4. RidGME estimates; asymmetric; 3σ (left), residuals ridge (right); $\varepsilon = 0.01$.

	$\widehat{\beta}_k$	CI95	$\widehat{\beta}_k$	CI95
$\widehat{\beta}_0$	3.4679	(3.4129, 3.5225)	3.5146	(2.7886, 4.1267)
$\widehat{\beta}_1$	2.1544	(2.1543, 2.1544)	2.1554	(2.1499, 2.1601)
$\widehat{\beta}_2$	1.1191	(1.1188, 1.1195)	1.1210	(1.1097, 1.1300)
$\widehat{\beta}_3$	0.7195	(0.7195, 0.7196)	0.7183	(0.7141, 0.7252)
$\widehat{\beta}_4$	0.4523	(0.4521, 0.4525)	0.4507	(0.4406, 0.4629)

Figure 1 illustrates the results from the previous tables, including the results from the ridge estimator with $\lambda = 0.3$, an acceptable shrinkage parameter from the analysis of the ridge trace, $\text{ridge}_{0.3}$, and with the value of λ that provides the

Table 5. RidGME estimates; symmetric; 3σ (left), residuals ridge (right); $\varepsilon = 0.01$.

	$\widehat{\beta}_k$	CI95	$\widehat{\beta}_k$	CI95
$\widehat{\beta}_0$	2.5959	(2.5081, 2.7404)	4.1958	(3.8777, 5.5966)
$\widehat{\beta}_1$	1.6065	(1.5133, 1.6403)	2.0736	(1.8891, 2.1715)
$\widehat{\beta}_2$	1.1303	(1.1290, 1.1385)	1.1389	(1.1125, 1.1552)
$\widehat{\beta}_3$	0.3536	(0.3341, 0.3810)	0.6229	(0.4585, 0.7009)
$\widehat{\beta}_4$	0.3390	(0.3172, 0.3454)	0.4539	(0.4081, 0.4846)

lowest cross-validation root mean squared error (CV.RMSE), $\text{ridge}_{\text{min}}$. Table 6 presents the CV.RMSE, the standard deviation (sd), and the root mean squared error (RMSE) of the RidGME, ridge and OLS estimators. All values are rounded to four decimals.

Table 6. CV.RMSE, sd and RMSE of RidGME, ridge and OLS estimators.

	CV.RMSE	sd	RMSE
$\text{RidGME}_{\text{asym};3\sigma;0.01}$	2.4464	0.6743	2.0026
$\text{ridge}_{\text{min}}$	2.4503	0.6284	1.9793
$\text{RidGME}_{\text{symm};\eta;0}$	2.4553	0.4548	1.9476
$\text{ridge}_{0.3}$	2.4592	0.6418	2.0109
$\text{RidGME}_{\text{asym};\eta;0.01}$	2.4705	0.7206	2.0071
$\text{RidGME}_{\text{asym};3\sigma;0}$	2.5025	0.5688	1.9425
$\text{RidGME}_{\text{asym};\eta;0}$	2.5170	0.6013	1.9529
$\text{RidGME}_{\text{symm};\eta;0.01}$	2.5422	0.5064	2.0475
OLS	2.7099	0.4310	1.9188
$\text{RidGME}_{\text{symm};3\sigma;0}$	6.2382	2.4842	5.1230
$\text{RidGME}_{\text{symm};3\sigma;0.01}$	14.7594	2.5972	12.4214

Except in two scenarios, both using symmetric support specification and the three-sigma rule, the results reveal that the RidGME estimator is competitive with the *best* choice of ridge estimators (namely, the one associated with λ returning the lowest cross-validation RMSE) and outperforms OLS in terms of CV.RMSE. In particular, the lowest cross-validation error is obtained for the asymmetric support approach, not including the OLS estimates ($\epsilon = 0$) and with the three-sigma rule. Generally, when $\varepsilon = 0$, the in-sample error decreases, with the exception mentioned before, and the lowest RMSE is obtained precisely for the OLS. A good balance between in- and out-of-sample errors seems to be obtained when the symmetric support specification is used alongside the OLS estimation and the residuals. The same finding is verified by analyzing the estimates and bootstrap confidence intervals, by taking the estimates of the param-

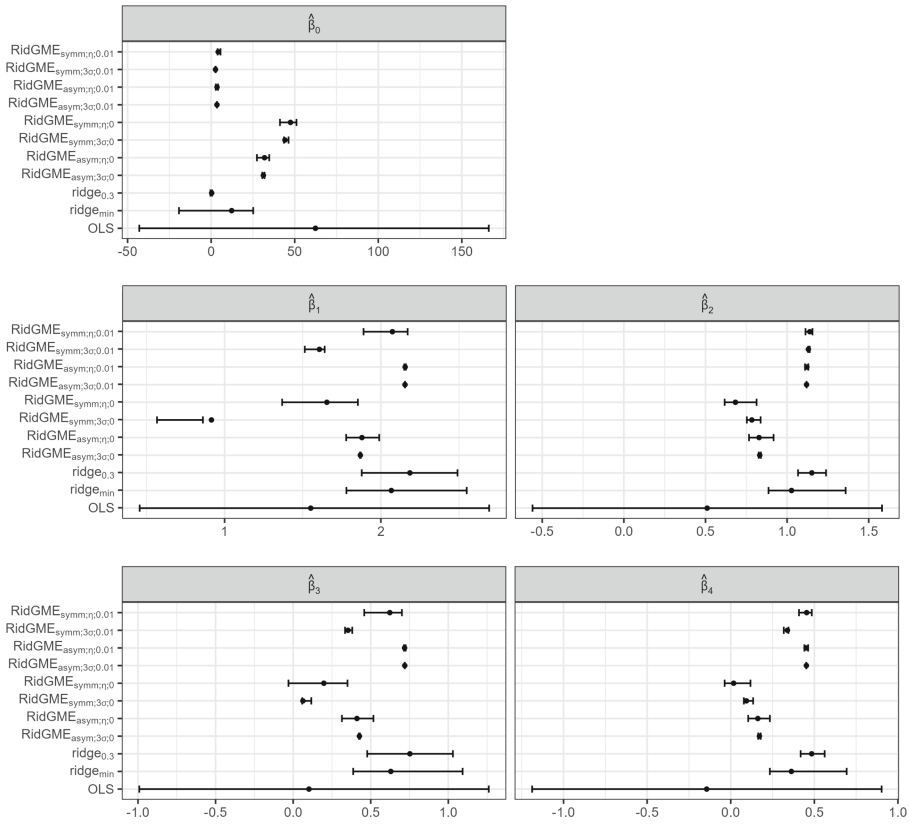

Fig. 1. Estimates and bootstrap CI for the RidGME, ridge and OLS estimators.

eters of the model without constant – the original model without collinearity – as a reference for the accuracy of the estimates. As expected, the OLS confidence intervals are the widest. Note that, for each estimate, these intervals contain all of the others. Asymmetric supports generally produce narrower confidence intervals. The intercept estimate is consistently lower when $\varepsilon = 0.01$. On the other hand, the remaining estimates are generally higher, with some ambiguity for the combination of specifications that produce the worst performance results. The RidGME estimator, although involving the GME estimator through the ridge trace information, is an alternative to the ridge estimator, so the analysis was carried out only in this context. An exhaustive simulation study, including a comparison with the traditional GME estimator, which involves arbitrary choices of supports (its main criticism that is avoided when it is inserted into the RidGME estimator), is left for future research.

5 Concluding Remarks

This work discusses the RidGME estimator as an alternative to ridge regression in ill-conditioned models. In terms of performance in deriving parameter estimates and confidence intervals, the outcomes of our analysis show that RidGME may represent a robust and interpretable approach to deal with the arbitrariness in the choice of the shrinkage parameter in ridge regression, yielding results that are comparable with its optimal choices. However, compared with different procedures available in the literature with a similar purpose, it should be stressed that the RidGME algorithm provides researchers with much more information besides estimation and inference. Indeed, the primary output of the GME steps in RidGME are the parameter, $\widehat{\boldsymbol{p}}$, and the error, $\widehat{\boldsymbol{w}}$, distributions; these distributions can be used as a means to compare the information-theoretic contributions of different parameters beyond their explained variance, e.g., selecting appropriate statistical divergences (e.g., Kullback-Leibler, chi-squared) or metrics defined between probability distributions. The relevance in highlighting individual and joint contributions is key in different application fields, including psychometrics, where multiple factors or latent variables/constructs may be combined to explain or predict manifest variables. Some user-friendly computational codes (MATLAB, R or Python) will be available upon request to the authors and will soon be available at GitHub.

Acknowledgments. This research is supported by the CIDMA (Center for Research and Development in Mathematics and Applications) under the FCT (Portuguese Foundation for Science and Technology) Multi-Annual Financing Program for R&D Units. M.A. also acknowledges the support provided by the project Small Area Estimation for data Quality in Health – SAEQHealth (Bando a cascata Programma PE GRINS – GRINS – Growing Resilient, Inclusive and Sustainable).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angelelli, M., Ciavolino, E., Pasca, P.: Streaming generalized cross entropy. Soft. Comput. **24**(18), 13837–13851 (2020)
2. Angelelli, M., Arima, S., Catalano, C., Ciavolino, E.: A robust statistical framework for cyber-vulnerability prioritisation under partial information in threat intelligence. Expert Syst. Appl. **255**, 124572 (2024)
3. Angelelli, M., Gervasi, M., Ciavolino, E.: Representations of epistemic uncertainty and awareness in data-driven strategies. Soft. Comput. **28**, 13763–13780 (2024). https://doi.org/10.1007/s00500-024-09661-8
4. Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics - Identifying Influential Data and Sources of Collinearity. Wiley, Hoboken, New Jersey (2004)
5. Caputo, M.R., Paris, Q.: Comparative statics of the generalized maximum entropy estimator of the general linear model. Eur. J. Oper. Res. **185**, 195–203 (2008)

6. Golan, A.: Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information. Oxford University Press, New York (2018)
7. Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley, Chichester (1996)
8. Hoerl, A.E., Kennard, R.W.: Ridge regression: applications to nonorthogonal problems. Technometrics **12**(1), 69–82 (1970)
9. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
10. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**(4), 620–630 (1957)
11. Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. **108**(2), 171–190 (1957)
12. Landau, L.D., Lifshitz, E.M.: Statistical Physics, vol. 5. Elsevier (2013)
13. Macedo, P.: Ridge regression and generalized maximum entropy: an improved version of the Ridge-GME parameter estimator. Commun. Stat. Simul. Comput. **46**(5), 3527–3539 (2017)
14. Macedo, P., Costa, M.C., Cruz, J.P.: Normalized entropy: a comparison with traditional techniques in variable selection. In: Simos, T.E., Tsitouras, C. (eds.) Symposium on Mathematical Methods for Artificial Intelligence. vol. 2425, p. 190002. AIP Conference Proceedings (2022)
15. Macedo, P., Scotto, M., Silva, E.: On the choice of the ridge parameter: a maximum entropy approach. Commun. Stat. Simul. Comput. **39**(8), 1628–1638 (2010)
16. McDonald, G.C.: Ridge regression. Wiley Interdiscip. Rev. Comput. Stat. **1**(1), 93–100 (2009)
17. Mukerji, S.: Understanding the nonadditive probability decision model. Econ. Theor. **9**, 23–46 (1997)
18. Muniz, G., Kibria, B.: On some ridge regression estimators: an empirical comparisons. Commun. Stat. Simul. Comput. **38**(3), 621–630 (2009)
19. Preckel, P.V.: Least squares and entropy: a penalty function perspective. Am. J. Agr. Econ. **83**(2), 366–377 (2001)
20. Pukelsheim, F.: The three sigma rule. Am. Stat. **48**(2), 88–91 (1994)
21. Ryan, T.P.: Modern Regression Methods, 2nd edn. Wiley, Hoboken, New Jersey (2009)
22. Sakallıoğlu, S., Kaçıranlar, S.: A new biased estimator based on ridge estimation. Stat. Pap. **49**(4), 669–689 (2008)
23. Shannon, C.E.: A mathematical theory of communication. Bell System Tech. J. **27**(3), 379–423 (1948)
24. Woods, H., Steinour, H.H., Starke, H.R.: Effect of composition of Portland cement on heat evolved during hardening. Ind. Eng. Chem. **24**(11), 1207–1214 (1932)
25. Zhang, R., McDonald, G.C.: Characterization of ridge trace behavior. Commun. Stat. Theory Methods **34**(7), 1487–1501 (2005)