Econometric Reviews

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: http://www.tandfonline.com/loi/lecr20

Interval estimation: An information theoretic approach

Amos Golan & Aman Ullah

To cite this article: Amos Golan & Aman Ullah (2017): Interval estimation: An information theoretic approach, Econometric Reviews, DOI: 10.1080/07474938.2017.1307573

To link to this article: http://dx.doi.org/10.1080/07474938.2017.1307573

Accepted author version posted online: 24 Mar 2017.
Published online: 24 Mar 2017.

ECONOMETRIC REVIEWS
https://doi.org/10.1080/07474938.2017.1307573

Taylor & Francis
Taylor & Francis Group

Interval estimation: An information theoretic approach

Amos Golan[a,b] and Aman Ullah[c]

[a]Department of Economics and Info-Metrics Institute, American University, Washington, District of Columbia, USA; [b]Santa Fe Institute, Santa Fe, New Mexico, USA; [c]Department of Economics, UC Riverside, Riverside, California, USA

ABSTRACT

We develop here an alternative information theoretic method of inference of problems in which all of the observed information is in terms of intervals. We focus on the unconditional case in which the observed information is in terms the minimal and maximal values at each period. Given interval data, we infer the joint and marginal distributions of the interval variable and its range. Our inferential procedure is based on entropy maximization subject to multidimensional moment conditions and normalization in which the entropy is defined over discretized intervals. The discretization is based on theory or empirically observed quantities. The number of estimated parameters is independent of the discretization so the level of discretization does not change the fundamental level of complexity of our model. As an example, we apply our method to study the weather pattern for Los Angeles and New York City across the last century.

KEYWORDS

Discretization; entropy; information; moment conditions; weather patterns

JEL CLASSIFICATION

C13; C51; C63

1. Introduction

There are many types of information used in applied sciences. Only a part of that information is observed. In fact, most often the question of interest is about some unobserved entity say preferences, whereas the observed information is in terms of the observed actions. A major issue is how to connect the two, and then what inferential procedure should be used to minimize the imposition of structures and assumptions that cannot be validated with the observed information. With the objective of inferring the complete interval, we develop here an alternative information theoretic method that is (i) the most conservative in terms of the amount of a-priori information imposed on the model and (ii) based on a minimal set of parameters needed to fully characterize the joint distribution. We view this method as a complement to other information theoretic approaches when very little information about the interval is observed.

We concentrate here on inferential problems in which all of the observed information is in terms of intervals. Examples include stock or commodity market data, such as low and high daily prices or returns, income data, weather information where we observe the minimal and maximal daily temperature at a certain location, and blood pressure data or other medical data. In this paper, we focus on the unconditional case where the observed interval information is in terms of the minimal and maximal values at each period, say the lowest and highest daily temperatures or individuals' expenses by types and periods, say minimal and maximal monthly expenses on food, travel, gas, etc. Given interval data, we infer the joint and marginal distributions of the interval variable and its ranges.

CONTACT Amos Golan agolan@american.edu Department of Economics and Info-Metrics Institute, American University, Washington, District of Columbia, USA; Santa Fe Institute, Santa Fe, New Mexico, USA.

Our inferential procedure is based on entropy maximization subject to multidimensional moment conditions and normalization in which the entropy is defined over discretized intervals. The discretization is based on theory or empirically observed quantities. The number of estimated parameters is independent of the discretization. Thus, the level of discretization does not change the fundamental level of complexity of our model.

There is a whole class of interval estimation approaches developed in the last decade. But they all are based on assumptions that we like to avoid. For example, researchers often transform interval-valued data into single points such as center, mean, ranges, and minimum and maximum bounds. Then descriptive statistics, principal components, clustering, and more recently regression methods are extended to analyze interval-valued data. Billard and Diday (2000) estimated parameters of linear regression models using the center points of the interval data and used lower and upper bounds of the explanatory variables to predict the lower and upper bounds of the dependent variable. This center method is well known to ignore the internal variations within observations. Others developed regression estimation using center and range method (De A. Lima Neto and de Carvalho, 2008; De Carvalho et al., 2004), considering center and range as bivariate random variable (Billard, 2007), developing Min/Max method (Billard and Diday, 2002; De A. Lima Neto et al., 2005) that fits two separate regressions for minimum and maximum observed values, exploring symbolic covariance method (Billard, 2002, 2008; Xu, 2010), and applying maximum likelihood and least-squares approaches as in Xu and Billard (2012) and Le-Rademacher and Billard (2010). Recently, developing and applying interval-valued data-based estimation method to economic and econometric issues have been initiated in the work by Arroyo and Gonzalez-Rivera (2012), Gonzalez-Rivera and Lin (2013), He et al. (2011), and Manski and Tamer (2003), among others. However, all the above work utilizes only the first few moments of the observed intervals, not their entire distribution. Thus, not all of the observed information is used. For more references and discussions on this, see Tuang (2016).

As an alternative, we propose here an information-theoretic (IT) inferential approach. That approach is simple to use and compute and works especially well with small or ill-behaved data. Its statistical properties are derived directly from the method of maximal entropy and are similar (under our proposed framework) to a special case of the maximum likelihood logit. We start by discretizing the interval-valued variable where the discretization is empirically based and justified by theory. Then, using the Boltzmann–Shannon entropy (Shannon (1948)), our approach builds directly on the classical maximum entropy (ME) formulation (Jaynes, 1957) and on extensions of that formulation (Golan, 1994, Golan 2008) to study the joint and marginal distributions of the interval variable and its ranges.

It is also important to mention the innovative work of Wu and Perloff (2005, 2007) that is directly connected to our proposed method. Their IT method extends the maximum entropy formalism for inferring the underlying distribution of a certain variable when the only available information is in terms of summary statistics of intervals rather than for the entire distribution. Though their approach builds on the same information theoretic foundations as our proposed method, the two are quite different in the underlying structure and assumptions imposed. In their approach, they assume a certain flexible functional form of the unknown density function, imposed some additional structure (that may come from economic theory) and use the differential entropy. In our approach, we do not use any structure except the observed interval data and we use Shannon entropy (not the differential one). In addition, we show below the simplicity, efficiency, and generality of our approach.

Our proposed model extends the traditional maximum entropy formulation in a number of ways. First, we discretize our model so each continuous value can be captured in a discrete way. Second, we extend the approach to interval estimation. Third, we can incorporate additional information in terms of priors or other direct or indirect conditional information. Overall, we view our approach as a complement to other IT approaches, especially for problems where the observed information is naturally bounded in a certain interval.

In the next section, we develop our IT approach for inference of interval information. Once we derive the inferred solution, we construct the concentrated model and compare it to the traditional log likelihood. In Section 3, we apply our method to study the weather pattern in New York City and

Los Angeles over a period of approximately one hundred years. We then, in Section 4, extend and generalize our method. We concentrate on two main issues. First, we show that under certain conditions in terms of the relationships among the intervals of interest, our inferred distribution reduces the familiar in Bose–Einstein distribution. We then extend it farther by allowing different functional relations (linear or nonlinear) among the intervals. Last, we extend it to include as many conditional variables or intervals as desired. We conclude in Section 5.

2. Interval estimation—an information-theoretic framework

2.1. Definitions, information and constraints

Consider the common case where all observed information is in terms of intervals and their expectation values (moments). Examples include financial data in which only the high and low prices are recorded daily, or commodities prices and volumes as recorded by government agencies, such as the United States Department of Agriculture (USDA), or even default data that are reported as intervals for given characteristic such as a Fair, Isaac and Company (FICO) score (or credit score of an individual in terms of the likelihood that this person will pay her/his debt or mortgage). There are a number of ways to represent each interval depending on the available information and the problem at hand. We concentrate here on the case where the only available information, over some well-defined scale, is in terms of the minimal and maximal points of an interval. The first question is how can we capture the interval information? Our objective is to do so with minimum structure and in a natural way.

We observe $i = 1, \ldots, n$ observations. For each i, we observe the maximal, Max (i), and minimal, Min (i), values, respectively. We define the center of each interval for observation i as $C_i = [\text{Max}\,(i) + \text{Min}\,(i)]/2$ and $\bar{C}$ as the sample's mean. We now specify a normalized center (deviation from the mean) as $D_i = C_i - \bar{C}$ for each $i = 1, 2, \ldots, n$. Next, we define the range of each interval as $R_i = \text{Max}\,(i) - \text{Min}\,(i)$. We have defined our interval in such a way that we can now capture all of the available information on each interval.

Given these expected values, we want to infer the joint probability distribution over location (deviations from mean) and range. As a first step, we discretize the D's and R's as follows:

$$D_{n_i} = (n_i)\, d \tag{1}$$

$$R_{k_i} = (k_i)\, r, \tag{2}$$

where n_i and k_i are the integers (for each i), d is a basic minimal unit of interest in terms of the location deviations (say, one standard deviation in units of D_i), and r is a basic minimal unit of interest in terms of the range (for example, one standard deviation in units of R_i). We note that given D_i and d, for the ith observation, n_i is just the integer value of D_i/d. We provide examples of the data transformation in Appendix A.

This discretization allows us to capture the joint, and marginal, distributions of the fundamental unit of interest. We view it as a "quantum" capturing the smallest unit of an entity. It can be a theoretical value, or in data analysis, it is the observed unit or the unit of interest. Once that unit is defined (theoretically or empirically), the distributions are defined on integers multiples of these units. In this way, we are able to discretize at any level, or bandwidth, of interest. In our empirical example, the units we use are fractions of standard deviations.

To simplify notations, but without loss of generality, from here on, we write $D_n = nd$, and similarly, $R_k = kr$ (ignoring the individual index i). With this formulation, our objective is to find the joint distribution of the integers n and k which captures the joint distribution of D and R. To do so, we define the quantity q_{nk} as the number of intervals with location in the range of nd to $(n+1)\,d$ and in the range of kr to $(k+1)\,r$. The integers can be negative and positive. We can now specify all of the observed

information in terms of three basic equations (or two basic moments).

$$\begin{aligned} \sum_{n,k} q_{nk} &= N \\ N\bar{D} &= \sum_{n,k} q_{nk} nd \\ N\bar{R} &= \sum_{n,k} q_{nk} kr. \end{aligned} \tag{3}$$

The first equation is just the total number of interval types, where *type* is defined as a function of some characteristics of the interval: $type = type\,(D_i,\ R_i)$. In our empirical example, it is in terms of standard deviations. The second is the conservation rule (constraint) for the location whereas the third is the equation of the range. We note, however, that the sums in (3) are taken such that they include all possible dependencies of k on n and are provided by the data or theory. This is another piece of information that is not always known or observed. If it is known, we should use it. For example, higher ranges of the interval may be associated with cases where observation i's daily mean temperature is much above (or below) the sample's (say annual) average, or similarly, when the spread of a stock return is larger for the relatively higher (in magnitude) returns. We provide examples below.

The set of Eq. (3) is the only information we have. We want to infer the high-dimensional quantity q_{nk}, while using only that information. But as we clearly see, the number of moments, called also conservation rules or commonly known as "constraints," is much smaller than the number of unknown quantities to be inferred. We do not have enough equations to determine the q_{nk}s uniquely. The problem is under determined. There are infinitely many q_{nk}s that satisfy these constraints. We must choose one. To do so, we convert this under-determined problem into a well-posed constrained optimization problem and choose a certain criterion that will pick out one of the solutions. The criterion we use is the familiar Boltzmann–Shannon entropy (Shannon, 1948). This inferential approach is known as the maximum entropy formulation (Jaynes, 1957). Our proposed method builds directly on that formulation.

2.2. Formulation—the constrained model

We now formulate our IT inferential model. Instead of using q_{nk}, we prefer to use the probabilities $p_{nk} \equiv q_{nk}/N$. Thus, we want to infer p_{nk}. We do so via the maximum entropy procedure. We maximize the joint entropy (over n and k) subject to the above two constraints [the first equation in (3) can just be subsumed within the other two] and a normalization constrained enforcing the inferred probability distribution to be proper. Specifically,

$$\begin{aligned} &\underset{\{P\}}{\text{Maximize}}\ \ H\,(P) = -\sum_{n,k} p_{nk} ln\left(p_{nk}/g_{nk}\right) \\ &\text{subject to} \\ &\bar{D} = \sum_{n,k} p_{nk} nd; \quad \bar{R} = \sum_{n,k} p_{nk} kr; \quad \sum_{n,k} p_{nk} = 1, \end{aligned} \tag{4}$$

where $H\,(P)$ is the Boltzmann–Shannon entropy, "ln" is the natural log, and for generality, we introduce in the entropy functional H the quantity g_{nk} capturing the multiplicity of state nk. In many problems, $g_{nk} = 1$ that is what we will assume in this paper. (This is equivalent for assuming uniform priors for each p_{nk}.)

Formulating the Lagrangian and solving, we get the optimal solution:

$$p^*_{nk} = \frac{exp\left(-\lambda^*_D nd - \lambda^*_R kr\right)}{\sum_{n,k} exp\left(-\lambda^*_D nd - \lambda^*_R kr\right)} \equiv \frac{exp\left(-\lambda^*_D nd - \lambda^*_R kr\right)}{\Omega\left(\lambda^*_D, \lambda^*_R\right)}, \tag{5}$$

where λ_D^* and λ_R^* are the inferred Lagrange multipliers associated with the constraints $\bar{D}$ (location) and $\bar{R}$ (range), respectively (the unknown parameters in the maximum entropy distribution), "$*$" stands for the inferred solution, and $\Omega\left(\lambda_D^*, \lambda_R^*\right) = \sum_{n,k} exp\left(-\lambda_D^* nd - \lambda_R^* kr\right)$ is a normalization factor. As always, the Lagrange multipliers are determined from the data.

Because the entropy is a well-behaved concave function and the constraints are linear, the maximum entropy solution p_{nk}^* (or P^*) is globally optimal (a known result which is not provided here). For detailed derivations (including the covariance and corresponding test statistics) and historical perspectives, see, for example, Jaynes (1957), Levine (1980), a recent review by Golan (2008), and a complete derivation with examples in Golan (2017). The concentrated function is derived in Appendix B.

Finally, we have emphasized earlier that the sums in (3), or the constraints in (4), are taken such that they include all possible dependencies of k on n provided by the theory or the data. These include conditionalities and correlations among the different quantities and intervals. More formally, we express the dependence below via the normalization function Ω. Although this dependency is problem specific, we specify it here as the dependence of the range on the location. (To simplify notations we omit the $*$s used in the optimal solution.)

$$\begin{aligned}\Omega\left(\lambda_D, \lambda_R\right) &= \sum_{n,k} e^{(-\lambda_D nd - \lambda_R kr)} = \sum_{n} e^{(-\lambda_D nd)} \left[\sum_{k(n)} e^{(-\lambda_R k(n) r)}\right] \\ &= \sum_{n} e^{(-\lambda_D nd)} \left[\sum_{k} \gamma_k(n) e^{(-\lambda_R k(n) r)}\right], \end{aligned} \tag{6}$$

where $\gamma_k(n)$ is some known function capturing the dependency of k on n (range on location). If, for example, we are modeling input, captured by the integers n, and output (captured by k) relationship, then the dependency function $\gamma_k(n)$ is just a multidimensional production function of any desired functional form. We provide an explicit example of such a function below.

2.3. The Lagrange multipliers

The Lagrange multipliers have two natural interpretations. They arise from two separate viewpoints or formulations: information theory and statistics. They connect the two. From information (and optimization) theoretic point of view, they capture the relative information of each one of the constraints. It is the marginal amount of information, a certain constraint contributed to the (reduction of the entropy of the) inferred distribution. The larger the magnitude of that multiplier, the larger its contribution relative to all other information used. But it is only a relative measure. It is relative to the information set used. If, on the other hand, an estimated Lagrange multiplier is practically zero, it means that there is no additional information in the corresponding equation. A hypothesis about these multipliers is a hypothesis about the relative informational content of the constraint.

The statistical interpretation is natural as well. The inferred Lagrange multipliers are the estimated values of the parameters in the probability distribution of interest—the exponential distribution with number of parameters equals the number of observed constraints (moments): two in our model. We can view it as optimizing an exponential likelihood function.

As always, with such problems, the quantity of interest is the marginal effects of p_{nk} with respect to n and k in which these marginal effects are calculated as discrete effects (between n and $n+1$ or similarly between k and $k+1$).

It is also insightful to transform the two Lagrange multipliers in the following way:

$$\begin{aligned} T &= 1/\lambda_R \\ w &= \lambda_D/\lambda_R. \end{aligned} \tag{7}$$

Then, building on (3)–(5) and the Lagrangian (Appendix B), we have

$$\bar{R} - TH + w\bar{D} = -T \ln \Omega(\lambda) \equiv z, \tag{8}$$

where $H(P)$ is the entropy defined above. We can now express the normalization function $\Omega(\lambda)$ as a function of the transformed multipliers:

$$\Omega(T, w) = \sum_{n,k} e^{(-wnd/T - kr/T)}. \tag{9}$$

From a more theoretical point of view, it is worth thinking of the interpretation of T and w. From (8), we have $T = \frac{\partial \bar{R}}{\partial H(P)}$ and $w = -\frac{\partial \bar{R}}{\partial \bar{D}}$. The first equation identifies T as the rate of change of the mean range with the entropy (uncertainty) of P where T is a monotone function of $\bar{R}$. It captures the sensitivity of the inferred distribution to the observed moment $\bar{R}$. The smaller the corresponding multiplier λ_R (the larger the T), the smaller is the impact of that constraint on the optimal solution. Stated differently, the larger the T, the more stable is our solution for small perturbations about the optimal solution. The exact relationship between $\bar{R}$ and H is problem specific. We will return to this in the empirical example.

The second equation captures the change in the expected range with respect to the expected location while holding H (entropy) constant. The sign of w is problem specific, though we expect it to be negative (for positive integers k) for the following reasons. The range of the interval will decrease as the expected value decreases toward its mean value ($w < 0$). Mathematically, it follows from (9) that for $T < 0$, the sum in that equation will not converge unless $w < 0$. Our empirical analysis confirms this argument. Finally, for a fix N, it is easy to verify that $\frac{\partial H(P)}{\partial \bar{D}} = 0 = \frac{\partial H(P)}{\partial \bar{R}} \frac{\partial \bar{R}}{\partial \bar{D}} + \frac{\partial H(P)}{\partial \bar{D}}$ which means $\frac{w}{T} = \frac{\partial H(P)}{\partial \bar{D}}$.

The above formulation has the following interpretation. $\bar{R}$ captures the mean range of the intervals. Therefore, it may be considered as a measure of variability (uncertainty). The parameter λ_R captures the impact of $\bar{R}$ on the optimal solution. T is the inverse of λ_R (information in R). As $T \to \infty$, the optimal solution is more stable. Consequently, large/small T implies low/high uncertainty. Similarly, w may be interpreted as the uncertainty in D relative to R.

Having transformed the multipliers, we can specify the normalization function and the dependence (6) of range on location as

$$\Omega(T, w) = \sum_{n} e^{-wnd/T} \left[\sum_{k(n)} e^{-k(n)r/T} \right] = \sum_{n} e^{-wnd/T} \sum_{k} \gamma_n(k) e^{-k(n)r/T}. \tag{10}$$

Equation (10) is general for all possible different conditionalities among the different variables (say $\bar{R}$ on $\bar{D}$, or k on n). In Section 4, we derive a special case where we know (or assume to know) the exact relationship between the range of the interval and its mean.

3. Illustrative examples: weather pattern analysis

To demonstrate the performance and simplicity of our approach, we provide here a concise empirical analysis of weather data.

3.1. Data and transformation

We use data for two locations: New York City (NYC) and Los Angeles (LA). In each case, we studied two full calendar years of data taken from the same weather station location. For each day, we observe the minimal and maximal temperatures (measured in tenth of degrees Celsius). The data for NYC consist of the years 1900 and 2013. For LA, the data are from 1931 and 2013. The data source is publically available at the National Oceanic and Atmospheric Administration website. Rather than using one standard deviation (σ) as the fundamental unit for both d and r, as specified in (1)–(2), we use 1/3 of σ for each one. This gives us a better resolution for analyzing the current data. Furthermore, to establish a common

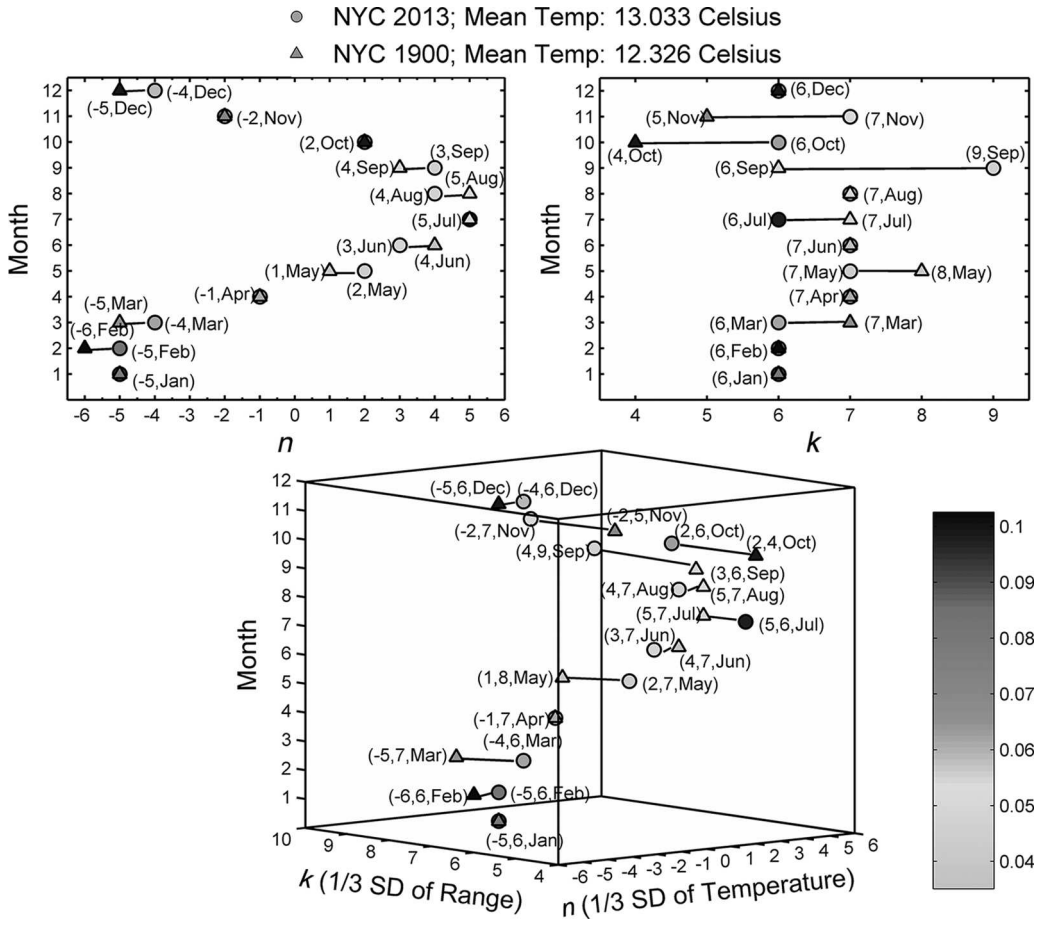

Figure 1. A heat map representation of the inferred IT interval estimation for NYC. The horizontal axis is the standard deviations of the range. The vertical axis is the standard deviations of the temperature. Both are in units of 1/3 standard deviations. Each point is in terms of deviations from the mean temperature and mean interval length for each month (and for each year). The gray scale capture the exact probability. The overall mean temperature has changed from 12.33°C (54.19 F) in 1900 to 13.03°C (55.45 F) in 2013. But the more interesting result is the change in the range of the temperature. This is easier to see in Fig. 2. *Note*: NYC, New York City.

support for both periods, for each city, we use the same d and r for both periods—those from the earlier period.

As a first step, we need to transform the original data (daily minimum and maximum temperature) into n and k and then construct the constraints. To do so, we follow the definitions of Eqs. (1)–(3). We demonstrate this transformation in great details in Appendix A where we use a full month of data (March, 1931) from the LA data set. We also show the corresponding q_{nk} table.

3.2. New York City

The inferred distribution for NYC is shown in Fig. 1. It shows the joint distribution over temperature and range. The horizontal axis is the standard deviations of the range. The vertical axis is the standard deviations of the temperature. Both are in units of 1/3 standard deviations. Each point is in terms of deviations from the mean temperature and mean interval length for each month (and for each year). The gray scale capture the exact probability as is described by the heat map. The figure also presents the average monthly values inferred for each one of the two periods (1900 and 2013). In addition to the slight increase in the mean temperature, we can easily see the change in the range during the warmer

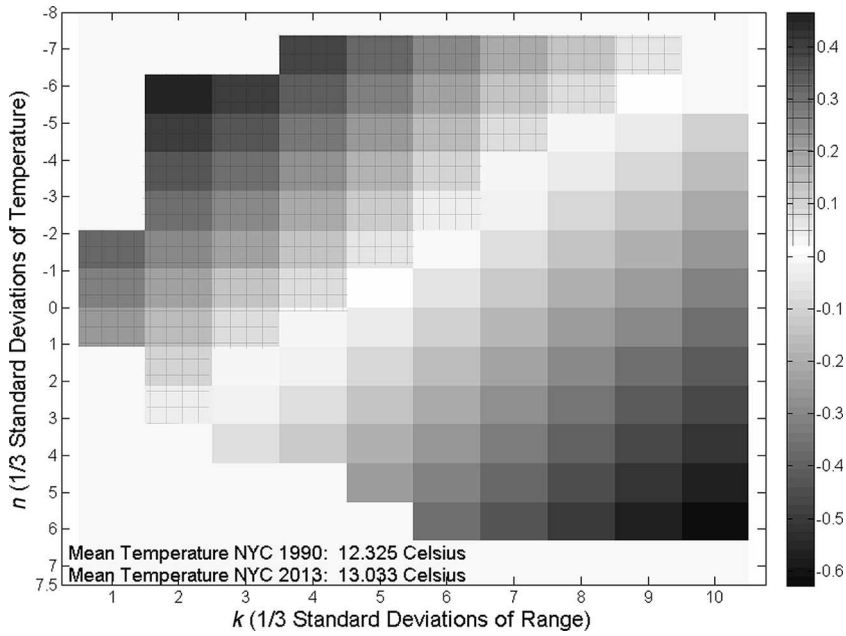

Figure 2. Surprisal analysis of the NYC weather data. The horizontal axis is the standard deviations of the range. The vertical axis is the standard deviations of the temperature. Both are in units of 1/3 standard deviations. The surprisal is defined as $I(n,k)_{\text{NYC}} = -\ln\left[p^*_{nk}(2013)/p^*_{nk}(1900)\right]$. Note that a negative surprisal (plain darker color) means an increase in the current probability relative to the initial (1900) one. The darker gray (with "squares") captures the opposite (current probability is smaller than the 1900 one). The interesting observation here is the symmetric change in range of the interval. It increases (dark gray) for temperature above average (warmer periods) while decreases (gray with "squares") for the cooler (below average) temperature. A deeper analysis, including the marginal and monthly distributions, is available upon request. *Note*: NYC, New York City.

period relative to the cooler period. The lower panel presents the joint distribution for both periods on a single picture. The top right panel is the marginal probability distribution of the range. The top left panel is the marginal distribution of the temperature. Both are in terms of 1/3 standard deviations, where, for example, the standard deviation of the R_i data is the usual square root of the sum of square of $\left(R_i - \bar{R}\right)^2/N$.

For an alternative estimator (for the standard deviations of the rate of return of a common stock based on interval data), not considered here, see Parkinson1980. He was interested in estimating the diffusion constant of a random walk process of stocks. He relates the range of the interval to the diffusion constant and then uses an extreme value approach to estimate that diffusion constant. He argued that relative to other approaches, his approach provides a better estimate of the diffusion constant that captures the variance of the rate of return of a common stock. In our case, we do not assume that our data were generated via a random walk.

It is worth noting that though the inferred distribution is defined on n and k, it is immediate to get back to the date and month (through the index i), calculate the estimated temperature and mean (in terms of deviations), and then average over the month. It is practically the reverse (with the estimated parameters) of the transformation shown in Appendix A.

The main results are easier to see in the NYC surprisal analysis presented in Fig. 2. The surprisal is defined as

$$I(n,k) = -\ln\left[p^*_{nk}(\text{current})/p^*_{nk}\left(\text{older period}\right)\right], \tag{11}$$

where 'ln' stands for the natural logarithm. The surprisal analysis for the NYC reveals some interesting phenomenon: a symmetric change in the range (of the temperature) from the cooler to the warmer

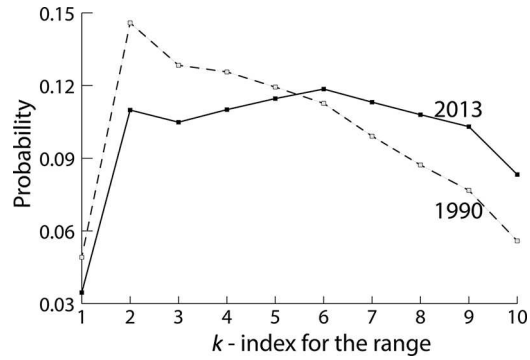

Figure 3. The marginal distribution of the range for NYC. The dashed line is the 1900 range distribution, while the solid dark one is the 2013 range distribution. *Note*: NYC, New York City.

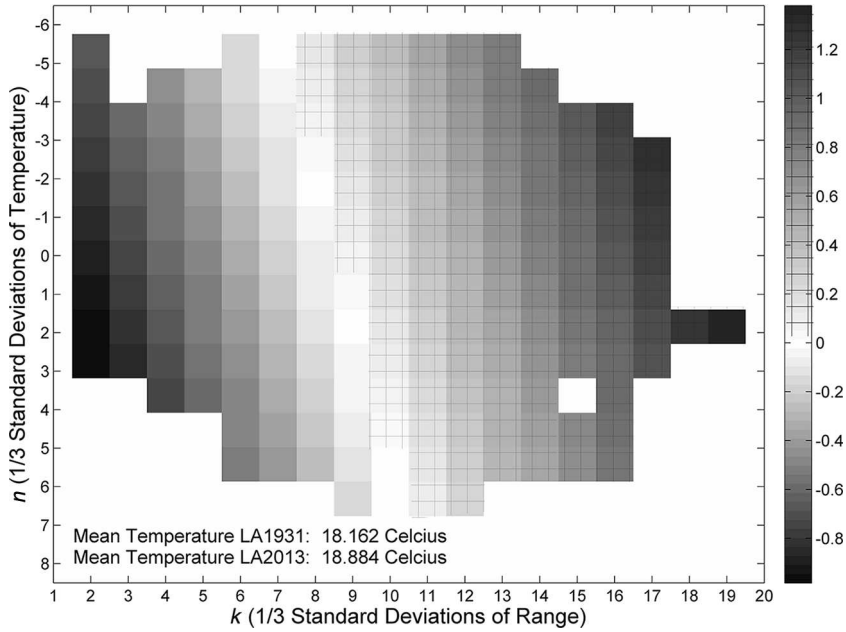

Figure 4. Surprisal analysis of the LA weather data. The horizontal axis is the standard deviations of the range. The vertical axis is the standard deviations of the temperature. Both are in units of 1/3 standard deviations. The surprisal is defined as $I(n,k)_{\mathrm{LA}} = -\ln\left[p^*_{nk}\ (2013)\ /p^*_{nk}\ (1931)\right]$. The mean temperature increased from 18.16°C (64.69 F) in 1931 to 18.88°C (65.98 F) in 2013. Unlike the NYC case, this figure shows a decrease in the range for all levels of temperature, meaning that the interval decreases for every period of the year.

periods of the year. The range increases for the warmer temperature periods while decreases during the cooler periods.

The estimated T and w for 1900 are −7.93 and −0.291, respectively. For 2013, these are −5.51, and −0.119, respectively. The signs of the w's are consistent with our earlier argument [below Eq. (9)].

Figure 3 presents the marginal distribution of the range for the two periods. The dashed line is 1900 whereas the solid dark one is the distribution of 2013. It is interesting to note the more even distribution in the later period.

3.3. Los Angeles

The LA analysis yielded quite a different result. We discussed it briefly here. Figure 4 shows the LA surprisal. In contrast to the NYC case, the LA surprisal reveals a different story. In this case, the range (or

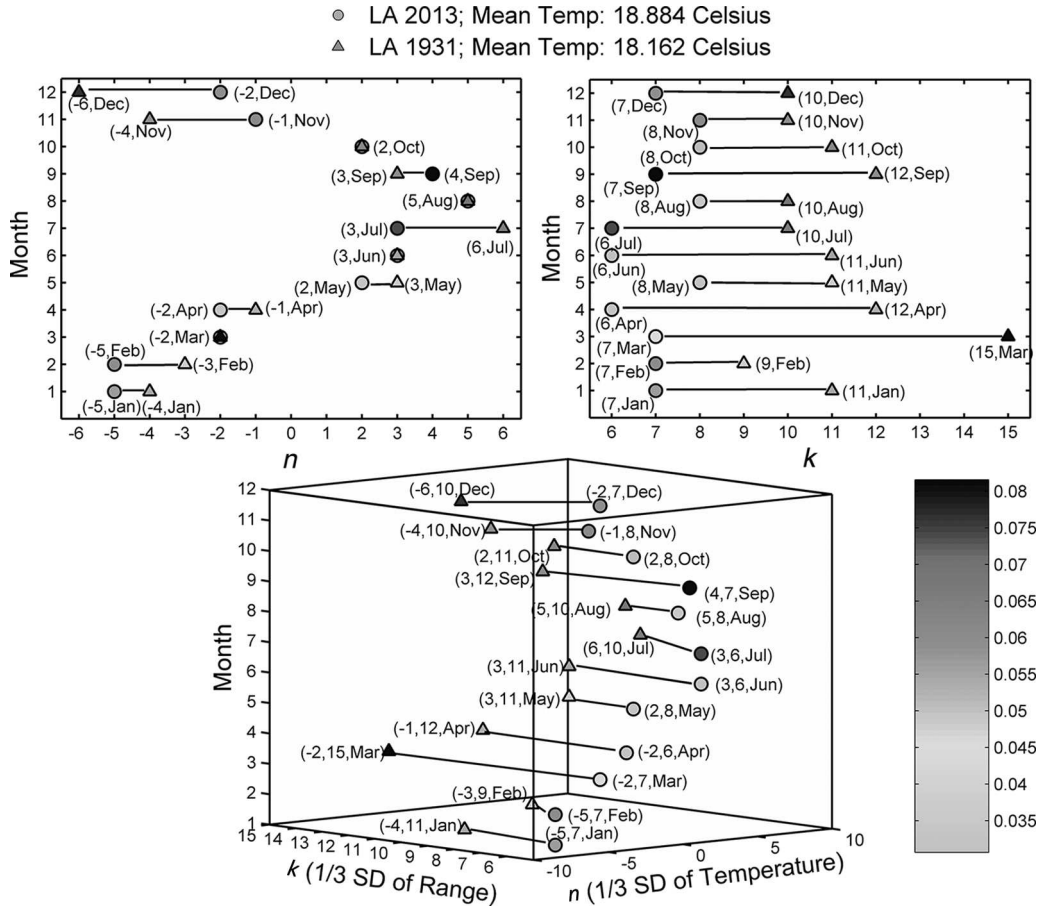

Figure 5. A heat map representation of the inferred IT interval estimation for LA. The horizontal axis is the standard deviations of the range. The vertical axis is the standard deviations of the temperature. Both are in units of 1/3 standard deviations. Each point is in terms of deviations from the mean temperature and mean interval length for each month (and for each year). The gray scale capture the exact probability. The lower panel presents the joint distribution and probability values (heat map) for both periods on a single picture. The top right panel is the marginal probability distribution of the range. The top left panel is the marginal distribution of the temperature. Both are in terms of 1/3 standard deviations.

variation about the mean) of the temperature decreased for all seasons and temperatures. The symmetric change between warmer (above mean temperature) and cooler periods that we observed in the NYC case is changed to a different symmetry: a practically similar decrease in the range throughout the seasons.

The estimated T and w in this case are -11.11 and 0.29, respectively, for 1931, and -142.86 and 5.29, respectively, for 2013. The signs of the ws in this case are also consistent with our earlier argument (Section 2.4) because the integer n is both negative and positive.

Figure 5 presents the inferred distribution for LA. It is interesting to contrast it with Fig. 1 (NYC). It shows the joint distribution over temperature and range. The horizontal axis is the standard deviations of the range, and the vertical axis is the standard deviations of the temperature. Both are in units of 1/3 standard deviations. Each point is in terms of deviations from the mean temperature and mean interval length for each month (and for each year). The colors capture the exact probability as is described by the heat map. The figure also presents the average monthly values inferred for each one of the two periods (1931 and 2013). Looking at the top right panel, together with the surprisal, reveals the consistent decrease in the range between 1931 and 2013.

4. Extensions

So far, we concentrated on the basic model. Having defined the elements of that model, the inferential procedure and solution, we now discuss possible extensions and special cases of our formulation. We concentrate on two cases. The first deals with an additional structure in which one interval is a certain function of the other. In case of the linear form, our model reduces to the celebrated Bose–Einstein distribution (e.g., Kittel, 1969).

The second extension deals with adding more information in terms of additional variables. In that case, one can directly condition the inferred distribution on other environmental or economic information.

4.1. The Bose–Einstein distribution—a special case of the interval distribution

We derive here a special, and familiar, distribution which is a special case of our generic interval framework. We show that if the range is a linear function of location, then p^*_{nk} is the familiar Bose–Einstein distribution function. To derive this, let $R_i = bD_i$, or in the notations of our basic units (and suppressing the index i), $kr = bnd$ where b is a scalar. Let $a = bd/r$ so $k = an$. Thus, $k$is a linear function of n, meaning that the range of the interval is a function of its location. One can think of it as a measure of "total risk" captured by the spread of the information, or in the weather case, we can think of it as a (linear) increase in variations for those days of extreme temperature. Within a production economic (or theory of the firm) framework, if D captures inputs and R captures outputs, this linear relationship is the constant technology which in this case is a constant returns to scale production process (Golan, 1994).

Substituting this linear relationship into Ω (and letting the sum goes to infinity) yields

$$\Omega = \sum_{n=0}^{\infty} e^{-n(\lambda_D d + \lambda_R a r)}.$$

To simplify the derivation, let $X = e^{-(\lambda_D d + \lambda_R a r)}$. Then, for $|X| < 1$, we obtain

$$\Omega = \sum_{n=0}^{\infty} X^n = \frac{1}{1-X}.$$

Substituting this into p^*_{nk} yields

$$p_{nk} = p_n = \sum_{n=0}^{\infty} nX^n \Big/ \sum_{n=0}^{\infty} X^n = X\frac{d}{dX}\sum_{n=0}^{\infty} X^n \Big/ \sum_{n=0}^{\infty} X^n$$

$$= \frac{X}{1-X} = \frac{1}{X^{-1}-1} = \frac{1}{e^{\lambda_D d(n) + \lambda_R a r(n)} - 1}.$$

This is the celebrated Bose–Einstein distribution, characterizing many physical systems, especially within thermodynamics.

4.2. Additional information and nonlinearities

We now extend the previous example in two ways. First, we show that it can be generalized to more complicated functional forms. Then, we show that it can be extended to include as many variables as needed. Rather than using the approach, we used to derive the Bose–Einstein distribution, and we derive the extensions via reformulations of the necessary moment conditions.

First, we show that the above derivation can also be done by directly imposing an additional constraint into (4) or simply by rewriting the moment $\bar{R}$. Recalling that $kr = bnd$, where b is a scalar, and $a = bd/r$ so $k = an$, we can express the two moments $\bar{D}$ and $\bar{R}$ as functions of n. Then, the inference problem from (4) gives a solution that is equivalent to the Bose–Einstein distribution derived above.

An immediate extension of the above is just $R_i = f(D_i) = \alpha + \beta D_i$ or $kr = \alpha + bnd$. A more interesting case is $R_i = f(D_i) = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \cdots$. Naturally, this can be extended to any desired form. For example, consider the quadratic form $R_i = b_0 + b_1 nd + b_2 (nd)^2$. Recalling that $R_i = k_i r$ we can express it in terms of k to get $k = \frac{b_0}{r} + \frac{b_1}{r} nd + b_2 \frac{(nd)^2}{r} = \alpha_0 + \alpha_1 n + \alpha_2 n^2$ for $\alpha_0 = b_0/r$, $\alpha_1 = b_1 d/r$ and $\alpha_2 = b_2 d^2/r$, respectively. Then, from (4), we now have $\bar{R} = \sum_{n,k} p_{nk} kr = \sum_n p_n n \left(\alpha_0 + \alpha_1 n + \alpha_2 n^2\right) r$.

Next, we consider the case of introducing more information in terms of additional covariates. These covariates can be in any form, including intervals. Let X be some variable. As before, we define a "quantum" x and use the notation j for the corresponding integer. We now have the additional constraint $\bar{X} = \sum_{n,k,j} p_{nkj} jx$ where we have p_{nkj} rather than p_{nk}. Maximizing the entropy (over n, k and j) subject to the three moment constraints and normalization yields the solution

$$p^*_{nkj} = \frac{exp\left(-\lambda^*_D nd - \lambda^*_R kr - \lambda^*_X jx\right)}{\sum_{n,k,j} exp\left(-\lambda^*_D nd - \lambda^*_R kr - \lambda^*_X jx\right)} \equiv \frac{exp\left(-\lambda^*_D nd - \lambda^*_R kr - \lambda^*_X jx\right)}{\Omega\left(\lambda^*_D, \lambda^*_R, \lambda^*_X\right)}.$$

We can then follow on previous derivations to express any function such as $R_i = f(D_i, X_i)$ or $R_i = f(X_i)$ and $D_i = g(X_i)$.

In terms of inference and testing our constraints and parameters, we can perform a χ^2 test on each one of the coefficients or constraints and an entropy-ratio test on the model as a whole. These types of tests are similar in nature to the empirical likelihood tests, or the entropy ratio tests. See, for example, Golan (2008) or Judge and Mittelhammer (2012). Using these statistics, we tested the linear and nonlinear functional forms of $R_i = f(D_i, X_i)$ discussed above on our weather data. All of these forms were rejected by the data. That makes sense for the weather data as it is hard to expect a constant relationship between the range and the mean daily temperature.

5. Discussion and summary

In this paper, we developed a simple, easy to use, efficient information-theoretic inferential method for analyzing interval data. Such data are often observed in the financial markets, in the natural sciences and in many economic problems as well. It is usually in terms of some minimal and maximal value per observation, per period. Our method builds directly on the classical maximum entropy formulation and is based on minimal statistical or distributional assumptions. We have also shown that if certain relationships exist between the observed pieces of information, we can build it directly into our inferential method. We can then test these relationships to confirm if indeed they are consistent with the observed data.

At times, we may have additional information in terms of priors arising from theory or from past experiments. Although we did not discuss it here, our method can be immediately extended to include these priors. In that case, we substitute the Kullback and Leibler (1951) relative entropy measure for the Shannon entropy one we used here. (See also Kullback, 1959).

The statistics and diagnostics for our proposed method, such as the pseudo-R^2, entropy ratio test, and other χ^2 statistics are direct extensions of the maximum entropy and generalized maximum entropy method. For a derivation and summary of the basic test statistics, see Soofi (1992) and Golan (1988, 2008).

The empirical illustration demonstrates the simplicity and applicability of our approach. It also provides some interesting insights into the joint distribution of temperature and range.

The results of this paper can be extended in a number of ways. First, it can be used, as we have shown, for inferring higher dimensional distributions. Second, it can be generalized for the realistic cases in which the expectation values are imperfect: stochastic (or noisy) moment conditions in line of the generalized maximum entropy estimator of Golanet al. (1996a); Golan et al. (1996b). Third, the assumption of knowing the relationship, like linear, between the observed pieces of information can be avoided by considering a nonparametric relationship. Fourth, when prior information is known, it can be

incorporated into our model using Kullback and Leibler (1951) criterion rather than Shannon's entropy. Last, this framework can be easily extended for inferring conditional distributions.

6. Appendix A: data transformation

In this example, we show the way we transform the original observed information, specified in terms of daily minimum and maximum temperatures, into n and k via Eqs. (1)–(3) and the definitions of D_i, C_i, R_i, $\bar{C}$, $\bar{D}$, and $\bar{R}$. We then derive the set of q_{nk}. To make these derivations easier to observe, rather than use the complete annual data set, we use a subset of only 31 observations (March) of our LA, 1931 daily data. We treat that subset though it is the full sample here. The mean average temperature, $\bar{C}$, (of these 31 observations) is 16.031. For instance on March 1, 1931, the maximum temperature is 28.3 and the minimum is 3.9. Thus, $C_i = 16.1$, $D_i = 0.0694$, and $R_i = 24.4$. Following Eq. (1), $D_i = (n_i)\, d$ where d is practically 1 (see below), we can calculate the value of the integer n_i. In this case, $n_i = 1$. Similarly, from Eq. (2) $k_i = 17$. On the last day of that month, we observed maximum of 26.7 and a minimum of 5. So, $C_i = 15.85$, $D_i = -0.18064$, $R_i = 21.7$, $n_i = -1$, and $k_i = 15$.

Table A1 presents the q_{nk} for March 1931 (using only the 31 observations). The corresponding moments are shown below the table. From the data and transformations, we have $\bar{C} = 16.031$, $\bar{D} = 0$, $\bar{R} = 21.1$, and the minimal units for D_i and R_i are $1/3\sigma(d) = 0.986$ and $1/3\sigma(r) = 1.490$, respectively.

7. Appendix B: the concentrated function and complexity

In this Appendix, we transform the maximum entropy problem (4) from a constrained optimization problem to an unconstrained, concentrated (or dual) problem. Like a concentrated likelihood function, the dual function concentrates on the minimal set of parameters: the Lagrange multipliers (one for each constraint). In addition to the increase in computational efficiency and the exponential decrease in parameters, the concentrated model allows for a direct comparison with the more traditional statistical methods.

Given the optimization problem (4), we start with the Lagrangian but without the normalization requirement. This is because the maximum entropy solution (5) already satisfies that requirement. We then substitute (5) into the Lagrangian which allows us to transform the problem from the probability space to the multipliers space. We show below the detailed derivation. In our notation underneath, we

Table A1. The Q (or q_{nk}) matrix for the March, 1931, LA data. These values are the number of intervals with location (as measured, for example, in units of 1/3 deviations from the location mean) in the range nd to $(n+1)\, d$ and range (measured, for example, in units of deviations from the mean range) in the range kr to $(k+1)\, r$. We use that matrix to calculate the frequencies $p_{nk} \equiv q_{nk}/N$ as defined in Section 2.

$\frac{k}{n}$	9	10	11	12	13	14	15	16	17	18	19	20	21
−6						1							
−5								1					
−4			1	1		1		1			1		
−3	1				1			1					
−2				1	1	1	1						
−1							1						
1		2			1				2				
2				1					1		1		
3										1			
4		1					1	1					
5										1			1
6							1						
7											1		

also emphasize explicitly the dependence of the p_{nk} on the λs: $p_{nk}(\lambda_D, \lambda_R)$.

$$\begin{aligned}
\ell(\lambda_D, \lambda_R) &= -\sum_{n,k} p_{nk} ln\left(p_{nk}\right) + \lambda_D \bar{D} + \lambda_R \bar{R} \\
&= -\sum_{n,k} p_{nk}(\lambda_D, \lambda_R)\, ln\left[\frac{exp\left(-\lambda_D nd - \lambda_R kr\right)}{\sum_{n,k} exp\left(-\lambda_D nd - \lambda_R kr\right)}\right] + \lambda_D \bar{D} + \lambda_R \bar{R} \\
&= -\sum_{n,k} p_{nk}(\lambda_D, \lambda_R)\left[\left(-\lambda_D nd - \lambda_R kr\right) - \sum_{n,k} exp\left(-\lambda_D nd - \lambda_R kr\right)\right] + \lambda_D \bar{D} + \lambda_R \bar{R} \\
&= \lambda_D \bar{D} + \lambda_R \bar{R} + \ln\left[\Omega(\lambda_D, \lambda_R)\right]. \qquad (12)
\end{aligned}$$

Looking at (12), we see that (i) the complexity level of our inferential model (number of Lagrange multipliers, or moment conditions) is independent of our choice of discretization and (ii) that our proposed IT method is similar to the multinomial logit. The first two elements to the right of the equality sign on the bottom line are just the familiar sum of the data and the unknown parameters (Lagrange multipliers here). The right-hand side term is just the log of the normalization of the exponential (logistic) distribution. Thus, this is exactly the log-likelihood logit of the discretized model that we developed here. For a complete comparison among the IT methods and multinomial logit, see Golanet al. (1996a); Golan et al. (1996b) and Golan (2008).

Acknowledgment

The authors are thankful to Huancheng Du for his help with the empirical analysis and code, to Tuang Tual for his help constructing the figures, and to the participants of the conference in honor of Essie Maasoumi at Emory University for their helpful comments. We also thank the co-editor Peter Phillips and the three reviewers.

References

Arroyo, J., Gonzalez-Rivera, G. (2012). Time series modeling of histogram-valued data: The daily histogram time series of S&P 500 intradaily returns. *International Journal of Forecasting* 28(1):20–33.

Billard, L. (2002). Dependence and variable components of symbolic interval valued data. In: Brito, P., Betrand, P., Cucumel, G., de Carvalho, F., eds. *Selected Contributions in Data Analysis and Classification*, pp. 3–12.

Billard, L. (2007). Dependencies and variation components of symbolic interval-valued data. In: Brito, P., Betrand, P., Cucumel, G., de Carvalho, F., eds. *Selected Contributions in Data Analysis and Classification*, pp. 3–12.

Billard, L. (2008). Sample covariance functions for complete quantitative data. In: *World Congress, International Association of Computational Statistics*, Yokohama, Japan.

Billard, L., Diday, E. (2000). Regression analysis for interval-valued data. In: Kiers, H. A. L., Rasson, J. P., Groenen, P. J. F., Schader, M., eds. *Data Analysis, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin, Heidelberg: Springer.

Billard, L., Diday, E. (2002). Symbolic regression analysis. In: Jajuga, K., Sokołowski, A., Bock, H. H., eds. *Classification, Clustering, and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin, Heidelberg: Springer.

Billard, L., Diday, E. (2012). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Vol. 654. Chichester, England: John Wiley & Sons, Inc.

De A. Lima Neto, E., de Carvalho, F. (2008). Center and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, 52:1500–1515.

De A. Lima Neto, E., de Carvalho, F., Freire, E. (2005). Applying constrained linear regression models to predict inter-valued data. In: Furbach, U., eds. *KI 2005: Advances in Artificial Intelligence*, pp. 92–106.

De Carvalho, F., de A Lima Neto, E., Tenorio, C. (2004). A new method to fit a linear regression model for interval-valued data. In: Biundo, S., Frühwirth, T., Palm, G., eds. *KI 2004: Advances in Artificial Intelligence*, Vol. 3238, Berlin, Heidelberg: Springer, pp. 295–306.

Golan, A. (1988). A discrete-stochastic model of economic production and a model of production fluctuations — Theory and empirical evidence. Ph.D. Dissertation. Berkeley: UC Berkeley.

Golan, A. (1994). A multivariable stochastic theory of size distribution of firms with empirical evidence. *Advances in Econometrics* 10:1–46.
Golan, A. (2008). Information and entropy econometrics—A review and synthesis. *Foundations and Trends in Econometrics* 2(1–2):1–145.
Golan, A., Judge, G., Miller, D. (1996a). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. New York: John Wiley & Sons.
Golan, A., Judge, G., Perloff, J. (1996b). A generalized maximum entropy approach to recovering information from multinomial response data. *Journal of the American Statistical Association* 91:841–853.
Golan, A. (2017). *Foundations of Info-Metrics*. Oxford University Press.
Gonzalez-Rivera, G., Lin, W. (2013). Constrained regression for inter-valued data. *Journal of Business & Economic Statistics* 31(4):473–490.
He, Y., Hong, Y., Han, A., Wang, S. (2011). Forecasting of inter valued crude oil prices with autoregressive conditional interval models. *International Statistical Review* (submitted).
Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Review* 106:620–630.
Judge, G., Mittelhammer, R. (2012). *An Information Theoretic Approach to Econometrics*. New York: Cambridge University Press.
Kittel, C. S. (1969). *Thermal Physics*. New York: John Wiley & Sons.
Kullback, S. (1959). *Information Theory and Statistics*. New York: John Wiley & Sons.
Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* 22:79–86.
Le-Rademacher, J., Billard, L. (2010). Maximum likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference* 141:1593–1602.
Levine, R. D. (1980). An information theoretical approach to inversion problems'. *Journal of Physics A* 13:91–108.
Manski, C. F., Tamer, E. (2003). Inference on regressions with interval data on a regressor or outcome. *Econometrica* 70(2):519–546.
Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business* 53(1):61–65.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27:379–423.
Soofi, E. S. (1992). A generalized formulation of conditional logit with diagnostics. *Journal of the American Statistical Association* 47:812–816.
Tuang, T. S. (2016). Interval estimation: An information theoretic approach. Ph.D. Thesis. Washington, DC: American University.
Wu, X., Perloff, J. (2005). China's income distribution, 1985–2001. *The Review of Economics and Statistics* 87(4):763–775.
Wu, X., Perloff, J. (2007). GMM estimation of a maximum entropy distribution with interval data. *Journal of Econometrics* 138(2):532–546.
Xu, W. (2010). Symbolic data analysis: Inter-valued data regression. Ph.D. Thesis. Athens, GA: University of Georgia.
Xu, W., Billard, L. (2012). A study of symbolic intervals and its application to regression. Unpublished manuscript.