

# A Data-Driven Weighted Generalized Maximum Entropy Estimator

Ximing Wu\*

September 14, 2009

(Preliminary and Incomplete)

## **Abstract**

The method of Generalized Maximum Entropy (GME), proposed in Golan, Judge and Miller (1996), uses an objective function that is the sum of the entropies for coefficient distributions and disturbance distributions. This method can be generalized to the weighted GME, where different weights are assigned to these two components. We propose a data-driven method to select the weights in the entropy objective function. We use the least squares cross validation to derive the optimal weight. Monte Carlo simulations demonstrate that the data-driven weighted GME (D-GME) estimator is comparable to and often outperforms the conventional GME estimator, which places equal weights on the coefficient and disturbance distributions.

---

\*Department of Agricultural Economics, Texas A&M University, College Station, TX 77843-2124. email: xwu@ag.tamu.edu

# 1 Introduction

Jaynes' Principle of Maximum Entropy prescribes a method of density construction based on limited information. This approach, and its generalization through minimization of the cross entropy by Kullback, Leibler and others, have found wide-spread applications in various fields of science. See, for example, Skilling (1989) and references therein. In particular, motivated by the famous Jaynes' die problem, this principle provides a solution to the "ill-posed" inverse problem. Golan, Judge and Mill (1996, GJM henceforth) generalizes this method to the regression framework. In particular, they reparameterize the coefficients and disturbances in a linear regression model as discrete random variables on a bounded support and maximize the sum of entropies of distributions on these supports. The coefficients of interest are then calculated as the expectation of random variables on the prescribed supports under the derived distributions of the entropy maximization. They further generalize this so-called Generalized Maximum Entropy (GME) method to a weighted one, in which different weights are assigned to the entropy of coefficient distributions and of disturbance distribution.

Although the specifications of the coefficient and disturbance supports can be guided by non-sample information and preliminary estimates, there is no clear guidance on how to select the weights placed on the entropies of the coefficient distributions and disturbance distributions. In this study, we propose a data-driven method of selecting this weight parameter, which balances the two components in the entropy maximization objective function in an automatic, objective manner. We use the least squares cross validation in our implementation of the proposed method. The results are shown to improve on the conventional GME estimator under various scenarios.

## 2 Generalized Maximum Entropy Estimator

In this section, we briefly review the literature on information entropy, the principle of maximum entropy and its applications to possibly ill-posed inverse problem. We then discuss the generalized maximum entropy estimator for linear regressions and its statistical properties.

## 2.1 The GME Estimator

GJM presented a novel information-theoretic estimator that is based on the celebrated principle of Maximum Entropy (ME). Let  $X$  be a random variable with possible outcome values  $x_k, k = 1, \dots, K$  and probabilities  $p_k$  such that  $\sum_{k=1}^K p_k = 1$ . Shannon (1948) defined the information entropy of the distribution of probabilities,  $\mathbf{p} = \{p_k\}_{k=1}^K$  as the measure

$$H(\mathbf{p}) = - \sum_{k=1}^K p_k \log p_k,$$

where  $0 \log 0 = 0$ . The entropy measures the uncertainty of a distribution and reaches a maximum when  $p_1 = p_2 = \dots = p_K = 1/K$  or, in other words, when the probabilities are uniform.

Jaynes (1957) proposed using the entropy concept in choosing the unknown distribution of probabilities. Under what Jaynes called the maximum entropy principle, one chooses the distribution for which the information (data) is just sufficient to determine the probability assignment. More precisely, one chooses the distribution, among those distributions consistent with known information, that maximizes the entropy. This maximum entropy formulation that is based on the work of Shannon (1948) and Jaynes (1957) has been extended by Kullback (1959), Levine (1980) and many others who are identified in the collection of papers in Levine and Tribus (1979). Axiomatic arguments for the justification of the ME principle have been made by Shore and Johnson (1980), Jaynes (1984), Skilling (1989) and Csiszár (1991). See GJM for an in-depth review of this literature.

Suppose that that  $E[X] = y$ . According to the ME principle, one can construct a density of  $X$  by maximizing

$$H(\mathbf{p}) = -\mathbf{p}' \log \mathbf{p}$$

subject to the data consistency and normalization-additivity requirements

$$\begin{aligned} y &= \mathbf{X}' \mathbf{p}, \\ \mathbf{p}' \mathbf{1} &= 1, \end{aligned}$$

where  $\mathbf{X}, \mathbf{p}$  are  $K \times 1$  vectors, and  $\mathbf{1}$  is a  $K \times 1$  vector of ones. The analytical solution to

the entropy maximization problem can be obtained by the Lagrangian function

$$\mathcal{L} = -\mathbf{p}' \log \mathbf{p} + \lambda'(y - \mathbf{X}'\mathbf{p}) + \mu(1 - \mathbf{p}'\mathbf{1}),$$

with optimality conditions

$$\partial\mathcal{L}/\partial\mathbf{p} = -\log \hat{\mathbf{p}} - \mathbf{1} - \mathbf{X}'\hat{\lambda} - \hat{\mu} = \mathbf{0},$$

$$\partial\mathcal{L}/\partial\lambda = y - \mathbf{X}'\hat{\mathbf{p}} = 0,$$

$$\partial\mathcal{L}/\partial\mu = 1 - \hat{\mathbf{p}}'\mathbf{1} = 0.$$

We can then solve for  $\hat{\mathbf{p}}$ , in terms of  $\hat{\lambda}$  to get

$$\hat{\mathbf{p}} = \exp(-\mathbf{X}'\hat{\lambda})/\Omega(\hat{\lambda}), \tag{1}$$

where

$$\Omega(\hat{\lambda}) = \sum_k \exp(-X'_k \hat{\lambda})$$

is a normalization factor that converts the relative probabilities into absolute probabilities.

The solution (1) establishes a unique non-linear relation between  $\hat{\mathbf{p}}$  and  $y$  through  $\hat{\lambda}$ . Unlike conventional regression methods such as the least squares estimator, the ME method can be used for inferences in the so-called ill-posed problem. For instance, let us look at the famous Jaynes' dice problem. Suppose that one is given a six-sided die that can take on the values  $k = 1, 2, \dots, 6$ , and asked to estimate the probabilities for each possible outcome given that the average outcome from a large number of independent rolls of the die was  $y$ . The ME formulation of this problem is as follows:

$$\max H(\mathbf{p}) = -\sum_{k=1}^6 p_k \log p_k$$

subject to

$$\sum_{k=1}^6 p_k x_k = y,$$

$$\sum_{k=1}^6 p_k = 1,$$

where  $x_k = k$  for each  $k = 1, \dots, 6$ . This is an inverse problem with one observation (the mean) and six unknowns and thus clearly ill-posed. Using the ME framework, one is able to assign unique probability to each possible outcome. For example, when the average outcome is 3.5, the ME method assigns equal weights to all six outcomes. If the average outcome is larger/smaller than 3.5, the ME method “tilts” the distribution smoothly such that the weights to each side of the die increases/decreases with number of dots on it.

GJM generalizes the ME solution for the inverse problem to the regression framework. Consider the linear model

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

where  $\mathbf{y}$  is a  $T$ -dimensional vector of observables,  $\mathbf{X}$  is a  $T \times K$  design matrix, and  $\boldsymbol{\beta}$  is a  $K$ -dimensional vector of unknown parameters. The unobservable disturbance vector  $\mathbf{e}$  may represent one or more sources of noise in the observed system, including sample and non-sample errors in the data, randomness in the behavior of the economic agents, and specification or modeling errors.

GJM reparameterize model (2) such that  $\boldsymbol{\beta}$  are represented by expectations of random variables with compact supports. In particular, suppose that  $\beta_k$  is a discrete random variable with a compact support and  $M$  possible outcomes  $\mathbf{z}_k = [z_{k1}, \dots, z_{kM}]'$ , where  $2 \leq M \leq \infty$ , and  $z_{k1}$  and  $z_{kM}$  are the plausible extreme values (upper and lower bounds) of  $\beta_k$ . We can express  $\beta_k$  as a convex combination

$$\beta_k = \mathbf{z}'_k \mathbf{p}_k,$$

where  $\mathbf{p}_k = [p_{k1}, \dots, p_{kM}]'$  is an  $M$ -dimensional vector of positives weights that sum to one. Further, these convex combinations may be assembled in matrix form so that  $\boldsymbol{\beta}$  may be

written as

$$\boldsymbol{\beta} = \mathbf{Z}\mathbf{p} = \begin{bmatrix} z'_1 & 0 & \cdot & 0 \\ 0 & z'_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & z'_K \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ p_K \end{bmatrix},$$

where  $\mathbf{Z}$  is a  $K \times KM$  matrix and  $\mathbf{p}$  is a  $KM$ -dimensional vector of weights.

One can then use the ME principle to estimate  $\mathbf{p}$  as follows:

$$\max H(\mathbf{p}) = -\mathbf{p}' \log \mathbf{p} \quad (3)$$

subject to

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{Z}\mathbf{p}, \\ \mathbf{1}_K &= (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p}, \end{aligned} \quad (4)$$

where  $\mathbf{1}_K$  and  $\mathbf{1}_M$  are  $K$  and  $M$ -dimensional vector of ones respectively, and  $\mathbf{I}_K$  is the  $K$ -dimensional identity matrix.

Since model (3) it is a nonlinear estimator and the data consistency requirement (4) attempts to force the model (without the error term) to hold for all observations, the entropy maximization problem may not necessarily converge. To circumvent this numerical difficulty, GJM further generalize model (3) by incorporating the unobservable disturbance term,  $\mathbf{e}$ , explicitly into the data consistency requirement. Assuming that  $\mathbf{e}$  is a random vector with finite location and scale parameters, one can represent his uncertainty about the outcome of the error process by treating each  $e_t$  as a finite and discrete random variable with  $2 \leq J \leq \infty$  possible outcomes. Suppose that there exist sets of error bounds,  $v_{t1}$  and  $v_{tJ}$ , for each  $e_t$  so that  $\Pr[v_{t1} < e_t < v_{tJ}]$  may be made arbitrarily small. One can then write

$$e_t = \mathbf{v}'_t \mathbf{w}_t,$$

where  $\mathbf{v}_t = [v_{t1}, \dots, v_{tJ}]'$  is a finite support for  $e_t$ , and  $\mathbf{w}_t = [w_{t1}, \dots, w_{tJ}]'$  is a  $J$ -dimensional vector of positive weights that sum to one. The  $T$  unknown disturbances may be written in

matrix form as

$$\mathbf{e} = \mathbf{V}\mathbf{w} = \begin{bmatrix} v'_1 & 0 & \cdot & 0 \\ 0 & v'_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & v'_T \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ w_T \end{bmatrix},$$

where  $\mathbf{V}$  is a  $T \times TJ$  matrix and  $\mathbf{w}$  is a  $TJ$ -dimensional vector of weights, which are strictly positive and sum to one for each  $t$ .

Using the reparameterized unknowns,  $\boldsymbol{\beta} = \mathbf{Z}\mathbf{p}$  and  $\mathbf{e} = \mathbf{V}\mathbf{w}$ , one can rewrite model (2) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}\mathbf{Z}\mathbf{p} + \mathbf{V}\mathbf{w}.$$

The Generalized Maximum Entropy (GME) estimator is then defined by

$$\max H(\mathbf{p}, \mathbf{w}) = -\mathbf{p}' \log \mathbf{p} - \mathbf{w}' \log \mathbf{w}, \quad (5)$$

subject to

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{Z}\mathbf{p} + \mathbf{V}\mathbf{w}, \\ \mathbf{1}_K &= (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p}, \\ \mathbf{1}_T &= (\mathbf{I}_T \otimes \mathbf{1}'_J)\mathbf{w}. \end{aligned}$$

This optimization problem can be solved using the Lagrangian method. The Lagrangian equation takes the form

$$\mathcal{L} = H(\mathbf{p}, \mathbf{w}) + \boldsymbol{\lambda}'[\mathbf{y} - \mathbf{X}\mathbf{Z}\mathbf{p} - \mathbf{V}\mathbf{w}] + \boldsymbol{\theta}'[\mathbf{1}_K - (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p}] + \boldsymbol{\tau}'[\mathbf{1}_T - (\mathbf{I}_T \otimes \mathbf{1}'_J)\mathbf{w}],$$

where  $\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\tau}$  are  $T \times 1, K \times 1, T \times 1$  vectors of Lagrangian multipliers respectively. Solving

the first order conditions yields

$$\hat{p}_{km} = \frac{\exp(z_{km}X'_k\hat{\boldsymbol{\lambda}})}{\Omega_k(\hat{\boldsymbol{\lambda}})},$$

$$\hat{w}_{tj} = \frac{\exp(v_{tj}\hat{\boldsymbol{\lambda}})}{\Psi_t(\hat{\boldsymbol{\lambda}})},$$

where

$$\Omega_k(\hat{\boldsymbol{\lambda}}) = \sum_{m=1}^M \exp(z_{km}X'_k\hat{\boldsymbol{\lambda}}),$$

$$\Psi_t(\hat{\boldsymbol{\lambda}}) = \sum_{j=1}^J \exp(v_{tj}\hat{\boldsymbol{\lambda}}).$$

Furthermore, this constrained optimization problem can be rewritten as an unconstrained problem, in which the objective function takes the form

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbf{y}'\boldsymbol{\lambda} - \sum_{k=1}^K \log(\Omega_k(\boldsymbol{\lambda})) - \sum_{t=1}^T \log(\Psi_t(\boldsymbol{\lambda})) \equiv \mathcal{M}(\boldsymbol{\lambda}). \quad (6)$$

The minimal value function,  $\mathcal{M}(\boldsymbol{\lambda})$ , may be interpreted as a constrained expected log-likelihood function. This dual version of the GME problem simplifies the estimation considerably. The analytical gradient of the dual problem

$$\nabla_{\boldsymbol{\lambda}}\mathcal{M}(\boldsymbol{\lambda}) = \mathbf{y} - \mathbf{XZ}\mathbf{p} - \mathbf{V}\mathbf{w}$$

is simply the model consistency constraint. The Hessian matrix of  $\mathcal{M}(\boldsymbol{\lambda})$  takes the form

$$\nabla_{\boldsymbol{\lambda}\boldsymbol{\lambda}'} = -\mathbf{XZ}\nabla_{\boldsymbol{\lambda}'}\mathbf{p}(\boldsymbol{\lambda}) - \mathbf{V}\nabla_{\boldsymbol{\lambda}'}\mathbf{w}(\boldsymbol{\lambda}) = -\mathbf{X}\boldsymbol{\Sigma}_Z(\boldsymbol{\lambda})\mathbf{X}' - \boldsymbol{\Sigma}_V(\boldsymbol{\lambda}), \quad (7)$$

where  $\boldsymbol{\Sigma}_Z(\boldsymbol{\lambda})$  and  $\boldsymbol{\Sigma}_V(\boldsymbol{\lambda})$  are covariance matrices for distributions  $\mathbf{p}(\boldsymbol{\lambda})$  and  $\mathbf{w}(\boldsymbol{\lambda})$  respectively. Both covariance matrices are strictly positive definite for any interior solution,  $(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ , which ensures the uniqueness of the solution.

## 2.2 Statistical Properties of GME

Under some mild regularity conditions, GJM establish the large sample properties of the GME estimation. They also analyze its small sample properties, both analytically for some special cases and numerically using Monte Carlo simulations.

The noise term,  $\mathbf{V}\mathbf{w}$ , effectively “loosens” the model constraints for a given set of observations, and thus an interior solution is more likely. On the other hand, because of the presence of  $\Sigma_{\mathbf{V}}(\boldsymbol{\lambda})$ , which is positive definite, in the Hessian matrix, the GME estimator behaves like the ridge estimator in the sense that the all coefficients are shrunk toward zero. Consider, for simplicity, the case where  $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}_T$  and  $\mathbf{X}$  is orthogonal. The approximate covariance matrix of the GME estimate  $\hat{\boldsymbol{\beta}}$  is

$$\sigma^2\Sigma_{\mathbf{Z}}(\Sigma_{\mathbf{Z}} + \Sigma_{\mathbf{V}})^{-2}\Sigma_{\mathbf{Z}}.$$

The finite sample performance of this estimator clearly depends on the specification of the error support  $\mathbf{V}$ . Intuitively, the wider is  $\mathbf{V}$ , the larger is the degree of shrinkage toward zero. GJM proposed to use the  $3\sigma$  rule for the error support, where  $\sigma$  refers to the standard deviation of the disturbance. In practice,  $\sigma$  is replaced by its consistent estimate, such as that based on the OLS regression.

A second factor that may influence the finite sample performance of the GME estimator is the specification of the coefficient support,  $\mathbf{Z}$ . The restrictions imposed on the parameter space through  $\mathbf{Z}$  reflect prior knowledge about the unknown parameters. However, such knowledge is not always available, and researchers may want to entertain a variety of plausible bounds on  $\boldsymbol{\beta}$ . As the parameter supports are widened, the GME risk functions modestly shift upward reflecting the reduced constraints on the parameter space. Hence, wide bounds may be used without extreme risk consequences, if one’s knowledge is minimal, to ensure that  $\mathbf{Z}$  contains  $\boldsymbol{\beta}$ . Intuitively, widening the bounds increases the impact of the data and decreases the impact of the support. On the other hand, narrowing the parameter supports only improves the risk as long as the true parameter vector is well in the interior of the support. GJM conducted Monte Carlo simulations on the impact of  $\mathbf{Z}$  by using different supports. They found modest impacts of varying the parameter support on the estimation.

For both the parameter and error supports, we need to select the number of points,  $M$  and  $J$ , respectively. Since the variance of the distributions  $\mathbf{p}(\boldsymbol{\lambda})$  and  $\mathbf{w}(\boldsymbol{\lambda})$  depends on the specification of the support, the dimension of the supports may affect the sampling properties of the estimator. Adding more points to the support of  $\mathbf{Z}$  should decrease the variance of the associated point estimator. On the other hand, it increases the computational burden of the problem. GJM reported an experiment showing that the estimator improves as the number of support points  $M$  increases for small and modest  $M$ . The greatest improvement is observed when  $M$  is increased from three to five.

The specifications of the parameter and error supports clearly affect the GME estimation results. In addition, the specification of the dual loss objective function (5) can also influence the estimator. By accounting for the unknown signal and noise components in the consistency relations, the GME estimates of the unknown parameter  $\boldsymbol{\beta}$  and disturbances  $\mathbf{e}$  are jointly determined. As a result, the entropy based objective reflects statistical losses in the sample space (prediction) and in the parameter space (precision). It is noted, however, the objective function (5) implicitly places equal weights on the parameter and error entropies.

To avoid arbitrarily assigning weights to the two loss components, GJM suggested a weighed GME estimator with the following objective function

$$H(\mathbf{p}, \mathbf{w}; \gamma) = -(1 - \gamma)\mathbf{p}' \log \mathbf{p} - \gamma\mathbf{w}' \log \mathbf{w}, \quad (8)$$

where  $\gamma \in (0, 1)$  controls the weights given to the two entropies. The corresponding unconstrained GME( $\gamma$ ) objective function is

$$\mathcal{M}(\boldsymbol{\lambda}; \gamma) = \mathbf{y}'\boldsymbol{\lambda} - (1 - \gamma) \sum_{k=1}^K \log(\Omega_k(\boldsymbol{\lambda}; \gamma)) - \gamma \sum_{t=1}^T \log(\Psi_t(\boldsymbol{\lambda}; \gamma)).$$

One can then show that

$$\hat{p}_{km} = \frac{\exp(z_{km}X'_k\hat{\boldsymbol{\lambda}}/(1 - \gamma))}{\Omega_k(\hat{\boldsymbol{\lambda}}; \gamma)},$$

$$\hat{w}_{tj} = \frac{\exp(v_{tj}\hat{\boldsymbol{\lambda}}/\gamma)}{\Psi_t(\hat{\boldsymbol{\lambda}}; \gamma)},$$

where  $\hat{\lambda}$  are functions of  $\gamma$ , and

$$\Omega_k(\hat{\lambda}; \gamma) = \sum_{m=1}^M \exp(z_{km} X'_k \hat{\lambda} / (1 - \gamma)),$$

$$\Psi_t(\hat{\lambda}; \gamma) = \sum_{j=1}^J \exp(v_{tj} \hat{\lambda} / \gamma).$$

GJM illustrated that the entropy optimization results are affected by  $\gamma$ . Furthermore, they reported that the effect of the weight on the estimation results cannot be determined unambiguously even for some very simple cases.

### 3 Data-driven GME (D-GME)

GJM demonstrate various merits of the GME estimator, especially its resistance to multicollinearity problem. The implementation of the GME estimation, however, requires several “human” decisions which are not required in the OLS. As discussed in the previous section, various factors may affect the performance of the GME estimator. GJM provides some guidance on the specifications of these factors. First, non-sample information can be useful. For example, it is not uncommon in practice that the sign, range or approximate multitude of coefficients in question are known *a priori*. This information provides useful guidance on the specification of the coefficient support. Similarly, non-sample information regarding the error distribution is sometimes available. For instance, it is well known that the error distributions in financial studies have fat tails. Accordingly, one can use a wider error support than the usual  $3\sigma$  rule.

A second useful principle is adaptation. Generally any consistent estimators can provide useful information on the coefficients and the distribution of the disturbance. Thus, one can tailor his specification of the coefficient and error supports based on preliminary consistent estimates. For example, to use the usual  $3\sigma$  rule, one can replace  $\sigma$  with a consistent estimate. In the spirit of adaptive estimation, one can further tailor the error support such that it reflects characterizations of the error distribution, such as skewness, fat-tailedness, and so on.

Lastly, the maximum entropy problem can be further generalized to the minimum cross entropy problem. The cross entropy, or Kullback-Leibler information criteria, for two distributions,  $\mathbf{p}$  and  $\mathbf{q}$  (with a common support) is defined as

$$D(\mathbf{p}, \mathbf{q}) = \sum_k p_k \log(p_k/q_k).$$

The cross entropy measures the discrepancy between  $\mathbf{p}$  and  $\mathbf{q}$ . It is non-negative and equals zero if and only if  $\mathbf{p} = \mathbf{q}$  almost everywhere. Suppose in addition to the model consistency requirement, prior information is available in the form of probability mass function on the discrete support for the coefficients and disturbance, it can be incorporated into the estimation by minimizing the cross entropy subject to the model consistency and additivity constraint. Denote the prior distributions for the coefficient and errors  $\mathbf{q}$  and  $\mathbf{u}$  respectively. The minimum cross entropy problem is formulated as follows:

$$\min D(\mathbf{p}, \mathbf{w}; \mathbf{q}, \mathbf{u}) = -\mathbf{p}' \log \mathbf{p}/\mathbf{q} - \mathbf{w}' \log \mathbf{w}/\mathbf{u},$$

subject to

$$\begin{aligned} \mathbf{y} &= \mathbf{XZ}\mathbf{p} + \mathbf{V}\mathbf{w}, \\ \mathbf{1}_K &= (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p}, \\ \mathbf{1}_T &= (\mathbf{I}_T \otimes \mathbf{1}'_J)\mathbf{w}. \end{aligned}$$

The solution takes the form

$$\begin{aligned} \tilde{\mathbf{p}}_{km} &= \frac{q_{km} \exp(z_{km} X'_k \tilde{\boldsymbol{\lambda}})}{\Omega_k(\tilde{\boldsymbol{\lambda}})}, \\ \tilde{\mathbf{w}}_{tj} &= \frac{u_{tj} \exp(v_{tj} \tilde{\boldsymbol{\lambda}})}{\Psi_t(\tilde{\boldsymbol{\lambda}})}, \end{aligned}$$

where

$$\Omega_k(\tilde{\boldsymbol{\lambda}}) = \sum_{m=1}^M q_{km} \exp(z_{km} X_k' \tilde{\boldsymbol{\lambda}}),$$

$$\Psi_t(\tilde{\boldsymbol{\lambda}}) = \sum_{j=1}^J u_{tj} \exp(v_{tj} \tilde{\boldsymbol{\lambda}}).$$

Clearly, the GME is a special case of the GCE, where the prior distributions  $\mathbf{q}$  and  $\mathbf{u}$  have been set to constant. The (GCE) estimator provides an effective way of incorporating non-sample information into the estimation process. For example, one can set  $\mathbf{q} = [.2, .6, .4]$  for a coefficient support  $\mathbf{z} = [-z, 0, z]$  to strengthen the shrinkage toward zero. Or if it is known that the error distribution is skewed, one can use a support such as  $\mathbf{V} = [-v, 0, 2v]$ . Obviously, the implicit uniform prior distribution of the GME does not ensure that the mean of the error distribution is centered at zero. To “re-center” the prior error distribution on this asymmetric support, one can use a prior distribution  $\mathbf{u} = [4/9, 1/3, 2/9]$ .

To summarize, one can use non-sample information and preliminary consistent estimate to aid the specification of the coefficient and error supports. The prior distributions on these supports can be further “tilted” exponentially by non-uniform prior distributions incorporated through the GCE framework. In contrast, there is no clear guidance for the selection of  $\gamma$  in the weighted GME problem. In this section, we propose a data-driven generalized maximum entropy (D-GME) method to select  $\gamma$ . We use the method of least squares cross-validation (LSCV), which is widely used in nonparametric estimations. In particular, this method is implemented as follows.

1. Given the coefficient support  $\mathbf{Z}$ , disturbance support  $\mathbf{V}$ , and weight  $\gamma$ , estimate  $\boldsymbol{\beta}$  using the weighed GME method (8), on  $T - 1$  observations, with the  $t^{\text{th}}$  observation omitted for  $t = 1, \dots, T$ . Denote the estimate  $\hat{\boldsymbol{\beta}}_{-t}(\gamma)$ .<sup>1</sup>
2. Calculate the squared prediction error  $\hat{s}_t(\gamma) = (y_t - \mathbf{x}_t \hat{\boldsymbol{\beta}}_{-t}(\gamma))^2$  for each  $t$ .
3. Select  $\gamma$  such that it minimizes the least squares sum of prediction errors  $\sum_{t=1}^T \hat{s}_t(\gamma)$ .

---

<sup>1</sup>For simplicity, we use uniform prior distributions for  $\mathbf{Z}$  and  $\mathbf{V}$ .

The LSCV method is known to produce asymptotically consistent estimate for the optimal smoothing or regularization parameter (Hall, 1984; Stone, 1984; Hall and Marron, 1987). By doing so, we allow the data to specify the weight for the coefficient uncertainty and the disturbance uncertainty.

## 4 Monte Carlo Simulations

To investigate the finite sample performance of the proposed D-GME method, we conducted some Monte Carlo simulations. The purpose of these experiments is to compare the D-GME with the conventional GME, which places equal weights on the entropy of the coefficient and that of the disturbance. It is not intended as an investigation of the GME method, where careful selection of the support for the coefficient and the disturbance is crucial to its performance. For simplicity, we choose not to use non-sample information in the specification of coefficient and error support, and to use the simple GME approach, or, uniform prior distributions in the GCE framework.

### 4.1 Regressions with normal errors

Following GJM’s Monte Carlo simulation setup, we investigate the performance of the D-GME on linear models where the design matrices vary in degree of multicollinearity. Recall that the GME is similar to the ridge regression as a robust estimator against multicollinearity. Thus we are interested in its performance in the presence of multicollinearity.

We measure a matrix’s multicollinearity using its condition number, which is the ratio between its largest and smallest eigenvalues. Let  $\mathbf{X}$  be a  $n \times 4$  matrix which is generated randomly from an i.i.d. standard normal distribution. To form a design matrix with a desired condition number,  $\kappa(\mathbf{X}'\mathbf{X}) = \mu$ , the singular value decomposition of  $\mathbf{X} = \mathbf{Q}\mathbf{L}\mathbf{R}$  was recovered. Then, the eigenvalues in  $\mathbf{L}$  were replaced with the vector

$$a = \left[ \sqrt{\frac{2}{1+\mu}}, 1, 1, \sqrt{\frac{2\mu}{1+\mu}} \right],$$

which has length  $K = 4$ . The new design matrix,  $\mathbf{X}_a = \mathbf{Q}\mathbf{L}_a\mathbf{R}$ , is characterized by

$\kappa(\mathbf{X}'_a \mathbf{X}_a) = \mu$ , and the condition number may be specified *a priori*. We then set

$$\mathbf{y} = \mathbf{X}_a \boldsymbol{\beta} + \mathbf{e},$$

where  $\boldsymbol{\beta} = [2, 1, -3, 2]'$ , and  $\mathbf{e}$  are  $T$  i.i.d. standard normal random errors.

In the Monte Carlo simulations, we consider three estimators: the OLS, GME and D-GME. For both GME estimators, we use a five-point support  $\mathbf{Z} = [-z, -z/2, 0, z/2, z]$  for  $z = 10, 20, 30, 50, 100$  respectively. For the error support, we set  $\mathbf{V} = [-\hat{\sigma}, -\hat{\sigma}/2, 0, \hat{\sigma}/2, \hat{\sigma}] \times 3$ , where  $\hat{\sigma}$  is the the standard error of the OLS residuals. We consider sample size  $n = 30$  and  $n = 50$ . Each experiment is repeated 500 times.

We use the LSCV to select the optimal weight  $\gamma$ . In particular, we use a line search over the interval  $[0, 1]$  to locate the  $\gamma$  that minimizes the squared prediction errors.<sup>2</sup> We report the Mean Squared Errors (MSE) of coefficient estimates,  $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ , in Table 1. When  $\mu = 1$ , the MSE of the OLS is close to 4, its theoretical value, in all cases. Not surprisingly, when  $\mathbf{X}$  is orthogonal, the OLS outperforms both GME estimators, which are shrinkage estimators and thus biased. On the other hand, in most cases where  $\mu > 1$ , the two GME estimators have smaller MSEs than the OLS does. This is consistent with the famous Stein's phenomenon that the OLS is dominated by some shrinkage estimators in multiple linear regressions.

Comparing the GME and D-GME, we note that when  $z = 10$ , or the coefficient support is defined on  $[-10, 10]$ , the MSEs of the GME are smaller than those of the D-GME. This result suggests that when a relatively precise coefficient support is used, the GME estimator has a smaller risk. Intuitively, with a narrow support for the coefficients that covers the true values, the coefficients can be estimated precisely, regardless the choice of the weight  $\gamma$  in a weighted GME framework. On the other hand, the potential benefit of the D-GME is largely offset by the additional variation entailed by the data-driven method of selecting the entropy weight  $\gamma$ . However, in practice, the improvement due to narrow coefficient supports is only obtainable if the supports contain the true unknown coefficient values. Without prior or non-sample information, using a narrow coefficient support increases the risk of missing the true values and renders the estimator inconsistent.

---

<sup>2</sup>We searched over an equally-spaced interval  $\boldsymbol{\rho} = [\log(0.01), \log(0.01) + h, \log(0.01) + 2h, \dots, \log(0.99)]$ , where  $h = (\log(0.99) - \log(0.01))/15$  and  $\gamma$  is set to  $\exp(\rho)$ .

Nonetheless, when  $z \geq 20$ , the D-GME outperforms the GME considerably. Furthermore, the performance of the D-GME relative to that of the GME improves with both the width of coefficient support and the condition number. The average ratios between the MSEs of the D-GME and those of the GME across two sample sizes are respectively  $[1.11, 0.87, 0.78, 0.69, 0.70]$  for  $z = [10, 20, 30, 50, 100]$ , while these ratios are respectively  $[1.00, 0.87, 0.76, 0.69]$  across the condition numbers  $\mu = [1, 10, 20, 50]$ . In addition, it is noted that the performance of the D-GME appears to stabilize for  $z \geq 50$ . In other words, its performance seems to be affected little when a wide coefficient support is further widened. In contrast, the MSE of the GME increases with the coefficient support and reaches the level of that of the OLS for  $z \geq 50$ . Given the fact that a narrow coefficient support increases the risk of inconsistent estimates, the stability of the D-GME under a wide range of coefficient supports is highly desirable.

Table 1: MSE of regressions with normal errors

$z$	$\mu$	$n = 30$			$n = 50$		
		OLS	D-GME( $\hat{\gamma}$ )	GME	OLS	D-GME( $\hat{\gamma}$ )	GME
10	1	3.84	4.06 (0.26)	3.57	3.84	4.26 (0.24)	3.67
	10	7.38	5.97 (0.24)	5.37	6.57	5.77 (0.23)	4.91
	20	10.83	7.29 (0.24)	6.78	10.63	7.06 (0.22)	6.49
	50	19.85	8.28 (0.23)	7.99	21.29	8.15 (0.23)	7.50
20	1	3.94	4.25 (0.08)	4.03	4.07	4.28 (0.08)	4.18
	10	8.35	6.81 (0.11)	7.59	7.55	6.55 (0.09)	7.27
	20	13.00	8.72 (0.11)	10.69	13.85	8.63 (0.10)	11.13
	50	27.58	12.81 (0.14)	16.05	25.91	11.88 (0.12)	16.27
30	1	3.86	4.00 (0.04)	3.92	4.10	4.42 (0.04)	4.53
	10	7.73	6.07 (0.05)	7.49	8.11	6.73 (0.05)	8.21
	20	13.31	8.29 (0.07)	12.11	12.35	7.90 (0.06)	11.33
	50	26.89	12.67 (0.08)	21.09	26.70	12.75 (0.08)	21.01
50	1	4.16	3.89 (0.02)	4.28	3.92	3.98 (0.02)	4.35
	10	7.44	5.55 (0.02)	7.58	8.70	6.91 (0.02)	9.90
	20	13.14	8.19 (0.03)	12.98	12.78	8.21 (0.03)	13.58
	50	28.40	13.87 (0.05)	26.65	31.14	15.29 (0.04)	28.83
100	1	4.02	3.82 (0.02)	4.15	3.74	4.39 (0.01)	4.74
	10	7.71	6.04 (0.02)	7.93	8.03	6.64 (0.01)	8.81
	20	12.21	7.93 (0.02)	12.14	12.77	8.32 (0.02)	13.47
	50	26.33	13.52 (0.03)	26.38	26.98	12.84 (0.02)	27.20

Next we turn our attention to the empirically determined weight  $\hat{\gamma}$  in the weighted entropy objective function. For each experiment, the average  $\hat{\gamma}$  is reported in parenthesis for the D-GME estimator. We observe two note-worthy features. First,  $\hat{\gamma}$  increases generally with  $\mu$ . Recalls that  $\gamma$  is the weight placed on the entropy of the disturbance distributions. Thus the more severe the “ill-posed” problem is, the larger is the weight selected by the LSCV. In other words, the data-driven method automatically relaxes the model consistency constraints when the underlying linear inverse problem associated with the OLS becomes problematic. Second,  $\hat{\gamma}$  decreases with the width of the coefficient support across all condition numbers. Intuitively, the wider is the coefficient support, the weaker are the restrictions imposed by the GME estimation procedure. Correspondingly, the smaller is the need to regulate the entropy, or uncertainty, of the disturbance distribution.<sup>3</sup>

Lastly, we note that the overall performance of the estimators in questions remains quite stable when the sample size is increased from 30 to 50. The average ratios, across all cases, in the MSE between the D-GME and GME are 0.834 and 0.828 respectively for  $n = 30$  and 50. It is well-known that data-driven methods normally require a sizeable sample to attain its theoretical advantages. Nonetheless our results demonstrate that the D-GME can outperform the GME with quite small sample sizes under various scenarios.

## 4.2 Regressions with non-normal errors

Next we investigate the performance of the proposed estimator when the errors are generated from non-normal distributions. Using the same sample design outlined above, we generated the errors instead from a  $\chi^2(4)$  and a  $t(3)$  distribution. The  $\chi^2(4)$  errors were centered by subtracting the mean (4), and all drawings were scaled to have unit variance by dividing each by the associated standard deviation ( $\sqrt{3}$  and  $\sqrt{8}$  respectively). Under a  $\chi^2$  error distribution, we set the disturbance support to  $\mathbf{V} = [-\hat{\sigma}, -\hat{\sigma}/2, 0, \hat{\sigma}, 2\hat{\sigma}] \times 3$  to account for the skewness of the  $\chi^2$  distribution. When the disturbance terms were generated from the  $t$  distribution, instead of using the  $3\sigma$  rule, we set  $\mathbf{V} = [-\hat{\sigma}, -\hat{\sigma}/2, 0, \hat{\sigma}/2, \hat{\sigma}] \times 5$  to account

---

<sup>3</sup>Recall that the GME estimator implicitly assumes a uniform prior distribution for the error support. Since we use a symmetric error support centered at zero, a uniform distribution over this support leads to a zero disturbance. A wider coefficient support means less restrictive constraints on  $\beta$  and thus the a smaller  $e = y - x\beta$  in absolute value. With the error terms more likely to be close to zero, the need to regulate the entropy of error term distributions is less.

for the fat-tailedness of the error distribution.

The estimation results for the  $\chi^2$  case are reported in Table 2. The overall pattern is similar to that with normal errors. With narrow coefficient supports, the GME has a smaller MSE than the D-GME. On the other hand, when  $z \geq 30$ , the D-GME outperforms the GME, and the performance gap generally increases with the condition number. The performance is quite similar between  $n = 30$  and  $n = 50$ . On the other hand, the average  $\gamma$  is larger than that for normal error case, indicating a heavier penalty for the uncertainty in error distribution when it is non-normal.

Lastly, Table 3 reports the results for  $t$  error distributions. With  $n = 30$ , the overall pattern is again similar to the first two cases. A noteworthy difference is that the MSEs for the two GME estimators increase substantially when the sample size is raised to 50. In contrast, the OLS does not seem to be affected by the change in sample size. Nonetheless, except when the condition number is small or a very wide coefficient support is used ( $z = 100$ ), the D-GME still outperforms the OLS. We also note that the weight  $\gamma$  is larger than that chosen under normal errors.

## 5 Concluding Remarks

The Generalized Maximum Entropy (GME) estimator is a robust estimator that is resistant to multicollinearity. Like other robust estimators, the estimator requires specification of some “tuning” parameters. In particular, it requires users to specify discrete supports for the coefficients and disturbance. In a more general weighted GME framework, one also needs to specify a weight that determines the relative weight placed on the entropy of the coefficient distributions and error distributions. Although the specifications of the coefficient and error support can be guided by non-sample information and preliminary consistent estimates, there is no clear guidance for choosing the weight in a weighted GME estimation.

In this study, we have presented a data-driven weighted GME estimator (D-GME). The conventional GME estimator places equal weights on the entropy of the coefficient distribution and disturbance distribution. Instead, we proposed to use the method of least squares cross validation to select this weight in a data-driven manner. We demonstrate numerically that the D-GME provides superior performance under various scenarios. Investigation on

Table 2: MSE of regressions with  $\chi^2(4)$  errors

$z$	$\kappa$	$n = 30$			$n = 50$		
		OLS	D-GME( $\hat{\gamma}$ )	GME	OLS	D-GME( $\hat{\gamma}$ )	GME
10	1	3.98	3.41 (0.42)	2.81	4.08	3.43 (0.44)	2.73
	10	8.07	5.64 (0.41)	4.18	7.75	5.06 (0.45)	3.67
	20	12.17	6.50 (0.40)	4.99	12.16	5.88 (0.43)	4.69
	50	29.33	7.70 (0.44)	5.66	27.46	7.09 (0.44)	5.55
20	1	4.03	3.54 (0.18)	3.18	3.92	3.39 (0.19)	2.92
	10	8.53	5.66 (0.19)	5.40	7.79	5.26 (0.20)	4.83
	20	12.12	7.27 (0.18)	6.76	12.58	7.16 (0.21)	6.08
	50	28.15	11.22 (0.22)	9.23	28.61	10.82 (0.22)	8.29
30	1	4.21	3.62 (0.11)	3.44	4.34	3.60 (0.10)	3.46
	10	8.09	5.66 (0.12)	5.82	8.41	5.85 (0.12)	6.05
	20	13.60	7.58 (0.12)	9.00	13.76	7.91 (0.13)	8.76
	50	26.99	11.45 (0.15)	13.06	25.87	10.95 (0.15)	11.57
50	1	4.11	3.46 (0.05)	3.35	3.90	3.23 (0.04)	3.14
	10	8.58	5.48 (0.05)	6.79	7.76	5.36 (0.05)	6.01
	20	13.15	6.97 (0.06)	10.27	13.34	7.39 (0.06)	9.89
	50	27.81	11.16 (0.08)	17.84	28.01	11.28 (0.08)	17.47
100	1	3.90	3.02 (0.02)	3.45	4.00	2.94 (0.01)	3.25
	10	7.55	4.70 (0.02)	6.35	8.54	5.12 (0.02)	6.65
	20	13.62	6.86 (0.02)	11.49	13.52	6.94 (0.03)	10.81
	50	28.27	11.29 (0.04)	22.14	29.26	12.03 (0.04)	21.93

combining the data-driven selection of the weight parameter and automatic specification of the discrete supports for the coefficients and errors to achieve adaptiveness and further improvement shall be of interest for future studies.

## References

- [1] Ciszár, I. (1991) Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics*. 19, 2032-2066.
- [2] Golan, A., G. Judge, D. Miller. (1996) *Maximum entropy econometrics: robust estimation with limited data*. 1996. John Wiley & Sons, Chichester.
- [3] Hall, P. (1984) Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*. 11, 1156-1174.

Table 3: MSE of regressions with  $t(3)$  errors

$z$	$\kappa$	$n = 30$			$n = 50$		
		OLS	D-GME( $\hat{\gamma}$ )	GME	OLS	D-GME( $\hat{\gamma}$ )	GME
10	1	3.84	4.09 (0.47)	3.61	3.36	4.71 (0.44)	4.02
	10	7.01	5.68 (0.44)	4.98	6.35	6.47 (0.42)	5.49
	20	10.65	6.41 (0.44)	5.47	10.31	7.31 (0.42)	5.91
	50	20.90	7.03 (0.44)	6.28	20.58	7.98 (0.43)	6.93
20	1	3.84	4.11 (0.21)	4.54	3.70	6.91 (0.18)	6.92
	10	7.46	6.07 (0.21)	6.30	7.18	8.12 (0.20)	8.95
	20	11.02	6.83 (0.21)	7.48	11.11	9.70 (0.20)	10.69
	50	24.31	9.98 (0.24)	9.37	24.20	11.64 (0.22)	12.03
30	1	4.60	4.33 (0.21)	5.39	3.59	6.23 (0.18)	6.50
	10	7.83	5.91 (0.21)	7.82	7.19	9.00 (0.20)	10.77
	20	12.03	7.73 (0.21)	9.84	11.38	9.71 (0.20)	13.32
	50	23.92	10.75 (0.24)	13.89	25.65	15.82 (0.22)	19.98
50	1	5.64	5.29 (0.05)	7.15	3.78	8.32 (0.04)	8.76
	10	7.86	5.74 (0.05)	9.28	7.80	12.54 (0.05)	17.37
	20	13.41	8.18 (0.06)	13.91	11.89	9.69 (0.05)	16.53
	50	24.07	11.05 (0.08)	21.30	25.82	20.27 (0.08)	32.39
100	1	3.96	4.01 (0.02)	5.19	3.80	11.75 (0.01)	12.88
	10	6.92	5.33 (0.02)	8.68	8.31	14.55 (0.02)	19.93
	20	14.25	7.78 (0.02)	17.64	11.64	17.64 (0.02)	25.50
	50	32.09	14.24 (0.03)	33.82	28.35	41.85 (0.04)	59.32

- [4] Hall, P., Marron J.S. (1987) Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields.* 74, 567-581.
- [5] Jaynes, E.T. (1957) *Information theory and statistical mechanics.* *Physics Review.* 106, 620-630.
- [6] Jaynes, E.T. (1984) Prior information and ambiguity in inverse problems. In D.W. McLaughlin (Ed.) *Inverse Problems*, pp. 151-166, SIAM Proceedings, American Mathematical Society, Providence, RI.
- [7] Kullback, J. (1959) *Information theory and statistics.* John Wiley, New York.
- [8] Levine, R.D. (1980) An information theoretical approach to inversion problems. *Journal of Physics A.* 13, 91-108.

- [9] Levine, r.D., Tribus, M. (1979) The Maximum Entropy Formalism. MIT Press, Cambridge.
- [10] Shannon, C.E. (1948) A mahtematical thoery of communication. Bell System Technical Journal. 27, 379-423.
- [11] Shore, J.E., Johnson, R.W. (1980) Axiomatic derivation of the principle of maximum entropy nad the principel of minimum cross-entropy. IEEE Transactions on Information Theory. 26, 26-37.
- [12] Skiling, J. (1989) The axioms of maximum entropy. In J. Skilling (Ed.) Maximum Entropy and Bayesian Methods in Science and Engineering, pp. 173-187, Kluwer, Dordrecht.
- [13] Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates. Annals of Statistics. 12, 1285-1297.