# INFORMATION-THEORETIC MODEL SELECTION AND ESTIMATION FOR INTERVAL DATA

AMOS GOLAN

DEPARTMENT OF ECONOMICS AND INFO-METRICS INSTITUTE, AMERICAN UNIVERSITY

TUAL TUANG

DEPARTMENT OF ECONOMICS, AMERICAN UNIVERSITY

AMAN ULLAH

DEPARTMENT OF ECONOMICS, UNIVERSITY OF CALIFORNIA AT RIVERSIDE

JULY, 2015

Abstract.

We develop an efficient information-theoretic estimator for analyzing interval-valued, and symbolic, data. Rather than applying the traditional least squares or likelihood methods to estimate some moments of the intervals (as often done), we use the complete information in the sample and identify the best model and related parameters that are consistent with the data generating process. It is an iterative approach. Our information-theoretic estimator imposes minimal statistical assumptions on the underlying distribution. We provide a large number of sampling experiments as well as a few empirical examples.

**Key Words:** Entropy, Information, Information Theory, Interval-Valued Data, Iterative Process

## 1. INTRODUCTION

Imagine the case where all of the observed information is in terms of intervals over some ranges. Examples include regression models where some set of known variables $\mathbf{X}$ has a causal effect on the observed dependent variable $\mathbf{y}$ and both the $\mathbf{y}$'s and the $\mathbf{X}$'s are observed as intervals. These types of data are very common across many disciplines. Examples include financial and weather data, histograms and more. The main issue for the researcher, or for the policy analyst, is how to identify the correct underlying structure of the model generating these interval data, and then, how policy implications or forecasting can be made. These types of problems are tough to analyze even for a large data set. A main difficulty is in separating the noise from the signal in the observed interval information such that the correct model (or class of models) is identified. Stated differently, facing interval data, a major concern is figuring out whether each point in the interval is generated via the same underlying process, and whether the impact of each point, within the interval of the independent variable, on the dependent variable is the same. Is there a single underlying process and if so what is it? If it is not a single process, can we identify the most probable processes? In this paper we develop an iterative Information-Theoretic (IT) method for simultaneously identifying the most appropriate model (out of a class of possible models) and for estimating its parameters for interval data problems. This is a flexible and efficient approach that imposes minimal assumptions on the underlying statistical process. Our proposed approach works well for small, large or ill-behaved data.

Generally speaking, the invention of high-performance computers together with the emergence of big data due to sophisticated data collection technologies necessitate researchers to think about processing, storing, and analyzing information that is often observed in terms of intervals, histograms or other types of symbolic style data. This led to much interest in the study of these types of data with significant innovations in the last several decades. However, though there is an increasing body of research that concentrates on estimation and inference of these types of data, most of the symbolic data research concentrates on computational statistics and data mining techniques. Our method falls within the former.

One of the more frequently studied symbolic type data is interval-valued data. In such data, instead of observing a single value for an observation, each observation is in terms of some interval (Billard and Diday, 2012). Much of the observed interval-valued data is due to aggregation of the original detailed data. Examples include intraday stock prices and returns which are routinely aggregated to obtain daily interval within the minimal and upper daily prices. Similarly, daily closing prices of stocks can be aggregated to form an interval with bounds of the lowest and highest weekly prices. At times, however, the observed intervals may be simply due to the intrinsic nature of the observable information. For example, an individual's blood pressure is measured as an interval of systolic and diastolic pressures. Other examples include risk analysis or weather prediction (e.g., high-frequency financial data, weekly

oil prices, and temperature patterns) where intervals are the natural state of each observation. Other examples of symbolic data include histogram-valued data and modal-valued data. We concentrate here on interval data.

Generally speaking, the classical regression methods fail to account for the internal variations within an observation. However, most existing interval-valued data estimation methods build on the more traditional methods for inferring separately the lower and upper bounds of the interval, or for inferring (separately) the center and ranges of the intervals. Then, these independent estimates are combined in a suitable way for prediction. Though these methods work well under some conditions, in practice these conditions don't hold. These conditions include the assumption that there is a unique relationship between the independent and dependent intervals across all points within the interval, and that the bounds of the intervals are perfectly known. Further, it is also often assumed that recovering the moments of the intervals is sufficient for uncovering the possibly more complex relationship among the intervals. Billard and Diday (2000) developed one of the first models to extend the classical method to deal with interval-valued data. Their innovation was to fit a regression line through the center points of the observations and use the estimated parameters, along with the lower and upper bounds of the explanatory variables, to predict the lower and upper bounds of the response variable. Though rather innovative, their model ignores the internal variations of observations as well as the ranges of the intervals. To resolve this issue, Lima Neto et. al. (2004, 2008) proposed running two separate regression models for the center points and ranges of the intervals and then, predict the centers and ranges, respectively. In line with that approach, Billard and Diday (2002) proposed fitting two separate regressions of the (minimum and maximum of) the intervals. A major drawback of these approaches is that the predicted lower bound of the response variable can be larger than the predicted upper bound. This is particularly true if the estimated coefficients are negative. Lima Neto and de Carvalho (2010) suggested imposing an extra constraint (that, unfortunately, may be inconsistent with the data generating process) that forces the estimated coefficients to be non-negative and guarantees an artificial "consistent" predicted lower and upper bounds.

Instead of transforming symbolic data to a classical data in order to use the classical estimator, Xu (2010) and Xu and Billard (2012) recently developed a "symbolic variation" least squares estimator using symbolic sample variance and covariance functions developed by Bertrand and Goupil (2000) and Billard (2007, 2008). In addition to the above least squares estimators, a maximum likelihood estimator (MLE) using a symbolic likelihood function was introduced by Le-Rademacher and Billard (2011), Xu (2010) and Xu and Billard (2012).

The information used for the inference in all of these studies consists solely of the first few moments of the intervals or on the minimally and maximally observed values. This means that not all of the observed information is used, leading to a decrease in efficiency and accuracy. Further, these methods

typically require (and assume) uniform or other distribution of points within the interval bounds. These are tough assumptions that, most often, cannot be verified. Instead of innovating on the above approaches we take a different route here. First, we relax these distributional assumptions on the within-interval behavior. Second, we use all of the observed information in the sample. But these two relaxations mean that the underlying problem is under-determined; there are infinitely many models that are consistent with the observed intervals. Therefore, we use all of the observed information, within an optimization framework, to identify the most probable (causal) model (or set of models) that is consistent with our observations. The most probable model is identified by an entropy measure.

The method we develop is an iterative Information-Theoretic, Generalized Maximum Entropy (GME) estimator (see Golan et. al.,1996). The estimation part (GME) has its roots in the intersection of information theory and statistical inference. It uses Shannon's entropy measure (Shannon, 1948) in conjunction with the principle of Maximum Entropy (Jaynes, 1957). It is easy to implement and program. Our proposed iterative process is as follows. First, we divide the observed interval-valued data (for each variable) into a number of mutually exhaustive, equally-spaced discrete sub-intervals. Second, we use a GME regression model to estimate the parameters of each one of the possible models (all possible combinations among the independent and the dependent sub-intervals). Third, using these inferred results we use our entropy measure to distinguish the "correct" (or best) model from the rest. Thus, our method identifies the best model and simultaneously infers the model's parameters. We use our estimates to predict the distribution of the response variable. This allows us to estimate and predict the full interval (distribution) of the response variable rather than just the lower and upper bounds (or means) as are captured via the other approaches.

Section 2 briefly discusses the interval-valued data and briefly summarizes some of the commonly used regression methods for analyzing these data. Section 3 provides the details of our iterative framework and discusses our Information-Theoretic method. In section 4 we present results from a large number of sampling experiments. In these experiments we contrast our method with its competitors. We also extend our proposed iterative framework to include other traditional estimators (instead of the GME) and compare with our method. In section 5 we provide two empirical examples. We conclude in Section 6.

## 2. CURRENT METHODS

This section briefly summarizes the nature of interval-valued data and discusses some of the existing (linear) estimation methods, including the Center Method (CM), the Center and Range Method (CRM), the Bivariate Center and Range Method (BCRM), the Symbolic Covariance Method (SCM), and others.

Following on Bertrand and Goupil's (2000), suppose we observe a sample of $N$ entities ($i = 1, ..., N$) of a random variable $\mathbf{X}$. For each observation $i$, there is an interval data point $[X]_i \equiv [X_i^L, X_i^U]$ and

$X_i^L \leq X_i^U$. Further, it is also assumed that values in the given interval ($X_i^L \leq x \leq X_i^U$) are uniformly distributed within the interval while each observation have the same probability $\frac{1}{N}$ of being observed. Then, the empirical density function $f_X(x)$ is a combination of $N$ uniform distributions:

$$\hat{f}_X(x) = \frac{1}{N} \sum_{i:x \in [x]_i} \frac{I(x \in [x]_i)}{\parallel [x]_i \parallel} = \frac{1}{N} \sum_{i:x \in [x]_i} \frac{1}{x_i^U - x_i^L}, x \in \Re \tag{1}$$

where $I(x \in [x]_i)$ is the indicator function of whether $x$ is inside (or not inside) the interval $[x]_i$ and $\parallel [x]_i \parallel$ is the length of that interval.

Based on the empirical density function (1), the sample mean and sample variance are:

$$\begin{aligned} \hat{E}(X) = \bar{X} = \int_{-\infty}^{\infty} x f(x) dx &= \frac{1}{N} \sum_{i:x \in [x]_i} \frac{1}{x_i^U - x_i^L} \int_{x_i^L}^{x_i^U} x dx \\ &= \frac{1}{2N} \sum_i (x_i^U + x_i^L) = \frac{1}{N} \sum_i x_i^C \end{aligned} \tag{2}$$

where $x_i^C$ is the center point of the interval $[x]_i$, and

$$\begin{aligned} S_X^2 = \int_{-\infty}^{\infty} (x - \bar{X})^2 f(x) dx &= \left( \int_{-\infty}^{\infty} x^2 f(x) dx \right) - \bar{X}^2 \\ = \frac{1}{3N} \sum_i ({x_i^U}^2 + x_i^U x_i^L + {x_i^L}^2) &- \frac{1}{4N^2} \left[ \sum_i (x_i^U + x_i^L) \right]^2 \end{aligned} \tag{3}$$

Based on the above formulas, Arroyo, Gonzalez-Rivera, and Mate (2010) noted that the sample mean of an interval random variable is the mean of all the center points in the sample. The sample variance captures not only the variations of the centers across observations, but also the variations within the interval. Naturally, when $X_i^L = X_i^U$, the interval data collapses to the classical single-valued data.

Finally, Billard (2007,2008) introduced a "symbolic" covariance function for analyzing "symbolic" regression analysis, while, Le-Rademacher and Billard (2011), Xu (2010) and Xu and Billard (2012) developed a "symbolic" likelihood function for interval-valued data.

We now summarize, very briefly, some of the currently used methods. Xu (2010) and Ahn et. al. (2012) discuss more in details about these existing methods. Following Ahn et. al. (2012)'s notations, let $X_1, ..., X_K$ be $K$ explanatory variables and $Y$ be the response variable. Assumes that $X_{ik} = [X_{ik}^L, X_{ik}^U]$ with $X_{ik}^L \leq X_{ik}^U$ and $Y_i = [Y_i^L, Y_i^U]$ with $Y_i^L \leq Y_i^U$, for $i = 1, ..., N$ and $k = 1, ..., K$. Consider the following linear regression model:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{4}$$

where $\mathbf{Y} = (Y_1, ..., Y_N)', \mathbf{X} = (X_1, ..., X_N)', X_i = (1, X_{i1}, ..., X_{iK})'$ for $i = 1, ..., N, \beta = (\beta_0, \beta_1, ..., \beta_K)', \epsilon = (\epsilon_1, ..., \epsilon_N)'$, $\epsilon_i \sim N(0, \sigma^2)$, and "$'$" stands for "transpose".

Billard and Diday (2000) proposed the center method (CM), which is one of the first major works in analyzing interval-valued data. The CM simply fits a linear regression line to the center points of the intervals. Let $X_1^c, ..., X_K^c$ be the center points of the intervals of explanatory variables $X_1, ..., X_K$ and $Y^c$ be the center point of a response variable $Y$. The CM essentially transforms the interval linear regression model of (4) into the following center-points model:

$$\mathbf{Y}^c = \mathbf{X}^c \beta^c + \epsilon^c \tag{5}$$

where $Y^c = (Y_1^c, ..., Y_N^c)', X^c = (X_1^c, ..., X_N^c)', X_i^c = (1, X_{i1}^c, ..., X_{iK}^c)'$ for $i = 1, ..., N, \beta^c = (\beta_0^c, \beta_1^c, ..., \beta_K^c)'$, and $\epsilon^c = (\epsilon_1^c, ..., \epsilon_N^c)'$.

The estimator $\hat{\beta}^c$ is estimated via the usual least squares method:

$$\min_{\hat{\beta}^c} = \sum_i \hat{\epsilon}_i^{c\,2} = \sum_i (Y_i^c - \hat{\beta}^{c\prime} X_i^c)^2 \quad \& \quad \hat{\beta}^c = (\mathbf{X}^{c\prime}\mathbf{X}^c)^{-1}\mathbf{X}^{c\prime}\mathbf{Y}^c \tag{6}$$

and standard statistical properties are readily adopted assuming the standard distributional assumptions on the error-term. Prediction of $\hat{Y} = [\hat{Y}^L, \hat{Y}^U]$ is given as:

$$\hat{Y}^m = x_0^m \hat{\beta}^c, \quad m = L, U \tag{7}$$

for a new observation $(x_0^L, x_0^U)$. Xu (2010) and Xu and Billard (2012) point out that a lower bound predicted response variable can be higher than an upper bound, and suggest the modified prediction is presented as:

$$\hat{Y}^L = \min(x_0^L \hat{\beta}^c, x_0^U \hat{\beta}^c) \quad \& \quad \hat{Y}^U = \max(x_0^L \hat{\beta}^c, x_0^U \hat{\beta}^c). \tag{8}$$

Although the CM uses the ranges of predictors for prediction of the lower and upper bounds, such bounds are ignored in estimating the parameters. That is, the variations within observations are not utilized.

In an attempt to capture the internal variations as well as variations across observations, Lima Neto et. al. (2004, 2008) propose the center and range method (CRM) that transforms the interval-valued data to single point centers and ranges of the interval variables and then regressing the center points and ranges separately. First, the CRM keeps the same model as in (5) for the center points. Then, obtain the ranges, $X_{ik}^r = (X_{ik}^U - X_k^L)$ and $Y_i^r = (Y_i^U - Y_i^L)$, using similar model. Let $X_1^r, ..., X_K^r$ be the $K$ ranges of the intervals of $X_1, ..., X_K$ and $Y^r$ be the range of Y and is given by the following:

$$\mathbf{Y}^r = \mathbf{X}^r \beta^r + \epsilon^r \tag{9}$$

where $Y^r = (Y_1^r, ..., Y_N^r)', X^r = (X_1^r, ..., X_N^r)', X_i^r = (1, X_{i1}^r, ..., X_{iK}^r)'$ for $i = 1, ..., N, \beta^r = (\beta_0^r, \beta_1^r, ..., \beta_K^r)'$, and $\epsilon^r = (\epsilon_1^r, ..., \epsilon_N^r)'$. Both $\hat{\beta}^c$ and $\hat{\beta}^r$ are estimated by minimizing the following objective function:

$$\min_{\hat{\beta}^c, \hat{\beta}^r} = \sum_i (\hat{\epsilon_i^c}^2 + \hat{\epsilon_i^r}^2). \tag{10}$$

They essentially perform two separate minimizations, one for the centers and another of the ranges, assuming that mid-points and ranges are independent. However, Ahn et. al. (2012) asserts that such assumption may not be true in general. Although CRM captures the interval variations through the ranges, Ahn et. al. (2012) also point out that it is not clear how these variations are transferred to the estimated coefficients. Prediction of $\hat{Y} = [\hat{Y}_L, \hat{Y}_U]$ is given as

$$\hat{Y}^L = \hat{Y}^c - \frac{\hat{Y}^r}{2} \quad \& \quad \hat{Y}^U = \hat{Y}^c + \frac{\hat{Y}^r}{2} \tag{11}$$

where $\hat{Y}^c$ and $\hat{Y}^r$ are predicted values from (5) and (9).

Similar to CRM, Billard and Diday (2007) proposes a bivariate center and range method (BCRM) which essentially utilized both the centers and ranges as predictors in the models, simultaneously. The center and range models can be presented as follows:

$$\mathbf{Y}^c = \mathbf{X}^{cr}\beta^c + \epsilon^c \quad \& \quad \mathbf{Y}^r = \mathbf{X}^{cr}\beta^r + \epsilon^r \tag{12}$$

where $X^{cr} = (X_1^{cr}, ..., X_N^{cr})', X_i^{cr} = (1, X_{i1}^c, ..., X_{iK}^c, X_{i1}^r, ..., X_{iK}^r)'$ for $i = 1, ..., N$. Potential problems arise in when all the explanatory intervals and response variables have the same centers and ranges. In this case, the previous methods will not work.

In addition to the previous approaches that utilize the classical regression estimators, Xu (2010) and Xu and Billard (2012) proposed a symbolic covariance method (SCM) using the "symbolic" covariance function introduced previously by Billard (2007,2008). Their model is

$$\mathbf{Y} - \bar{\mathbf{Y}} = (\mathbf{X} - \bar{\mathbf{X}})\beta + \epsilon \tag{13}$$

where $\hat{\beta} = \{(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})\}^{-1}(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{Y} - \bar{\mathbf{Y}}) = \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$, $\mathbf{S}_{XX}$ is the symbolic sample variance-covariance matrix of the predictors and $\mathbf{S}_{XY}$ is the vector of the symbolic sample covariance between Y and the predictors. As in Billard (2007,2008), the symbolic sample covariance between interval-valued variables $X_j$ and $X_k$ is defined as

$$\begin{aligned} Cov(X_j, X_k) = (6N)^{-1} \sum_{i=1}^{N} [2(X_{ij}^L - \bar{X}_j)(X_{ik}^L - \bar{X}_k) + (X_{ij}^L - \bar{X}_j)(X_{ik}^U - \bar{X}_k) \\ + (X_{ij}^U - \bar{X}_j)(X_{ik}^L - \bar{X}_k) + 2(X_{ij}^U - \bar{X}_j)(X_{ik}^U - \bar{X}_k)] \end{aligned} \tag{14}$$

where the symbolic sample mean is defined as in Bertrand and Goupil (2000).

Recently, Ahn et. al. (2012) proposed a Monte-Carlo method (MCM) that (i) generate a large number of samples by randomly selecting, uniformly, a single-valued data point for each observed intervals, (ii) fit a classical linear regression model on each single-valued sample, (iii) calculate the mean estimated coefficients over the fitted models. Then, use this mean of estimated coefficients to predict the response variable.

Using interval-valued data is gaining some tractions in economics and econometrics. Examples include forecasting stock prices and returns with interval and histogram-valued times series data (Arroyo, Gonzalez-Rivera, and Mate (2010), Arroyo and Gonzalez-Rivera (2012), and Arroyo, Gonzalez-Rivera, Mate, and San Roque (2011)). Other examples include the work of Hea et. al. (2011) on forecasting interval-valued crude oil prices with autoregressive conditional interval models, as well as the work of Manski and Tamer (2003), and Magnac and Maurin (2007, 2008) on interval data to deal with partial identification issues.

## 3. AN ITERATIVE INFORMATION-THEORETIC FRAMEWORK FOR INTERVAL ESTIMATION

In this section, we develop an iterative Information-Theoretic Method for estimating interval-valued data. Our method is an iterative version of the Generalized Maximum Entropy (GME) estimator (Golan et. al., 1996). First, we develop our iterative framework. Then, we briefly summarize GME estimator for the linear regression.

### 3.1. **The Iterative Framework**

The proposed iterative framework is simple and easy to implement. First, we relax distributional assumptions on the within-interval behavior. Second, we divide both interval-valued response and explanatory variables into a number of mutually exhaustive, equally-spaced sub-intervals and use these observed information in the sample by iteratively fitting a GME regression model through all possible combinations of the observed sub-intervals. Using the estimated entropy measure, we identify the most probable (causal) model (or set of models) that is consistent with our observations. The most probable model is a model with the largest estimated entropy measure.

To introduce the idea, consider first the simplest case. Suppose that there is only one explanatory variable $\mathbf{X}_i = [X_i^L, X_i^U]$ and one response variable $\mathbf{Y}_i = [Y_i^L, Y_i^U]$. We can divide $\mathbf{X}_i$ and $\mathbf{Y}_i$ into $M-1$ sub-intervals of equal size to get $X_i^m = [X_i^1, X_i^2, ..., X_i^M]$ and $Y_i^m = [Y_i^1, Y_i^2, ..., Y_i^M]$. $M$ represents the endpoints of sub-intervals for each $\mathbf{X}_i$ and $\mathbf{Y}_i$, or we can view it for now as $M$ observed points for each observation $(\mathbf{X}_i, \mathbf{Y}_i)$. For example, for observation $i = 1$ and $M = 5$ sub-intervals with one interval-valued regressor $\mathbf{X}_i$, observing $\mathbf{X}_1 \equiv [X_1^1, X_1^2, X_1^3, X_1^4, X_1^5] = [2, 4, 6, 8, 10]$ and $\mathbf{Y}_1 \equiv [Y_1^1, Y_1^2, Y_1^3, Y_1^4, Y_1^5] = [6, 10, 14, 18, 22]$ would correspond to $Y_i^m = 2 + 2X_i^m$. For $K = 2$ with the

underlying model $Y_i^m = 2 + 2X_{i1}^m - 3X_{i2}^m$, we observe $\mathbf{X}_{11} \equiv [X_{11}^1, X_{11}^2, X_{11}^3, X_{11}^4, X_{11}^5] = [2, 4, 6, 8, 10]$, $\mathbf{X}_{12} \equiv [X_{12}^1, X_{12}^2, X_{12}^3, X_{12}^4, X_{12}^5] = [3, 5, 7, 9, 11]$ and $\mathbf{Y}_1 \equiv [Y_1^1, Y_1^2, Y_1^3, Y_1^4, Y_1^5] = [-3, -5, -7, -9, -11]$. The two examples assume that each sub-interval $\mathbf{m}(= m_1, m_2)$ of $\mathbf{X}$ causes linearly the $m$th sub-interval of $\mathbf{Y}$ and the impacts are the same across different points within the intervals.

But now, to be more realistic, assume we know that $\mathbf{X}$ causes $\mathbf{Y}$ linearly but we do not know the correct model. We do not know if each point within the interval of $\mathbf{X}$ affects $\mathbf{Y}$ in the same way. Maybe, for example, points on the lower part of $\mathbf{X}$ affect $\mathbf{Y}$ in a different way than points around the mean or the upper portion of the interval. More precisely, we do not know which combination of intervals (or points) $\mathbf{m}(= m_1, m_2, ..., m_K)$ of $K$ explanatory variables $\mathbf{X}_i$ correspond to which interval (or point) $m$ of $Y_i^m$, where $m_k$ is an interval (or point) of $\mathbf{X}_{ik}$ and can takes the value of $m_k = 1, ..., M$. Given this complex setting, we want to infer the relationship between each one of the $\mathbf{X}$ sub-interval and the $\mathbf{Y}$ sub-interval. To do so with minimal assumptions on the underlying distribution, we apply the GME for all possible combination of sub-intervals within the interval-valued random variables $\mathbf{X}$ and $\mathbf{Y}$.

[*Insert Figure 1 here*]

Consider, for example, $M = 5$ and $K = 1$ as shown in Figure 1. Then, using $M \times M^K$ combinations of pairs such as $(X_1^1, Y^1)$, $(X_1^1, Y^2)$,...,$(X_1^1, Y^5)$, $(X_1^2, Y^1)$, $(X_1^2, Y^2)$,...,$(X_1^5, Y^5)$, we obtain 25 sets of estimated coefficients via a linear GME estimator, while $K = 2$ produces 125 sets of estimated coefficients. Iteratively fitting the linear GME model for the above combination assumes that the ordering of sub-intervals within $Y_i^m$ is known or does not matter for the analysis. There could be cases in which more complicated orderings or permutations could be required.

Using these estimated regression coefficients together with the empirical distribution of total entropy values for each one of the models, we are able to distinguish the best model. It is the one with the largest entropy. Thus, we not only identifies the best model but also obtain the inferred model's parameters. Generally speaking, the choice for $M$ is an empirical issue. Since the number of iterations required for our iteration approach increases multiplitively with $M$ and the number of regressors, $K$, one wants relatively low $M$. It mainly depends on the trade-off between the amount of variations obtained by increasing $M$ and the increase in iterations required. In here, we use $M = 5$ for both experiment and empirical sections.

### 3.2. **An Information-Theoretic Estimator - A Brief Summary**

Consider the linear regression model with $N$ observations and $K$ explanatory variables:

$$\mathbf{Y} = X\beta + \epsilon \tag{15}$$

where $\mathbf{Y}$ is a $N$-dimensional vector of observed random variable, $\mathbf{X}$ is a $N \times K$ matrix of regressors, $X_{ik} = [X_{ik}^L, X_{ik}^U]$ with $X_{ik}^L = X_{ik}^U$ is degenerated to a point value (instead of interval as in (4)), $\beta$ is $K$-dimensional vector of the unknown coefficients, and $\epsilon$ is a $N$-dimensional vector of unobserved and unobservable random errors. Golan et. el. (1996) proposed an Information-Theoretic (IT) estimator, which is a member of the IT class of estimators (eg., Golan 2008, Judge and Mittelhammer 2011). This estimator, called GME, uses minimal distributional assumptions and proved to perform well, relative to other methods, especially for small, ill-behaved and other complex data. Rather than treating the unknown quantities as point estimates, under the GME the complete probability distribution of each unknown quantity is estimated. They reformulated $\beta$ and $\epsilon$ such that

$$Y_i = \sum_k \sum_s z_{ks} p_{ks} x_{ik} + \sum_j v_j w_{ij} \tag{16}$$

where $\mathbf{z}$ and $\mathbf{v}$ are the support spaces for the signal $\beta$ and error $\epsilon$ respectively, $\mathbf{p}_k$ is an $S$-dimensional normalized probability distribution for each $\beta_k$, and $\mathbf{w}$ is a normalized probability distribution for each $\epsilon_i$. The support $\mathbf{z}_k$ may be different for each $\beta_k$ while $\mathbf{v}$ is symmetric around zero and similar for all $\epsilon_N$. Unless more information is known, each $\mathbf{z}_k$ should be specified to be symmetric about zero. For example, for $S = 3$, $\mathbf{z}_k = (-C, 0, C)$ for some large $C$. The bounds for the $\mathbf{v}$'s are +/- $3\sigma_y$, where $\sigma_y$ is the sample standard deviation. See Golan, et. al. (1996) or Golan (2008) for further discussions and examples.

Having reparameterized the model, it is clear that the number of unknown $\mathbf{p}$'s and $\mathbf{w}$'s exceeds the number of observable information. The problem is under-determined. To solve it, we follow on the classical Maximum Entropy (ME) formalism (Jaynes 1957). Let $H(\mathbf{p})$ and $H(\mathbf{w})$ be Shannon's entropies for $\mathbf{p}$ and $\mathbf{w}$, respectively, then the generalized GME is just

$$\max_{\{\mathbf{p},\mathbf{w}\}} \left\{ H(\mathbf{p},\mathbf{w}) = H(\mathbf{p}) + H(\mathbf{w}) \equiv -\sum_k \sum_s p_{ks} log(p_{ks}) - \sum_i \sum_j w_{ij} log(w_{ij}) \right\} \tag{17}$$

subject to

$$\begin{aligned} Y_i &= \sum_k \sum_s z_{ks} p_{ks} x_{ik} + \sum_j v_j w_{ij} \\ \sum_s p_{ks} &= 1 \\ \sum_j w_{ij} &= 1 \end{aligned} \tag{18}$$

where the second set of constraints are the normalization of both $\mathbf{p}$ and $\mathbf{w}$. Forming the Lagrangian and solving yields the estimated probabilities for $\beta$

$$\hat{p}_{ks} = \frac{exp(-z_{ks} \sum_i \hat{\lambda}_i x_{ik})}{\sum_s exp(-z_{ks} \sum_i \hat{\lambda}_i x_{ik})} \equiv \frac{exp(-z_{ks} \sum_i \hat{\lambda}_i x_{ik})}{\Omega_k(\hat{\lambda}_i)},$$

and the estimated probabilities for $\epsilon$

$$\hat{w}_{ij} = \frac{exp(-\hat{\lambda}_i v_j)}{\sum_j exp(-\hat{\lambda}_i v_j)} \equiv \frac{exp(-\hat{\lambda}_i v_j)}{\Psi_i(\hat{\lambda}_i)}$$

where $\lambda$ are the Lagrange multipliers associated with (18). The estimated values of $\beta$ and $\epsilon$ are

$$\hat{\beta}_k \equiv \sum_s z_{ks}\hat{p}_{ks} \tag{19}$$

$$\hat{\epsilon}_i \equiv \sum_j v_j \hat{w}_{ij},$$

where $\Omega_k(\hat{\lambda}_i) = \sum_s exp(-z_{ks}\sum_i \hat{\lambda}_i x_{ik})$ and $\Psi_i(\hat{\lambda}_i) = \sum_j exp(-\hat{\lambda}_i v_j)$ are the partition functions (also known as the normalization functions).

The concentrated (unconstrained/dual) GME model is:

$$\begin{aligned}\max_{p\in P, w\in W} H(P,W) &= \min_{\lambda\in D}\left\{\sum_i Y_i\lambda_i + \sum_k log\Omega_k(\lambda_i) + \sum_i log\Psi_i(\lambda_i)\right\} \\ &= \min_{\lambda\in D}\left\{\sum_i Y_i\lambda_i + \sum_k log\left[\sum_s exp(-z_{ks}\sum_i \lambda_i x_{ik})\right] + \sum_i log\left[\sum_j exp(-\lambda_i v_j)\right]\right\}.\end{aligned} \tag{20}$$

Solving for $\lambda$'s provide the estimated $\mathbf{p}$'s and $\mathbf{w}$'s which in turn provide $\hat{\beta}$ and $\hat{\epsilon}$. For more details, examples, theoretical and empirical applications, as well as different support structures see Golan (2008).

### 3.3. **Inference and Diagnostics**

Under the GME approach, we maximize the joint entropies of the signal and the noise. Keeping in mind that the estimated errors are minimized when the entropy of the noise, $H(\mathbf{w})$, is maximized and the estimated probability distributions of each $\beta_k$ are pushed to uniformity within its pre-specified support $z_k$, the natural statistic used for identifying the best possible (or "correct") model is the value of the objective function (17), or similarly (20). Note that, similar to the maximum likelihood and empirical likelihood approaches, the objective function provides also the foundations for the traditional entropy-ratio statistics and other $\chi^2$ tests. See Golan (2008) for summary of these tests.

To summarize, we use our proposed iterative IT method to simultaneously choose the best model (a set of models) and estimate the parameters of these models.

## 4. SAMPLING EXPERIMENTS

To demonstrate the behavior of our iterative approach, we use different data generating approaches. The first is simple (call it, "Simple Data") while the others ("Random Data") are much more realistic and consistent with the other studies.

In our experiments and empirical applications, we also compare our estimator with other information-theoretic estimators including the Empirical Likelihood (EL), classical Maximum Entropy (ME), and Ordinary Least Squares (OLS). The linear regression model of Empirical Likelihood (EL) approach can be written as below,

$$
\begin{aligned}
& l(\beta, \theta; \mathbf{Y}) = \max_{p,\beta} \left\{ \sum_{i=1}^{N} log(p_i) \right\} \\
& s.t. \\
& \sum_{i=1}^{N} p_i \mathbf{x}_i (Y_i - \sum_k x_{ik}\beta_k) = \mathbf{0} \\
& \sum_{i=1}^{N} p_i = 1; p_i \geq 0
\end{aligned}
\tag{21}
$$

where the $N$-dimensional $p$ is different from the previous GME method. The classical Maximum Entropy (ME) optimizes the Shannon entropy function with respect to the same constraints as in the EL method above. The ME model is

$$
\begin{aligned}
& l(\beta, \theta; \mathbf{Y}) = \max_{p,\beta} \left\{ -\sum_{i=1}^{N} p_i log(p_i) \right\} \\
& s.t. \\
& \sum_{i=1}^{N} p_i \mathbf{x}_i (Y_i - \sum_k x_{ik}\beta_k) = \mathbf{0} \\
& \sum_{i=1}^{N} p_i = 1; p_i \geq 0
\end{aligned}
\tag{22}
$$

### 4.1. **The Simple Data Generation Process**

Let $\mathbf{X}$ be the interval-valued explanatory variable while $\mathbf{Y}$ be the interval-valued dependent variable.

(1) Suppose that $X_i^m$ is the $m$th interval of the explanatory variable, $\mathbf{X}$, and $Y_i^m$ be the $m$th interval of the dependent variable of the $i$th observation. Assume $X_i^m$ and $Y_i^m$ have a linear relationship:

$$Y_i^m = \beta_1 + \beta_2 X_i^m + \epsilon_i \tag{23}$$

where the true regression coefficients, $\beta = (2, -3)$ and the error distribution follows $\epsilon_i \sim N(0, \sigma^2)$.

(2) For the $i$th observation, randomly generate $X_i^1$ from uniform(0, 20); $X_i^2$ from uniform(21, 40); $X_i^3$ from uniform(41, 60); $X_i^4$ from uniform(61, 80); and $X_i^5$ from uniform(81, 100).

(3) $\epsilon_i \sim N(0, \sigma^2)$, where $\sigma = 2$

(4) Generate $Y_i^m$ according to (23) for $m = 1, 2, ..., M$ and $i = 1, 2, ..., N$

First, notice that each interval $m$ of $X_i^m$ causes its corresponding $m$ of $Y_i^m$. Second, the intervals of $\mathbf{X}$ is increasing, $X_i^1 < X_i^2 < ... < X_i^5$, such that the underlying interval-ordering of $\mathbf{Y}$ follows $Y_i^1 \leq Y_i^2 \leq ... \leq Y_i^5$.

Using $M = 5$ and $N = 50$, we ran a sampling experiment consisting of 1000 samples. For the GME estimations, the parameter support space for both $\beta_1$ and $\beta_2$ are $z = [-500, 250, 0, 250, 500]$ while the errors support is $v_m = [-3\sigma_{Y^m}, 0, 3\sigma_{Y^m}]$.

In this "Simple Data" experiment, although we observed all 5 intervals of $X_i^1, ..., X_i^5$ and $Y_i^1, ..., Y_i^5$, we do not know which interval $m$ of $X_i^m$ causes which interval $m$ of $Y_i^m$ for estimation purposes. First, we assume that there is a "One-to-One" correspondence between an interval $m$ of $\mathbf{X}$ and $\mathbf{Y}$, $Y_i^m = f(X_i^m)$. Results associated with this baseline results for the "Simple Data" experiments are omitted since it's essentially the classical (trivial) problem. There are 5 sets of estimated values for each one of the parameters, $\beta_1$ and $\beta_2$, per sample, so the distribution of all 1000 samples is based on 5000 estimated values. As expected, the estimated coefficients are concentrated around the true value of $\beta_1 = 2$ and $\beta_2 = -3$. In a second set of experiments, we relax the assumption of "One-to-One" correspondence and iterate over all the possible combinations of the intervals. In this combination case, there are 25 estimated values for each one of the parameters per sample and the complete distribution is based on 25,000 estimated values.

[*Insert Figures 2a - 2d here*]

Figures 2a - 2d provide the results for the iterations IT-GME model. The histogram of estimated coefficients for the intercept, $\hat{\beta}_1$, is omitted. From Figure 2a, we demonstrate that it is very easy to identify the correct set of models generating the estimated coefficients of $\beta_2$ by looking at the two distinct humps: one with a very small variance concentrated around the true $\beta_2$ and a much larger variance of estimated coefficients distributed around zero. The histogram of the objective function value, and $H(\hat{p}, \hat{w})$ versus $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, can clearly distinguish the two humps (Figures 2b, 2c and 2d). We use the objective value, specified in terms of the total entropy, as our statistics for identifying the best model.

We note that similar figures and conclusions are derived from analyzing $H(\hat{p})$ vs. $\hat{\beta}_2$, and $H(\hat{w})$ vs. $\hat{\beta}_2$ for all and the top 5% of $H(\hat{p}, \hat{w})$, $H(\hat{p})$, and $H(\hat{w})$. However, similar scatter plots for the lowest 5% of $H(\hat{p}, \hat{w})$, $H(\hat{p})$, and $H(\hat{w})$ cannot recover the correct parameters. We present here only the statistic we discuss earlier- the objective function $H(\hat{p}, \hat{w})$. Also omitted is a plot of our objective function vs. Squared-Errors of $\hat{\beta}_2$ ($SE(\hat{\beta}_2)$) that suggests that the models that maximize the objective function are associated with lower squared-errors.

These results demonstrate that our statistic - the value of the objective function (total entropy) - is able to identify the correct model and therefore identify the true sub-distribution of each parameter within the overall empirical distribution. We are able to simultaneously identify the best underlying model and infer its parameters.

However, our first set of results is based on a generation process that is simple and unrealistic. Nevertheless, it is trivial to identify the "correct" model and estimated parameters values with our proposed method. But it demonstrates the strength of our approach.

Having shown the basic idea via a simple case, we move on to the more realistic case. In the next set of experiments, we follow on the experimental design of Ahn et. al. (2012). We call the first experiment "Random 1."

### 4.2. **The Random Data Generation Process**

#### 4.2.1. ***Case 1***

Let $X_1$ be the interval-valued explanatory variable and $Y$ be the interval-valued dependent variable. We generate our data as follows.

(1) Suppose that $X_{ik}^c$ is the center of $k$th explanatory variable, and $Y_i^c$ is the center of the dependent variable of the $i$th observation. Assume $X_{i1}^c$ and $Y_i^c$ have a linear relationship:

$$Y_i^c = \beta_1 + \beta_2 X_{i1}^c + \epsilon_i^c \tag{24}$$

and the regression coefficients, $(\beta_1, \beta_2) = (2, -3)$, and an error distribution that follows $\epsilon_i^c \sim N(0, \sigma_c^2)$ with $\sigma_c = 1$.

(2) For the $i$th observation, randomly generate $X_{i1}^c$ from uniform$(0, 100)$. Generate $Y_i^c$ according to (24) for $i = 1, ..., N$.

(3) Suppose that $X_{ik}^r$ is the range (or half-range) of $k$th explanatory variable, and $Y_i^r$ is the range of the dependent variable of the $i$th observation. For the $i$th observation, generate $X_{i1}^r$ randomly from uniform$(1, 5)$; and generate the range $Y_i^r$ randomly from uniform$(1, 10)$ for $i = 1, ..., N$.

(4) Calculate the bounds of the $i$th observation:

$$[X_{i1}^{min}, X_{i1}^{max}] = [X_{i1}^c - X_{i1}^r, X_{i1}^c + X_{i1}^r] \qquad \text{and} \qquad [Y_i^{min}, Y_i^{max}] = [Y_i^c - Y_i^r, Y_i^c + Y_i^r]$$

First, notice that the only causality is through the center point for each observation. Second, since the ranges of $\mathbf{X}$ and $\mathbf{Y}$ are independently generated from two different uniform distributions, the actual causality of the intervals is much harder to infer. Different specifications of the ranges, where $Y_i^r$ is a linear function of $X_{i1}^r$, and heteroskedastic errors are also considered, but omitted here for brevity. Contrary to the previous case, we only observed the lower and upper bounds of $\mathbf{X}$ and $\mathbf{Y}$, which we divided into $M$ equally-spaced intervals before the GME method procedure is applied. In all of our

experiments, we choose $M = 5$. As expected, this problem is harder than the previous one. The empirical distributions are more complex and we do not observe the perfect two humps, as in the "Simple Data" case, where the "correct" and "incorrect" are distinctly separated. There is now an overlapping of "correct" and "incorrect" $\hat{\beta}_2$ estimates. The correct subset of models lies within the incorrect one. Nonetheless, the value of the objective function still provides a good statistic for identifying the true model and its parameters. We preset these results in Figures 3a - 3d.

[*Insert Figures 3a - 3d here*]

In Figure 3a (the $\hat{\beta}_2$'s distribution) it is much harder to distinguish between the "correct" and "incorrect" sub-distributions. It seems to suggest that there is an overlapping between "correct" and "incorrect" $\hat{\beta}_2$ estimates with a much larger variance of estimated coefficients distributed around $\beta_2 = -3$. Looking at the empirical distribution of the objective values, there is little concentration at the top while scatter plots of the objective value versus $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, does not distinguish the humps (Figures 3b, 3c and 3d). However, the variances of $\hat{\beta}_2$ decrease slightly at higher levels of objective values. Scatter plot of objective function vs. Squared-Errors of $\hat{\beta}_2$ also confirm the merits of the proposed statistic - lower $SE(\hat{\beta}_2)$ are associated with larger $H(\hat{p}, w)$.

4.2.2. ***Case 2***

In order to investigate the sensitivity of our proposed method across the parameter space, we repeated the previous set of experiments but with different values of the true parameters. In this case, the true parameters are farther away from zero while all support spaces ($\mathbf{z}$) remain as before. We repeat the data generation processes done in the previous case but with one simple modification: the original true-parameters are multiplied by 2, 5, 10, and 20, for each set of experiments, respectively. Therefore, the true parameters are (4, -6), (10, -15), (20, -30) and (40, -60), respectively for $(\beta_1, \beta_2)$, for each one of the four sets of experiments. The results are shown in Figures 4a - 4d.

[*Insert Figures 4a - 4d here*]

Figures 4a - 4d provide histograms of the estimated $\beta_2$ for different values of the parameters. It is clear that as the parameter value increases, the distributions of the "correct" and "incorrect" models become clearly distinctive. For example in Figure 4d, with $\beta = (40, -60)$, it is much easier to identify the "correct" from "incorrect" sub-distributions.

[*Insert Figures 5a - 5d here*]

Figures 5a - 5d provide scatter plots of the objective function and $\hat{\beta}_2$ that correspond to Figures 4a - 4d. Again the value of the objective function proved to be the statistic that can identify the best model leading to estimates with lower risk and variance.

4.2.3. ***Case 3 - Multiple X's***

We now investigate the performance of our approach with multiple **X**'s. In that case the basic data generation process is just a direct extension of the previous case with an additional **X**. The true $\beta_1$, $\beta_2$, and $\beta_3$ are 2, -3, and 4, respectively; $\mathbf{X}_1^r$ is randomly generated from $U(1,5)$ and $\mathbf{X}_2^r$ from $U(2,3)$; and $\mathbf{Y}^r$ from $U(1,10)$. The estimation uses only the first 50 ($N/2$) observations while predictions and performance evaluations use the second 50 observation. The estimated coefficients of $\beta_2$ and $\beta_3$ are nicely distributed around the true values of -3 and 4, respectively. The results are qualitatively similar to the previous case. We summarize it in Table 1. In this case, we compare our approach with its competitors. These competitors include the other methods used for interval estimation as well as other estimators for the linear model within the iterative approach we propose here.

[*Insert Table 1 here*]

Table 1 provides results from the proposed iterative IT-GME approach for this experiment, as well as other information-theoretic estimators such as Empirical Likelihood (EL), classical Maximum Entropy (ME), see (21) & (22); iterative Ordinary Least Squares (OLS); and the existing methods such as Center Method (CM), Center and Range Method (CRM), Constrained Center and Range Method (CCRM), and Min and Max Method (using R's `iRegression` package). The estimators for the CM and CRM are given (6) and (9) while the Min/Max is just running a separate regression for the minimum and maximum bounds of the variables. The CCRM imposes a positive constraint on the range coefficients of the CRM model. The estimated coefficients, Standard Errors, Mean Square Errors, out-of-sample Root Mean Squared Errors (ORMSE), and out-of-sample Mean Absolute Errors (OMAE) are average over the 1000 samples. The generic RMSEs and MAEs are specified as the differences between the predicted values $[\hat{Y}_i^L, \hat{Y}_i^U]$ and the observed values $[Y_i^L, Y_i^U]$ (see Lima Neto et. al. (2008)):

$$RMSE.m = \sqrt{\frac{\sum_{i=1}^{N}(Y_i^m - \hat{Y}_i^m)^2}{N}} \quad \& \quad MAE.m = \frac{\sum_{i=1}^{N}|Y_i^m - \hat{Y}_i^m|}{N}, \quad m = L, U \tag{25}$$

and the predicted values are calculated as in other methods discussed: (8) for CM, (11) for CRM, and Min/Max is given as

$$\hat{\mathbf{Y}}^m = \mathbf{x}_0^m \hat{\beta}^m, \quad m = L, U \tag{26}$$

The column name **maxObj** selects the model that maximizes the objective function or total entropies over iterative models. That is, $\hat{\beta}_{kt}^* = \{\hat{\beta}_{kt} | H_{tj}(\hat{\mathbf{p}}, \hat{\mathbf{w}}) = H_t^*(\hat{\mathbf{p}}, \hat{\mathbf{w}})\}$ where $t = 1, ..., T$ samples,

$j = 1, ..., J$ iterations, $H_t^*(\hat{\mathbf{p}}, \hat{\mathbf{w}}) = \max\{H_{t1}(\hat{\mathbf{p}}, \hat{\mathbf{w}}), H_{t2}(\hat{\mathbf{p}}, \hat{\mathbf{w}}), ..., H_{tJ}(\hat{\mathbf{p}}, \hat{\mathbf{w}})\}$ of sample $t$ and $H_{tj}(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ is the objective value of $t$th sample and $j$th iteration. Here we are choosing the best model, hence **maxObj**, for each sample. For the iterative OLS, **minObj** is used as a criteria since the best (or most probable) model minimizes the least-square errors (objective value). The first 50 observations are used to estimate the parameters while the second 50 observations are used for prediction. Results presented in Table 1 are out-of-sample averages over $T = 1000$ samples, where their corresponding in-sample definitions as provided below:

$$\mathbf{bk} = \frac{\sum_{t=1}^{T} \hat{\beta}_{kt}^e}{T}, for\ \hat{\beta}_{kt}^e = \hat{\beta}_{kt}^*, \hat{\beta}_{kt}^C, \hat{\beta}_{kt}^R, \hat{\beta}_{kt}^L, \hat{\beta}_{kt}^U$$

$$\mathbf{obj} = \frac{\sum_{t=1}^{T} H_t^*(\hat{p}, \hat{w})}{T}$$

$$\mathbf{RMSE.u} = \frac{\sum_{t=1}^{T} RMSE.u_t}{T}$$

$$\mathbf{MAE.u} = \frac{\sum_{t=1}^{T} MAE.u_t}{T}$$

$$\mathbf{numIn} = \frac{1}{T \times N} \sum_{t,n} 1(\hat{Y}^L < Y^C \& Y^C < \hat{Y}^U)$$

$$\mathbf{numHCov} = \frac{1}{T \times N} \sum_{t,n} 1(Y^U < \hat{Y}^U)$$

$$\mathbf{StdErr(bk)} = \frac{\sum_{t=1}^{T} SE(\hat{\beta}_{kt}^e)}{T}, for\ \hat{\beta}_{kt}^e = \hat{\beta}_{kt}^*, \hat{\beta}_{kt}^C, \hat{\beta}_{kt}^R, \hat{\beta}_{kt}^L, \hat{\beta}_{kt}^U$$

$$\mathbf{MSE(bk)} = \frac{\sum_{t=1}^{T} (\hat{\beta}_{kt}^e - \beta_k)^2}{T}, for\ \hat{\beta}_{kt}^e = \hat{\beta}_{kt}^*, \hat{\beta}_{kt}^C, \hat{\beta}_{kt}^R, \hat{\beta}_{kt}^L, \hat{\beta}_{kt}^U$$

$$\mathbf{RMSE.l} = \frac{\sum_{t=1}^{T} RMSE.l_t}{T}$$

$$\mathbf{MAE.l} = \frac{\sum_{t=1}^{T} MAE.l_t}{T}$$

$$\mathbf{numCov} = \frac{1}{T \times N} \sum_{t,n} 1(\hat{Y}^L < Y^L \& Y^U < \hat{Y}^U)$$

$$\mathbf{numLCov} = \frac{1}{T \times N} \sum_{t,n} 1(\hat{Y}^L < Y^L)$$

$$\mathbf{numOverLap} = \frac{1}{T \times N} \sum_{t,n} \frac{|min(\hat{Y}^U, Y^U) - max(\hat{Y}^L, Y^L)|}{|max(\hat{Y}^U, Y^U) - min(\hat{Y}^L, Y^L)|} \tag{27}$$

On average, the model with maximum objective value, **maxObj** or **max(**$H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$**)**, estimated via GME, EL, and ME and least-square errors for iterative OLS produce consistent estimates of $\beta_2$ and $\beta_3$. Also, the estimated coefficients of $\beta_2$ and $\beta_3$ for the center, minimum, and maximum points are very close to the true value of -3 and 4. Overall, both CRM and CCRM produce the lowest ORMSE, OMAE, and perform the best in terms of all the new coverage statistics such as OnumIn, OnumCov, OnumHCov, OnumLCov, and OnumOverLap for both upper and lower intervals. Among the iterative information-theoretic estimators and iterative OLS, the iterative IT-GME performs the best in terms of all the statistics provided. It is also comparable with CM and slightly worse off than the MinMax on the new coverage statistics while its ORMSE and OMAE are smaller.

Given that the true data generation process follows the center points and their respective ranges, the existing CM, CRM, CCRM, and MinMax are the correct model by design while all the iterative

procedures are agnostics about the true underlying model. Even with such minimal assumption, the iterative estimators performed well in identifying the true or most probable model with the proposed criteria - maximum of the objective value and minimum for iterative OLS. It is worth noting that the CRM, CCRM, and MinMax require estimating two sets of parameters - coefficients for center and range or lower and upper. Since the proposed procedures require iteratively fitting all the possible models, it is natural to obtain and make use of the empirical distribution of the intercepts. One way is to modify the predicted intervals of lower and upper bounds. Among others, using the minimum and maximum of the empirical distribution of the intercepts, we can shift the predicted lines down for the lower and up for the upper bounds, respectively, while keeping the slope parameters chosen by **maxObj**. Let's call this criteria **maxObjI**. Then, the predicted lower and upper bounds are:

$$\hat{Y}^m = x_0^m \hat{\beta}_j^{m*}, m = L, U \tag{28}$$

where $\hat{\beta}_j^{L*} = [\min(\hat{\beta}_{1j}), \hat{\beta}_{2j}^*, \hat{\beta}_{3j}^*, ..., \hat{\beta}_{Kj}^*]$ and $\hat{\beta}_j^{U*} = [\max(\hat{\beta}_{1j}), \hat{\beta}_{2j}^*, \hat{\beta}_{3j}^*, ..., \hat{\beta}_{Kj}^*]$.

The results from the above intercept modifications greatly improve the coverage statistics. Among the iterative IT estimators and iterative OLS, the iterative IT-GME still preformes the best. Comparing with the existing methods, the iterative IT-GME out-performs them in terms of the coverage statistics, except for out-of-sample overlapping areas between observed and predicted intervals, while it under-performs in terms of the ORMSEs and OMAEs. Since we shift the lower and upper bounds to the extreme via using the minimum and maximum of the intercepts, $\hat{\beta}_1$, it is the most conservative lower and upper predictions one can make (over-coverage), hence the low overlapping areas and ORMSEs.

#### 4.2.4. ***Case 4 - Cauchy Distribution***

Since the errors and ranges from previous data generations were drawn from normal distributions, the center point is a good measure of centrality. However, that is not the norm in real applications, rather an exception. We now repeat Case 3, but now, the errors and ranges are randomly drawn from a Cauchy distribution: $\mathbf{e}^c \sim Cauchy(0, 2)$; $\mathbf{X}_1^r$, $\mathbf{X}_2^r \sim Cauchy(2.5, 2)$; and $\mathbf{Y}^r \sim Cauchy(5, 2)$. Our results are presented in Table 2.

[*Insert Table 2 here*]

On average, the model with maximum objective value, **maxObj**, for iterative OLS and IT estimators still produces the best estimates of $\beta_2$ and $\beta_3$. The existing methods also produce good estimates, except for the MinMax method with slope-parameters averages of [-2.26, 3.08] and [-2.29,3.04] for lower and upper bounds, respectively. Unlike the results from "Random Data 3", both CRM and CCRM no longer perform the best in all of the new coverage statistics. The iterative IT-GME out-performs CRM and CCRM in three out-of-sample coverage measures: percentages of complete coverage (OnumCov),

upper-bound coverage (OnumHCov), and lower-bound coverage (OnumLCov). The iterative IT-GME also out-performs MinMax and other iterative IT estimators on the majority of performance measures, except for ORMSEs and overlapping areas, while its performance is comparable with both iterative OLS and CM.

In the intercept modifications, the iterative IT-GME still out-performs other iterative IT estimators on the majority of measures, except for ORMSEs, while it is comparable with the iterative OLS. Comparing with the existing methods, the iterative IT-GME still out-performs them on coverage statistics, except for overlapping areas, while it under-performs on ORMSEs and OMAEs. We note that results from small sample versions of previous random data with small and large noises as well as correlations are largely the same with results presented here in this section.

### 4.3. **Mixed Models**

All the experiments presented above assume that both the upper and lower bounds are coming from the same underlying model (process). To relax such assumption, we generate a "Mixed Models Data" in which there exist two "correct" sub-samples within the data: one for the lower bounds and another for the upper bounds.

$$\begin{aligned} Y_i^{min} &= (-\beta_1) + (-\beta_2)X_{i1}^{min} + (-\beta_3)X_{i2}^{min} + \epsilon_i \\ Y_i^{max} &= \beta_1 + \beta_2 X_{i1}^{max} + \beta_3 X_{i2}^{max} + \epsilon_i \end{aligned} \tag{29}$$

where the true $\beta_1$, $\beta_2$, and $\beta_3$ are 2, -3, and 4, respectively; for the $i$th observation, $\epsilon_i \sim N(0,1)$; randomly generate $X_{ik}^c$ from $U(0,100)$, $X_{i1}^r$ from $U(1,5)$, $X_{i2}^r$ from $U(2,3)$, and $[X_{ik}^{min}, X_{ik}^{max}] = [X_{ik}^c - X_{ik}^r, X_{ik}^c + X_{ik}^r]$ for $k = 1, 2$. That is, the true parameters are -$\beta$ for $Y_i^{min}$ and $\beta$ for $Y_i^{max}$; and MinMax method is the correct model for this data generation process. We present the results in Figures 6a - 6d.

[*Insert Figures 6a - 6d here*]

Figures 6a and 6c provide scatter plots of $H(\hat{p}, \hat{w})$ and $\hat{\beta}_2$ and $\hat{\beta}_3$, respectively, for all 125000 estimated models. There are five distinct concentrations of $\hat{\beta}_2$ and $\hat{\beta}_3$ that are symmetric and center around 0. When looking at the same scatter plots for estimated parameters with maximum objective values ($H^*(\hat{p}, \hat{w})$) in 6b and 6d, the estimated values of $\beta_2$ and $\beta_3$ are concentrated only on -3 or 3 and -4 or 4. Such distinct distributions allow us to separate the two underlying sub-samples: estimated coefficients associated with $\beta$ and $-\beta$ within each sample. That is, for each sample, we pick two sets $(-\hat{\beta}, \hat{\beta})$ of estimated coefficients with maximum objective values ($H^*(\hat{\mathbf{p}}, \hat{\mathbf{w}})$). We present the different

statistics of this experiment in Table 3.

[*Insert Table 3 here*]

Table 3 provides comparison of results for the "Mixed Models Data". For all the iterative IT estimators and iterative OLS, we can distinguish the two sub-distributions: **bkn** for $-\beta$ or lower bounds and **bk** for $\beta$ or upper bounds. On average, iterative IT-GME performs the best among iterative IT estimators. It is comparable with iterative OLS and the "correct" model of MinMax methods on all measures. CM, CRM, and CCRM peform worse in recovering the correct parameters or identifying the two sub-samples, while they do fine with complete coverage, upper-bound coverage, and lower-coverage measures.

### 4.4. **Multi-Mixed Models**

To generalize the above "Mixed Models Data", we generate a "Multi-Mixed Data" in which there exist five "correct" sub-samples within the data: one for the lower bounds, another for the upper bounds, and three distinct models between the extreme bounds. That is,

$$Y_i^m = (\beta_1^m) + (\beta_2^m) X_{i1}^m + \epsilon_i \tag{30}$$

where the true ($\beta_1^m$, $\beta_2^m$) pairs are $(-2,-6), (-1,-3), (0,2), (1,3)$, and $(2,6)$ for each $m = 1, ..., 5$, respectively; for the $i$th observation, $\epsilon_i \sim N(0, 25)$; randomly generate each $X_{i1}^m$ from $U(0,20)$, $U(21,40)$, $U(41,60)$, $U(61,80)$, and $U(81,100)$. MinMax method no longer captures the whole correct models for this data generation process. We present the results in Figures 7a and 7b.

[*Insert Figures 7a and 7b here*]

Figure 7a provides scatter plots of $H(\hat{p}, \hat{w})$ and $\hat{\beta}_2$, for all 25000 estimated models. There are five distinct concentrations of $\hat{\beta}_2$ around the true $\beta_2$ values, as well as four other "incorrect" concentrations. When looking at the same scatter plots for estimated parameters with maximum objective values ($H^*(\hat{p}, \hat{w})$) in 7b, the estimated values of $\beta_2$ are concentrated only on -6, -3, 2, 3, and 6. Such distinct distributions allow us to separate the two underlying sub-samples: estimated coefficients associated with each $\beta^m$ within each sample. That is, for each sample, we pick five sets ($\hat{\beta}^m, m = 1, ..., 5$) of estimated coefficients with maximum objective values ($H^*(\hat{\mathbf{p}}, \hat{\mathbf{w}})$).

For all the iterative IT estimators and iterative OLS, we can distinguish the five sub-distributions corresponding all five correct models. On average, iterative IT-GME performs the best among iterative IT estimators. It is comparable with iterative OLS on all measures while the MinMax method can

captures only two of the five correct models. CM, CRM, and CCRM peform worse in recovering the correct parameters or identifying the five sub-samples.

## 5. EMPIRICAL EXAMPLES

In this section, we analyze two empirical examples using the iterative IT Generalized Maximum Entropy (GME) estimator and other competing approaches. The data are obtained from Billard and Diday (2012).

### 5.1. **Cholesterol Data**

Gillard and Diday (2012) provides interval-valued observations for an independent variable $X = Age$ and two dependent random variables $Y_1$=Cholesterol and $Y_2$=Weight for a certain population. There are seven of such interval-valued observations.

The parameter and error supports specifications for the proposed iterative IT GME estimator are as follows. For $\beta_1$ and $\beta_2$, the supports are $\mathbf{z} = [-500, -250, 0, 250, 500]$. The error supports are set to be symmetric around zero, $v_m = [-3 \times \sigma_{Y^{mc}}, 0, 3 \times \sigma_{Y^{mc}}]$ for each sub-$m$th interval, where $Y^{mc}$ is the center points of sub-interval $Y^m$. So, they are the empirical standard deviation of the center points of the dependent variable's sub-intervals, not the end-points. The intuition is that since the minimum bounds of each sub-interval would potentially be different from the centers of those sub-intervals, the centers of sub-intervals would capture more accurately the true variances of the intervals. Our results are robust to changes in the end points of the $\mathbf{z}$ supports.

[*Insert Table 4 here*]

Table 4 provides results from iterative IT estimators, iterative OLS, CM, CRM, CCRM, and Min/Max Method, where Cholesterol as a dependent variable and Age as a single predictor. Since there is no variations in the Age interval across observations, we cannot use the ranges of Age. Therefore,the lower bound of Age is used instead for CRM and CCRM. That is, the ranges of Cholesterol is regressed against Age-Low. The estimated coefficients across different methods are statistically significant and consistent in terms of the signs and magnitudes. Overall, CRM, CCRM, and MinMax perform the best in terms of RMSEs, MAEs, and coverage statistics. The iterative IT-GME performs similarly with (or marginally better than) iterative OLS and CM on all measures of fitness. For other iterative IT estimators, RMSEs and MAEs for the lower-bound prediction are similar to the existing methods while the upper-bound predictions are generally worse.

### 5.2. **Blood Pressure Data**

We use data provided by Billard and Diday (2012) that have interval-valued observations for a dependent variable $Y$ = Pulse Rate and two independent random variables $X_1$=Systolic Pressure and $X_2$=Diastolic Pressure for a certain population. Billard and Diday (2012) provide more information about the data. Xu (2010) discussed the need to apply interval-valued data rules such as Diastolic Pressure must be less than Systolic Pressure. Therefore, the data used here is the same as in Xu (2010) where there are 11 of such interval-valued observations.

The parameter and error support specifications for the proposed iterative IT GME estimator for this data are: $\mathbf{z} = [-500, -250, 0, 250, 500]$. The results are robust to choices in the end points of $\mathbf{z}$ supports. The errors supports are derived as before and symmetric around zero.

[*Insert Table 5 here*]

Table 5 provide results using iterative IT estimators, iterative OLS, and existing methods such as CM, CRM, CCRM and Min/Max. The estimated coefficients are similar in terms of the signs and magnitude. However, only Systolic Pressure seems to be statistically significant across the different methods.

## 6. CONCLUSION

In this paper, we proposed an iterative Information-Theoretic (IT) method for fitting a linear regression model for interval-valued data. Since there could be infinitely many models that are consistent with the observed interval-valued data, the proposed method identifies the most probable (causal) model that is consistent with the observed information. The estimation method we use is an information-theoretic GME estimator. The iteration process searches through a finite set of possible models within the observed data. To identify the best model we use an entropy criterion. Our approach allows us to relax distributional assumptions of within the interval behavior, as well as those of the underlying statistical process, and fully utilize all observed information that captures internal variations in the interval-valued observations. We provided numerous sampling experiments, robustness and sensitivity analyses as well as empirical analyses of different data sets. We also contrasted our results with those of competing models.

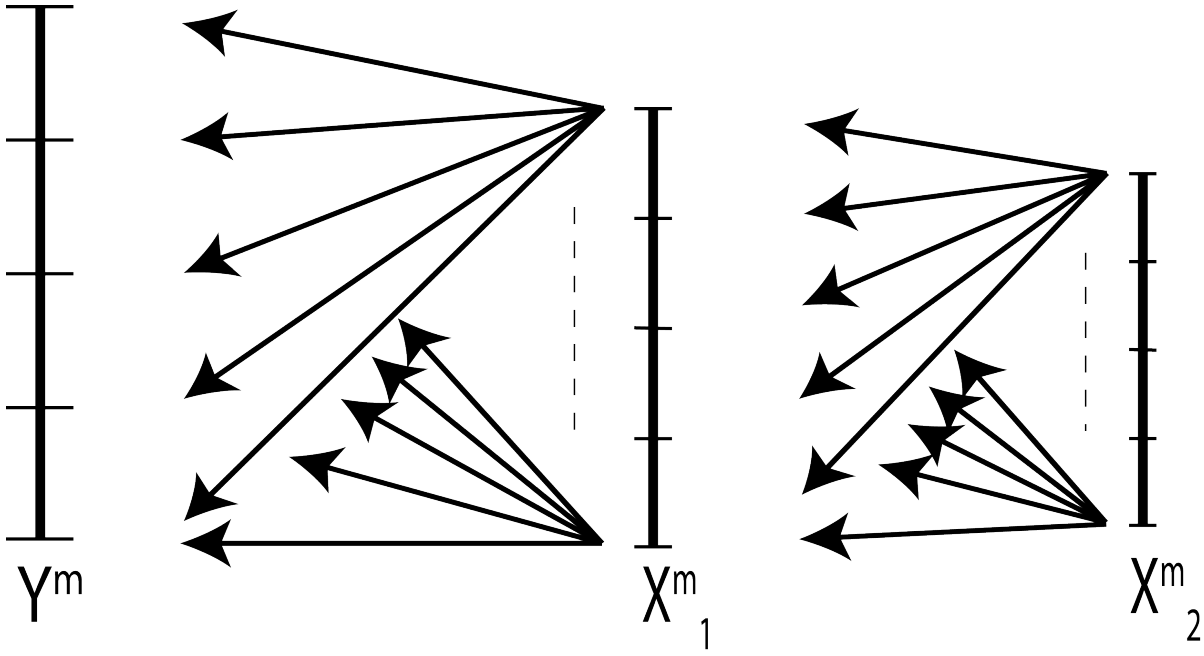
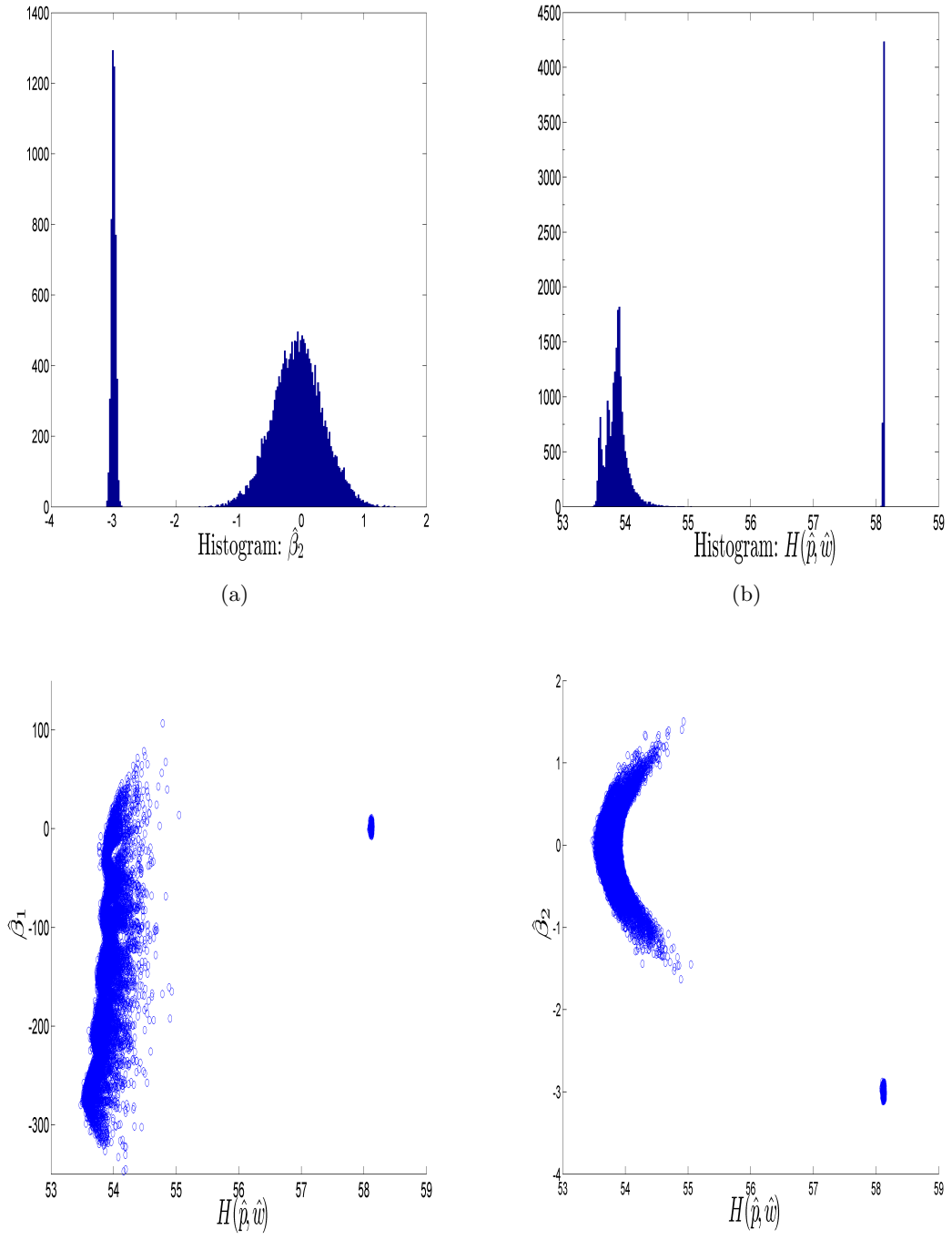


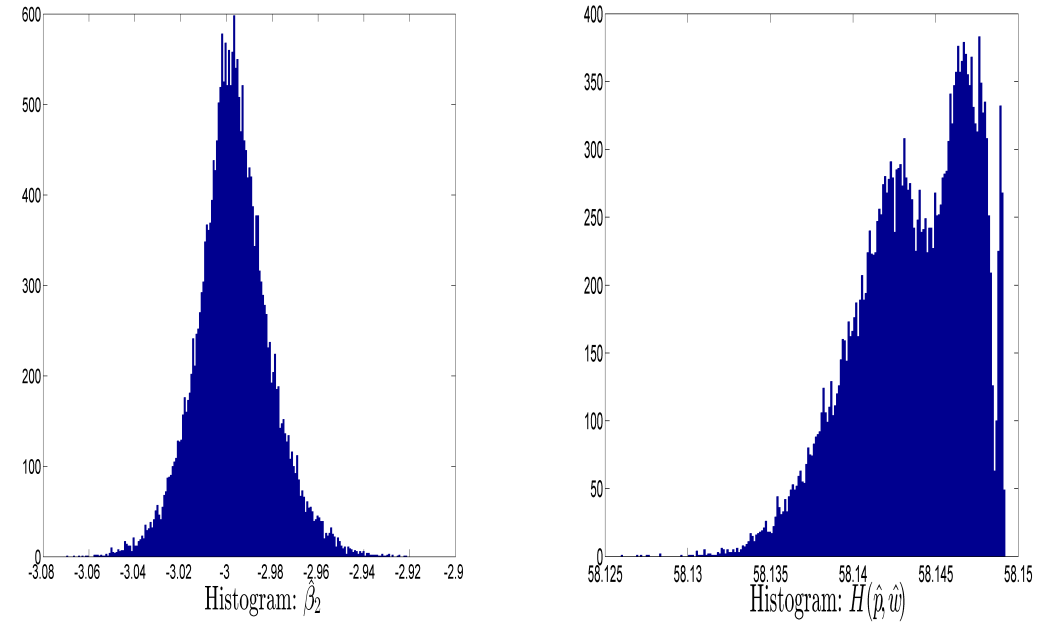

FIGURE 1. The Proposed iterative IT-GME approach is to iterate through all possible combination of sub-intervals within the interval-valued random variables $\mathbf{X}$ and $\mathbf{Y}$. For $M = 5$ and $K = 1$, we iterate through 25 models since we have 25 sub-interval pairs of $\mathbf{X}$ and $\mathbf{Y}$, including $(X_1^1, Y^1)$, $(X_1^1, Y^2)$,..., $(X_1^1, Y^5)$, $(X_1^2, Y^1)$, $(X_1^2, Y^2)$,...,$(X_1^5, Y^5)$. For $M = 5$ and $K = 2$, it requires 125 iterations since the number of all possible combination is $M \times M^K$.



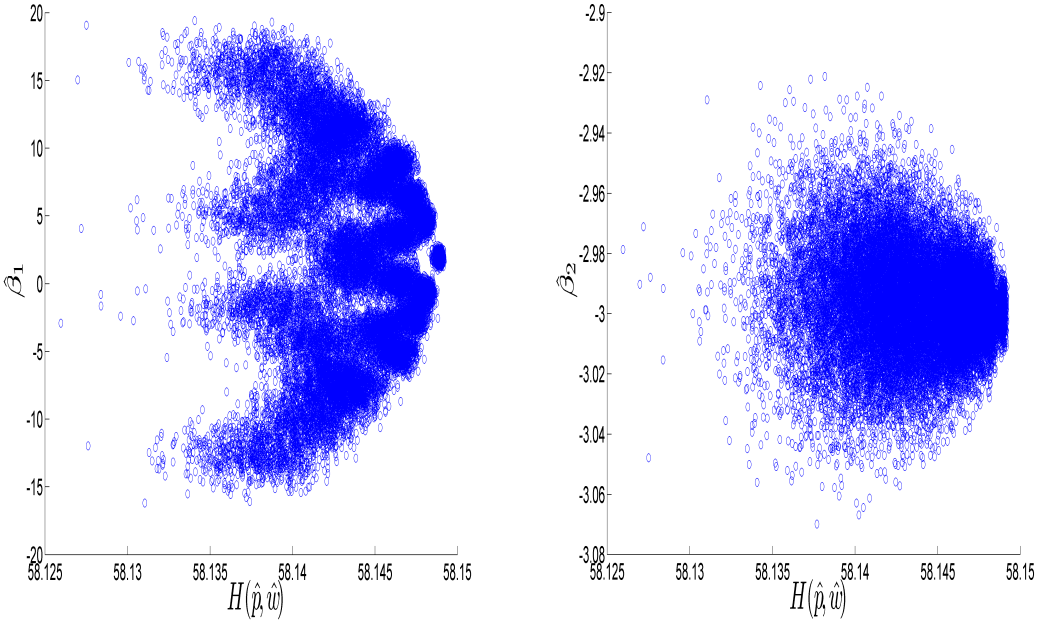

(a) (b)

(c) (d)

FIGURE 2. Results from 1000 samples of "Simple Data" with $N = 50$, $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$. Subfigures (a) and (b) show histograms of $\hat{\beta}_2$ and $H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$; (c) and (d) show scatter plots of $H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ vs. $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.




(a) (b)




(c) (d)

FIGURE 3. Results from 1000 samples of "Random Data 1" with $N = 50$, $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$. Subfigures (a) and (b) show histograms of $\hat{\beta}_2$ and $H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$; (c) and (d) show scatter plots of $H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ vs. $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

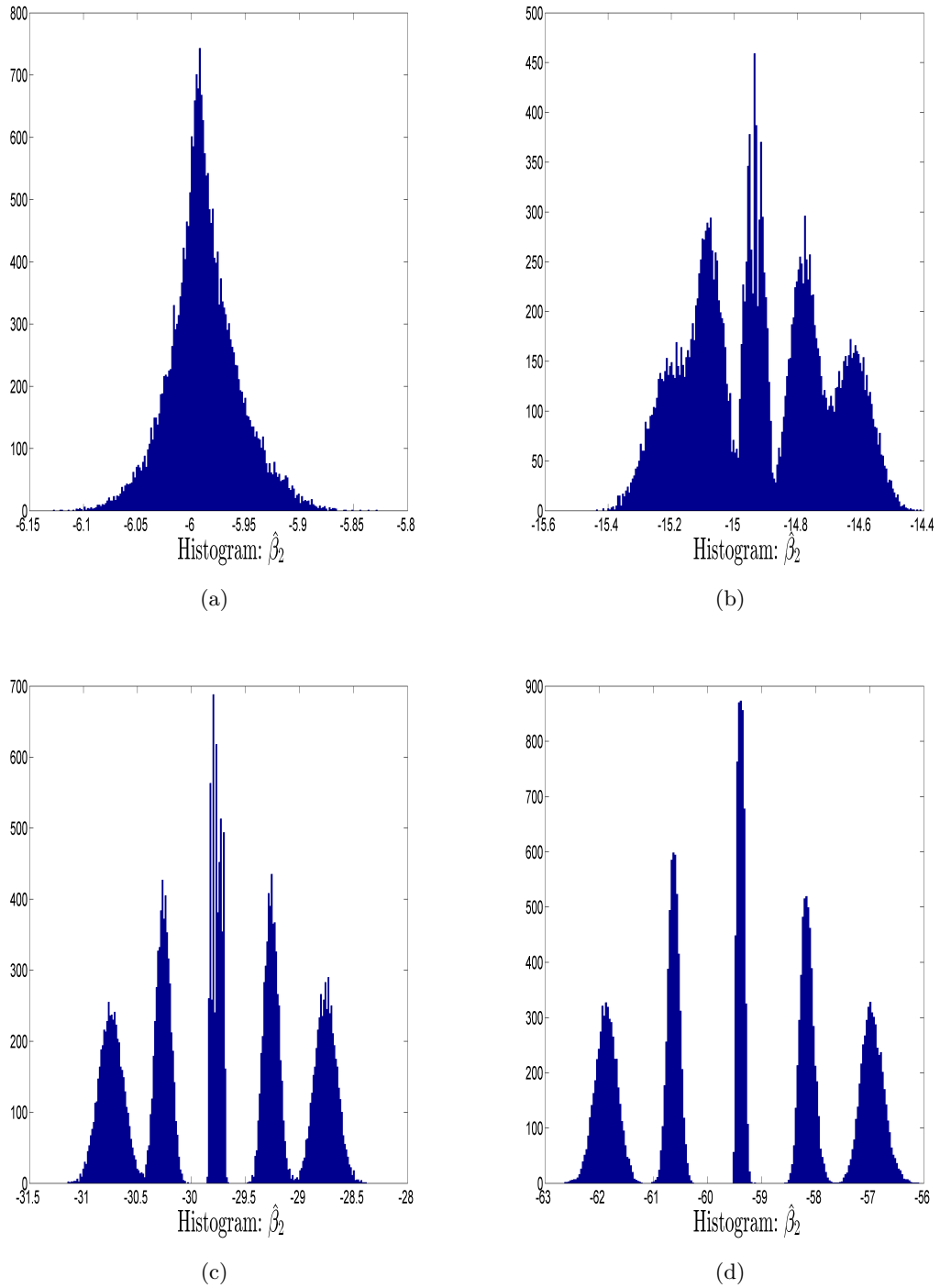
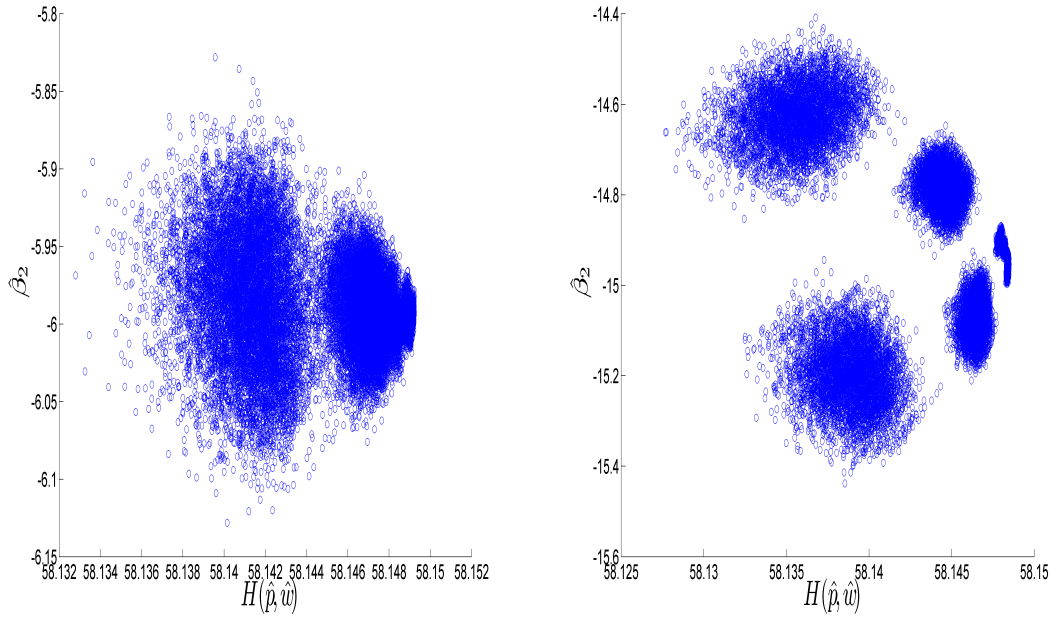
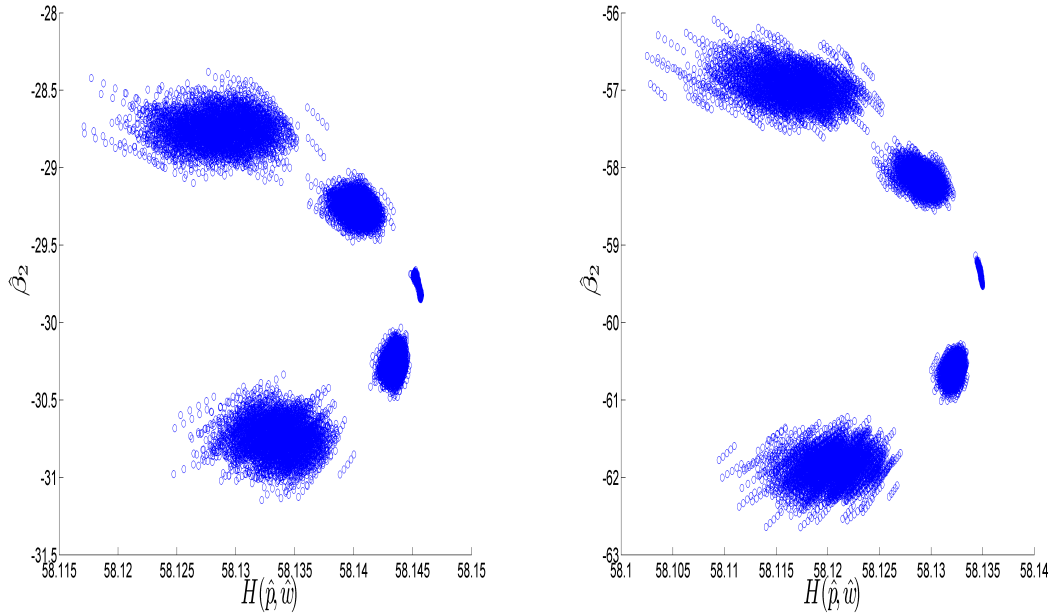



FIGURE 4. Results from 1000 samples of "Random Data 2" with different true-parameters: (a) $\beta = 2 \times [2, -3]'$, (b) $\beta = 5 \times [2, -3]'$, (c) $\beta = 10 \times [2, -3]'$, and (d) $\beta = 20 \times [2, -3]'$; and $N = 50$, $\mathbf{z} = (-500, 250, 0, 250, 500)$, and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$. Subfigures (a), (b),(c) and (d) show histograms of $\hat{\beta}_2$ for samples with $2\times\beta$, $5\times\beta$, $10\times\beta$, and $20\times\beta$, respectively.




(a) (b)




(c) (d)

FIGURE 5. Results from 1000 samples of "Random Data 2" with different true-parameters: (a) $\beta = 2 \times [2, -3]'$, (b) $\beta = 5 \times [2, -3]'$, (c) $\beta = 10 \times [2, -3]'$, and (d) $\beta = 20 \times [2, -3]'$; and $N = 50$, $\mathbf{z} = (-500, 250, 0, 250, 500)$, and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$. Subfigures (a), (b),(c) and (d) show scatter plots of $H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ vs. $\hat{\beta}_2$ for samples with $2\times\beta$, $5\times\beta$, $10\times\beta$, and $20\times\beta$, respectively.

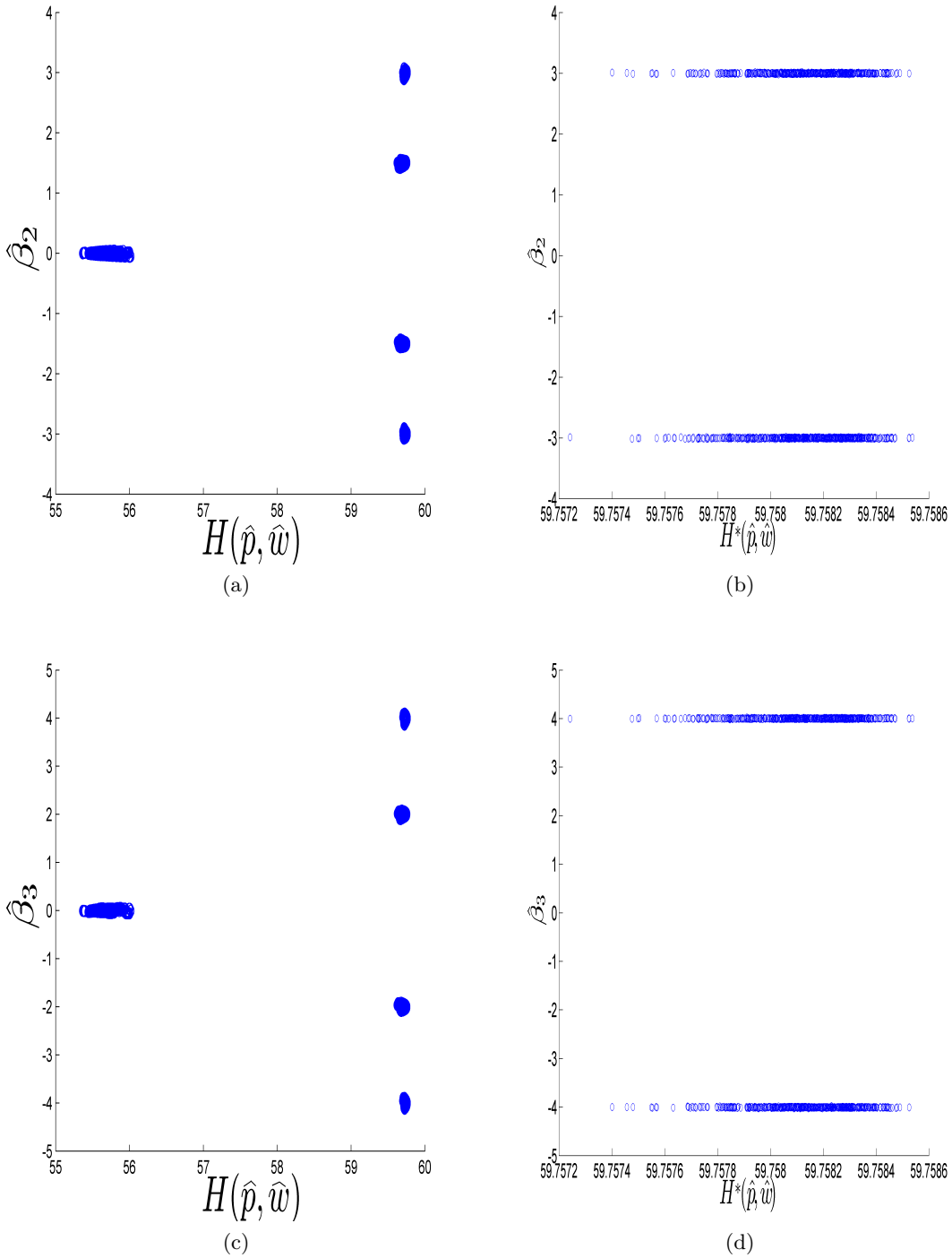
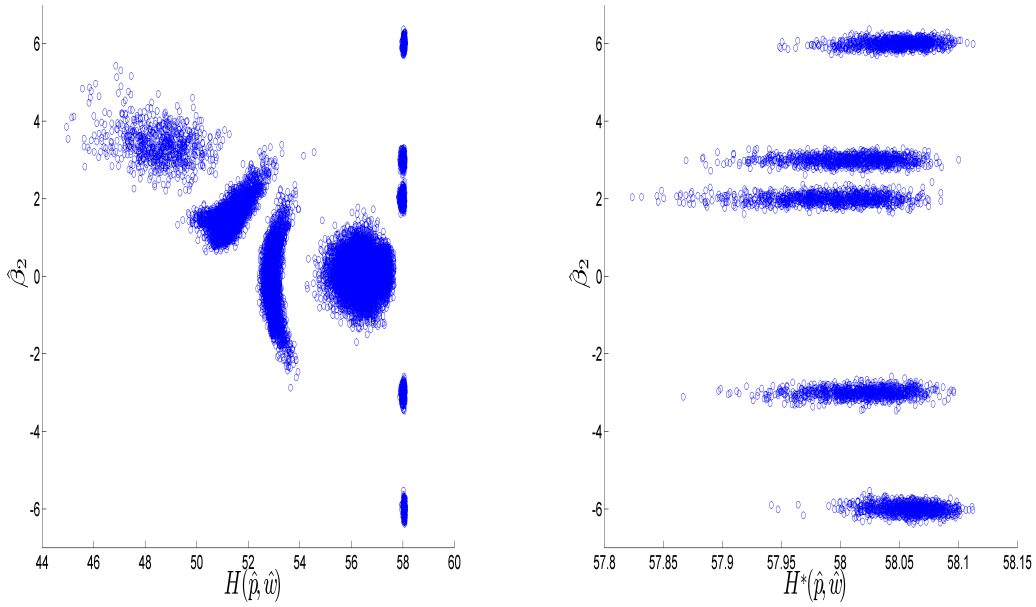



FIGURE 6. Results from 1000 samples of "Mixed Models Data" for the iterative GME model with $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$. Subfigures (a) and (c) show scatter plots of $H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ vs. $\hat{\beta}_2$ and $\hat{\beta}_3$, respectively; and (b) and (d) show scatter plots of $H^*(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ vs. $\hat{\beta}_2$ and $\hat{\beta}_3$, respectively.




(a) (b)

FIGURE 7. Results from 1000 samples of "Multi-Models Data" for the iterative GME model with $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$. Subfigures (a) shows scatter plot of $H(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ vs. $\hat{\beta}_2$ and (b) shows scatter plot of $H^*(\hat{\mathbf{p}}, \hat{\mathbf{w}})$ vs. $\hat{\beta}_2$.

TABLE 1. Random Data 3 - Multiple X's (T=1000): Results from iterative IT, OLS, and existing methods

| | **iGME** | **iEL** | **iME** | **iOLS** | **CM** | **CRM** | **CCRM** | **MinMax** |
|---|---|---|---|---|---|---|---|---|
| | **maxObj** | **maxObj** | **maxObj** | **minObj** | **Est.C** | **Est.C** | **Est.C** | **Est.L** |
| **b1** | 1.6966 | -0.5795 | -2.8436 | 2.0128 | 1.9906 | 1.9906 | 1.9906 | -2.7078 |
| **b2** | -2.9974 | -2.9970 | -2.9950 | -2.9997 | -2.9997 | -2.9997 | -2.9997 | -2.9950 |
| **b3** | 4.0018 | 4.0000 | 3.9990 | 3.9996 | 3.9996 | 3.9996 | 3.9996 | 3.9989 |
| **StdErr(b1)** | 0.3720 | 0.0420 | 0.4532 | 0.3854 | 0.3817 | 0.3817 | 0.3817 | 1.6596 |
| **StdErr(b2)** | 0.0049 | 0.0056 | 0.0608 | 0.0050 | 0.0050 | 0.0050 | 0.0050 | 0.0227 |
| **StdErr(b3)** | 0.0049 | 0.0055 | 0.0605 | 0.0051 | 0.0050 | 0.0050 | 0.0050 | 0.0229 |
| | | | | | **Est.R** | **Est.R** | **Est.R** | **Est.U** |
| **b1** | - | - | - | - | - | 11.1920 | 7.9046 | 6.2708 |
| **b2** | - | - | - | - | - | -0.0148 | 0.1194 | -2.9953 |
| **b3** | - | - | - | - | - | -0.0154 | 0.4814 | 3.9996 |
| **StdErr(b1)** | - | - | - | - | - | 6.8597 | 6.9394 | 1.8260 |
| **StdErr(b2)** | - | - | - | - | - | 0.3262 | 0.3300 | 0.0227 |
| **StdErr(b3)** | - | - | - | - | - | 1.3052 | 1.3203 | 0.0229 |
| **obj** | 59.7586 | 0.0296 | 3.9120 | 1.1101 | - | - | - | - |
| **ORMSE.l** | 4.1105 | 3.8250 | 4.3305 | 4.2246 | 4.1254 | 2.8680 | **2.8438** | 4.5154 |
| **ORMSE.u** | 4.1831 | 5.9863 | 7.8051 | 4.2088 | 4.1314 | 2.8545 | **2.8288** | 4.7076 |
| **OMAE.l** | 3.3588 | 3.1416 | 3.5486 | 3.4705 | 3.3714 | 2.4216 | **2.4068** | 3.7051 |
| **OMAE.u** | 3.4317 | 5.1596 | 7.0142 | 3.4592 | 3.3831 | 2.4117 | **2.3953** | 3.8501 |
| **MSE(b1/c/l)** | 0.3251 | 9.7618 | 26.3107 | 1.2212 | **0.1493** | **0.1493** | **0.1493** | 25.0772 |
| **MSE(b2/c/l)** | **0.0000** | 0.0003 | 0.0006 | **0.0000** | **0.0000** | **0.0000** | **0.0000** | 0.0006 |
| **MSE(b3/c/l)** | **0.0000** | 0.0003 | 0.0006 | **0.0000** | **0.0000** | **0.0000** | **0.0000** | 0.0006 |
| **MSE(b1u)** | - | - | - | - | - | - | - | 21.4991 |
| **MSE(b2u)** | - | - | - | - | - | - | - | 0.0005 |
| **MSE(b3u)** | - | - | - | - | - | - | - | 0.0005 |
| **N** | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| **OnumIn** | 0.8583 | 0.5757 | 0.2922 | 0.8357 | 0.8628 | **1.0000** | **1.0000** | 0.8707 |
| **OnumCov** | 0.1853 | 0.0931 | 0.0294 | 0.1798 | 0.1868 | 0.4141 | 0.4130 | **0.4348** |
| **OnumHCov** | 0.2563 | 0.1099 | 0.0296 | 0.2694 | 0.2618 | **0.5036** | 0.5022 | 0.5018 |
| **OnumLCov** | 0.2710 | **0.5138** | 0.7216 | 0.2676 | 0.2633 | 0.5064 | 0.5053 | 0.5073 |
| **OnumOverLap** | 0.4558 | 0.3662 | 0.2702 | 0.4489 | 0.4577 | 0.6381 | **0.6397** | 0.4857 |

Results from 1000 samples of "Random Data 3 - Multiple X's" for iterative information-theoretic (IT) estimators, iterative Ordinary Least Squares (iOLS), and existing methods. The first 50 observations ($N/2$) are used for estimation and the second-half are utilized for prediction. The estimated coefficients; Standard Errors; Mean Square Errors; and out-of-sample Root Mean Squared Errors (ORMSE), Mean Absolute Errors (OMAE), and coverage statistics are averaged over the 1000 samples. The iterative IT-GME uses $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$.

TABLE 2. Random Data 4 - Cauchy Distribution (T=1000): Results from iterative IT, OLS, and existing methods

| | iGME | iEL | iME | iOLS | CM | CRM | CCRM | MinMax |
|---|---|---|---|---|---|---|---|---|
| | maxObj | maxObj | maxObj | minObj | Est.C | Est.C | Est.C | Est.L |
| b1 | 1.1293 | 8.1065 | 4.1413 | 1.7283 | 1.7267 | 1.7267 | 1.7267 | 6.4063 |
| b2 | -2.9862 | -2.4345 | -2.2569 | -2.9994 | -2.9996 | -2.9996 | -2.9996 | -2.2570 |
| b3 | 4.0080 | 3.2782 | 3.1171 | 4.0056 | 4.0063 | 4.0063 | 4.0063 | 3.0767 |
| StdErr(b1) | 8.5724 | 0.7777 | 6.3383 | 8.6889 | 8.7448 | 8.7448 | 8.7448 | 23.7561 |
| StdErr(b2) | 0.1122 | 0.1141 | 0.9808 | 0.1138 | 0.1147 | 0.1147 | 0.1147 | 0.3293 |
| StdErr(b3) | 0.1120 | 0.1275 | 1.0529 | 0.1136 | 0.1144 | 0.1144 | 0.1144 | 0.3150 |
| | | | | | Est.R | Est.R | Est.R | Est.U |
| b1 | - | - | - | - | - | 19.4388 | 17.1238 | 19.4813 |
| b2 | - | - | - | - | - | 0.0042 | 0.1002 | -2.2915 |
| b3 | - | - | - | - | - | -0.0643 | 0.0915 | 3.0433 |
| StdErr(b1) | - | - | - | - | - | 8.4814 | 8.5241 | 29.2028 |
| StdErr(b2) | - | - | - | - | - | 0.3594 | 0.3613 | 0.3272 |
| StdErr(b3) | - | - | - | - | - | 0.3675 | 0.3701 | 0.3145 |
| obj | 59.5655 | 0.2140 | 3.9120 | 1.73E+07 | - | - | - | - |
| ORMSE.l | 431.8441 | 356.9115 | 343.0341 | 430.1148 | 430.1033 | 140.8712 | **135.3134** | 354.5416 |
| ORMSE.u | 432.1196 | 356.9037 | 341.8422 | 430.2478 | 430.2468 | 140.8954 | **135.2343** | 357.3542 |
| OMAE.l | 90.4475 | 94.8480 | 96.4670 | 88.7617 | 88.2033 | 32.0808 | **30.6785** | 97.7056 |
| OMAE.u | 90.4922 | 94.6514 | 95.2497 | 88.7384 | 88.1963 | 32.2018 | **30.6638** | 99.3937 |
| MSE(b1/c/l) | 773.203 | 7.2E+03 | 6.9E+03 | 182.803 | **138.742** | **138.742** | **138.742** | 6.9E+03 |
| MSE(b2/c/l) | 0.2509 | 1.1947 | 1.5853 | 0.0299 | **0.0289** | **0.0289** | **0.0289** | 1.5898 |
| MSE(b3/c/l) | 0.1152 | 1.9114 | 2.2714 | 0.0277 | **0.0250** | **0.0250** | **0.0250** | 2.4134 |
| MSE(b1u) | - | - | - | - | - | - | - | 9.9E+03 |
| MSE(b2u) | - | - | - | - | - | - | - | 1.6029 |
| MSE(b3u) | - | - | - | - | - | - | - | 2.4634 |
| N | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| OnumIn | 0.6341 | 0.3385 | 0.2832 | 0.6400 | 0.6633 | 0.6799 | **0.6947** | 0.2587 |
| OnumCov | 0.4095 | 0.2267 | 0.1921 | 0.4116 | **0.4261** | 0.3160 | 0.3206 | 0.1876 |
| OnumHCov | 0.6270 | 0.5523 | 0.5209 | 0.6302 | **0.6318** | 0.5958 | 0.6056 | 0.5035 |
| OnumLCov | 0.6239 | 0.6018 | 0.6119 | 0.6232 | **0.6260** | 0.6003 | 0.6009 | 0.6055 |
| OnumOverLap | 0.3140 | 0.3805 | 0.3975 | 0.3069 | 0.3106 | 0.4071 | 0.4092 | **0.4371** |

Results from 1000 samples of "Random Data 4 - Cauchy Distribution" for iterative information-theoretic (IT) estimators, iterative Ordinary Least Squares (iOLS), and existing methods. The first 50 observations ($N/2$) are used for estimation and the second-half are utilized for prediction. The estimated coefficients; Standard Errors; Mean Square Errors; and out-of-sample Root Mean Squared Errors (ORMSE), Mean Absolute Errors (OMAE), and coverage statistics are averaged over the 1000 samples. The iterative IT-GME uses $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$.

TABLE 3. Mixed Models Data (T=1000): Results from iterative IT, OLS, and existing methods

| | **iGME** | **iEL** | **iME** | **iOLS** | **CM** | **CRM** | **CCRM** | **MinMax** |
|---|---|---|---|---|---|---|---|---|
| | **maxObj** | **maxObj** | **maxObj** | **minObj** | **Est.C** | **Est.C** | **Est.C** | **Est.L** |
| **b1n/c/l** | -1.8430 | -0.0365 | 0.1159 | -1.8917 | 0.9634 | 0.9634 | 0.9634 | -2.0052 |
| **b2n/c/l** | 2.9998 | 1.5133 | 0.9064 | 3.0000 | 0.0005 | 0.0005 | 0.0005 | 3.0000 |
| **b3n/c/l** | -4.0010 | -2.0182 | -1.2088 | -3.9999 | 0.0004 | 0.0004 | 0.0004 | -3.9999 |
| **StdErr(b1n/c/l)** | 0.3927 | 0.0579 | 0.4518 | 0.4092 | 1.6129 | 1.6129 | 1.6129 | 0.4050 |
| **StdErr(b2n/c/l)** | 0.0056 | 0.0081 | 0.0635 | 0.0059 | 0.0223 | 0.0223 | 0.0223 | 0.0058 |
| **StdErr(b3n/c/l)** | 0.0065 | 0.0067 | 0.0525 | 0.0068 | 0.0259 | 0.0259 | 0.0259 | 0.0067 |
| | | | | | **Est.R** | **Est.R** | **Est.R** | **Est.U** |
| **b1/r/u** | 1.8930 | 1.0462 | 2.3674 | 2.1028 | - | 277.0595 | 177.0981 | 1.9950 |
| **b2 /r/u** | -2.9998 | -2.2729 | -3.0000 | -3.0001 | - | -0.1308 | 4.4624 | -3.0000 |
| **b3 /r/u** | 4.0013 | 3.0300 | 3.9999 | 4.0001 | - | 1.5484 | 16.0689 | 4.0001 |
| **StdErr(b1/r/c)** | 0.4221 | 0.0348 | 0.0917 | 0.4396 | - | 249.7190 | 252.4322 | 0.4354 |
| **StdErr(b2/r/u)** | 0.0056 | 0.0050 | 0.0128 | 0.0059 | - | 11.8550 | 11.9834 | 0.0058 |
| **StdErr(b3/r/u)** | 0.0065 | 0.0041 | 0.0106 | 0.0068 | - | 47.5491 | 48.0667 | 0.0067 |
| **objn** | 59.7581 | 0.0377 | 3.9120 | 1.0581 | - | - | - | - |
| **obj** | 59.1008 | 0.0267 | 3.9120 | 1.0574 | - | - | - | - |
| **RMSE.L** | 1.0233 | 84.7854 | 117.8113 | 1.3904 | 170.1424 | 91.6257 | 92.4881 | **0.9576** |
| **RMSE.U** | 4.6535 | 43.0419 | 1.2683 | 1.6837 | 170.1427 | 91.6255 | 92.7471 | **0.9576** |
| **MAE.L** | 0.8352 | 70.7774 | 98.2036 | 1.1470 | 141.9623 | 76.3258 | 77.2664 | **0.7678** |
| **MAE.U** | 3.8142 | 36.1484 | 1.0841 | 1.3868 | 141.9658 | 76.3316 | 77.4477 | **0.7679** |
| **MSE(b1n/c/l)** | 0.5342 | 6.3724 | 10.4829 | 0.7574 | 3.7780 | 3.7780 | 3.7780 | **0.1809** |
| **MSE(b2n/c/l)** | **0.0000** | 2.9889 | 4.8966 | **0.0000** | 9.0035 | 9.0035 | 9.0035 | **0.0000** |
| **MSE(b3n/c/l)** | **0.0001** | 5.3239 | 8.6960 | **0.0001** | 15.9978 | 15.9978 | 15.9978 | **0.0001** |
| **MSE(b1/u)** | **0.1987** | 2.8612 | 1.9436 | 0.7592 | - | - | - | 0.2088 |
| **MSE(b2/u)** | **0.0000** | 1.3908 | **0.0000** | **0.0000** | - | - | - | **0.0000** |
| **MSE(b3/u)** | **0.0000** | 2.4746 | **0.0000** | 0.0001 | - | - | - | 0.0001 |
| **N** | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| **numIn** | 0.9873 | 0.8710 | 0.7982 | 0.9951 | 0.0101 | **1.0000** | **1.0000** | 0.9984 |
| **numCov** | 0.0066 | 0.0056 | 0.0030 | 0.0031 | 0.0000 | 0.5342 | **0.5373** | 0.0032 |
| **numHCov** | 0.4898 | 0.2722 | 0.5384 | 0.5097 | 0.0056 | 0.5467 | **0.5487** | 0.4994 |
| **numLCov** | 0.4875 | 0.0461 | 0.0059 | 0.4884 | 0.0047 | 0.5456 | **0.5480** | 0.5006 |
| **numOverLap** | 0.9726 | 0.6127 | 0.6419 | 0.9795 | 0.0043 | 0.5782 | 0.5750 | **0.9838** |

Results from 1000 samples of "Mixed Models Data" for iterative information-theoretic (IT) estimators, iterative Ordinary Least Squares (iOLS), and existing methods. For all iterative IT estimators and iOLS, two models are chosen within each sample: the highest entropy (objective) values for sets of $\beta$ and $-\beta$, respectively. The estimated coefficients; Standard Errors; Mean Square Errors; and in-sample Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE); and coverage statistics are averaged over the 1000 samples. The iterative IT-GME uses $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$.

TABLE 4. Cholesterol Data: Results from iterative IT and existing methods

| | iGME | iEL | iME | iOLS | CM | CRM | CCRM | MinMax |
|---|---|---|---|---|---|---|---|---|
| | maxObj | maxObj | maxObj | minObj | Est.C | Est.C | Est.C | Est.L |
| b1 | 116.1925 | 96.1071 | 96.1071 | 128.4643 | 124.0536 | 124.0536 | 124.0536 | 96.1071 |
| b2 | 0.9341 | 0.6607 | 0.6607 | 0.8821 | 0.8821 | 0.8821 | 0.8821 | 0.6607 |
| StdErr(b1) | 8.7251 | 0.4476 | 1.1842 | 9.8120 | 9.5375 | 9.5375 | 9.5375 | 10.1621 |
| StdErr(b2) | 0.1380 | 0.0585 | 0.1548 | 0.1822 | 0.1630 | 0.1630 | 0.1630 | 0.1887 |
| | | | | | Est.R | Est.R | Est.R | Est.U |
| b1 | - | - | - | - | - | 64.7143 | 64.7143 | 149.7857 |
| b2 | - | - | - | - | - | 0.4429 | 0.4429 | 1.1036 |
| StdErr(b1) | - | - | - | - | - | 10.3956 | 10.3956 | 12.0234 |
| StdErr(b2) | - | - | - | - | - | 0.1930 | 0.1930 | 0.1901 |
| N | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |

Results from "Cholesterol Data" for iterative information-theoretic (IT) estimators, iterative Ordinary Least Squares (iOLS), and existing methods. For all iterative IT estimators and iOLS, we choose the model with the highest entropy (objective) values. The iterative IT-GME uses $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$.

TABLE 5. Blood Pressure Data: Results from iterative IT and existing methods

| | iGME | iEL | iME | iOLS | CM | CRM | CCRM | MinMax |
|---|---|---|---|---|---|---|---|---|
| | maxObj | maxObj | maxObj | minObj | Est.C | Est.C | Est.C | Est.L |
| b1 | 16.6215 | 22.5767 | 22.5767 | 16.7927 | 21.1708 | 21.1708 | 21.1708 | 22.5766 |
| b2 | 0.3187 | 0.2654 | 0.2654 | 0.3203 | 0.3289 | 0.3289 | 0.3289 | 0.2654 |
| b3 | 0.2098 | 0.1953 | 0.1953 | 0.2062 | 0.1699 | 0.1699 | 0.1699 | 0.1953 |
| StdErr(b1) | 15.6136 | 0.2344 | 0.7774 | 19.6352 | 18.4290 | 18.4290 | 18.4290 | 23.6213 |
| StdErr(b2) | 0.0877 | 0.0234 | 0.0776 | 0.1103 | 0.1074 | 0.1074 | 0.1074 | 0.1579 |
| StdErr(b3) | 0.1115 | 0.0258 | 0.0856 | 0.1402 | 0.1327 | 0.1327 | 0.1327 | 0.1626 |
| | | | | | Est.R | Est.R | Est.R | Est.U |
| b1 | - | - | - | - | - | 20.2149 | 17.9556 | 34.0810 |
| b2 | - | - | - | - | - | -0.1467 | 0.0000 | 0.3062 |
| b3 | - | - | - | - | - | 0.3480 | 0.2072 | 0.1089 |
| StdErr(b1) | - | - | - | - | - | 9.5907 | 9.8768 | 21.0029 |
| StdErr(b2) | - | - | - | - | - | 0.2107 | 0.2170 | 0.1089 |
| StdErr(b3) | - | - | - | - | - | 0.4431 | 0.4563 | 0.1658 |
| N | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |

Results from "Blood Pressure Data" for iterative information-theoretic (IT) estimators, iterative Ordinary Least Squares (iOLS), and existing methods. For all iterative IT estimators and iOLS, we choose the model with the highest entropy (objective) values. The iterative IT-GME uses $\mathbf{z} = (-500, 250, 0, 250, 500)$ and $\mathbf{v} = (-3\hat{\sigma}_Y, 0, 3\hat{\sigma}_Y)$.